Reinforcement Learning for Non-Stationary Markov Decision Processes: The Blessing of (More) Optimism

Wang Chi Cheung [1] **David Simchi-Levi** [2] **Ruihao Zhu** [2]

Abstract

We consider un-discounted reinforcement learning (RL) in Markov decision processes (MDPs) under drifting non-stationarity, *i.e.*, both the reward and state transition distributions are allowed to evolve over time, as long as their respective total variations, quantified by suitable metrics, do not exceed certain *variation budgets*. We first develop the Sliding Window Upper-Confidence bound for Reinforcement Learning with Confidence Widening (SWUCRL2-CW) algorithm, and establish its dynamic regret bound when the variation budgets are known. In addition, we propose the Bandit-over-Reinforcement Learning (BORL) algorithm to adaptively tune the SWUCRL2-CW algorithm to achieve the same dynamic regret bound, but in a *parameter-free* manner, *i.e.*, without knowing the variation budgets. Notably, learning non-stationary MDPs via the conventional optimistic exploration technique presents a unique challenge absent in existing (non-stationary) bandit learning settings. We overcome the challenge by a novel confidence widening technique that incorporates additional optimism.

1. Introduction

Consider a general sequential decision-making framework, where a decision-maker (DM) interacts with an initially unknown environment iteratively. At each time step, the DM first observes the current state of the environment, and then chooses an available action. After that, she receives an instantaneous random reward, and the environment transitions to the next state. The DM aims to design a policy that maximizes its cumulative rewards, while facing the following challenges:

- **Endogeneity:** At each time step, the reward follows a reward distribution, and the subsequent state follows a state transition distribution. Both distributions depend (solely) on the current state and action, which are influenced by the policy. Hence, the environment can be fully characterized by a discrete time Markov decision process (MDP).
- **Exogeneity:** The reward and state transition distributions vary (independently of the policy) across time steps, but the total variations are bounded by the respective variation budgets.
- **Uncertainty:** Both the reward and state transition distributions are initially unknown to the DM.
- **Bandit/Partial Feedback:** The DM can only observe the reward and state transition resulted by the current state and action in each time step.

It turns out that many applications, such as real-time bidding in advertisement (ad) auctions, can be captured by this framework (Cai et al., 2017; Flajolet & Jaillet, 2017; Balseiro & Gur, 2019; Guo et al., 2019; Han et al., 2020). Besides, this framework can be used to model sequential decision-making problems in transportation (Zhang & Wang, 2018; Qin et al., 2019), wireless network (Zhou & Bambos, 2015; Zhou et al., 2016), consumer choice modeling (Xu & Yun, 2020), ride-sharing (Taylor, 2018; Gurvich et al., 2018; Bimpikis et al., 2019; Kanoria & Qian, 2019), healthcare operations (Shortreed et al., 2010), epidemic control (Nowzari et al., 2016; Kiss et al., 2017), and inventory control (Huh & Rusmevichientong, 2009; Bertsekas, 2017; Zhang et al., 2018; Agrawal & Jia, 2019; Chen et al., 2019a).

There exists numerous works in sequential decision-making that considered part of the four challenges. The traditional stream of research (Auer et al., 2002a; Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2018) on stochastic *multi-armed bandits* (MAB) focused on the interplay between uncertainty and bandit feedback (*i.e.*, challenges 3 and 4), and (Auer et al., 2002a) proposed the classical *Upper Confidence Bound* (UCB) algorithm. Starting from (Burnetas & Katehakis, 1997; Tewari & Bartlett, 2008; Jaksch et al., 2010), a volume of works (see Section 3) have been

[1] Department of Industrial Systems Engineering and Management, National University of Singapore [2] MIT Institute for Data, Systems, and Society. Correspondence to: Wang Chi Cheung <isecwc@nus.edu.sg>, David Simchi-Levi <dslevi@mit.edu>, Ruihao Zhu <rzhu@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

	Stationary	Non-stationary
MAB	OFU (Auer et al., 2002a)	OFU + Forgetting (Besbes et al., 2014; Cheung et al., 2019b)
RL	OFU (Jaksch et al., 2010)	**Extra optimism + Forgetting (This paper)**

Table 1. Summary of algorithmic frameworks of stationary and non-stationary online learning settings.

devoted to *reinforcement learning* (RL) in MDPs (Sutton & Barto, 2018), which further involves endogeneity. RL in MDPs incorporate challenges 1,3,4, and stochastic MAB is a special case of MDPs when there is only one state. In the absence of exogeneity, the reward and state transition distributions are invariant across time, and these three challenges can be jointly solved by the *Upper Confidence bound for Reinforcement Learning* (UCRL2) algorithm (Jaksch et al., 2010).

The UCB and UCRL2 algorithms leverage the *optimism in face of uncertainty* (OFU) principle to select actions iteratively based on the entire collections of historical data. However, both algorithms quickly deteriorate when exogeneity emerge since the environment can change over time, and the historical data becomes obsolete. To address the challenge of exogeneity, (Garivier & Moulines, 2011b) considered the *piecewise-stationary* MAB environment where the reward distributions remain unaltered over certain time periods and change at unknown time steps. Later on, there is a line of research initiated by (Besbes et al., 2014) that studied the general *non-stationary* MAB environment (Besbes et al., 2014; Cheung et al., 2019a;b), in which the reward distributions can change arbitrarily over time, but the total changes (quantified by a suitable metric) is upper bounded by a *variation budget* (Besbes et al., 2014). The aim is to minimize the *dynamic regret*, the optimality gap compared to the cumulative rewards of the sequence of optimal actions. Both the (relatively restrictive) piecewise-stationary MAB and the general non-stationary MAB settings consider the challenges of exogeneity, uncertainty, and partial feedback (*i.e.*, challenges 2, 3, 4), but endogeneity (challenge 1) are not present.

In this paper, to address all four above-mentioned challenges, we consider RL in *non-stationary* MDPs where bot the reward and state transition distributions can change over time, but the total changes (quantified by suitable metrics) are upper bounded by the respective variation budgets. We note that in (Jaksch et al., 2010), the authors also consider the intermediate RL in *piecewise-stationary* MDPs. Nevertheless, we first demonstrate in Section 4.1, and then rigorously show in Section 6 that simply adopting the techniques for non-stationary MAB (Besbes et al., 2014; Cheung et al., 2019a;b) or RL in piecewise-stationary MDPs (Jaksch et al., 2010) to RL in non-stationary MDPs may result in poor dynamic regret bounds.

1.1. Summary of Main Contributions

Assuming that, during the T time steps, the total variations of the reward and state transition distributions are bounded (under suitable metrics) by the variation budgets B_r (> 0) and B_p (> 0), respectively, we design and analyze novel algorithms for RL in non-stationary MDPs. Let $D_{\max}$, S, and A be respectively the maximum diameter (a complexity measure to be defined in Section 2), number of states, and number of actions in the MDP. Our main contributions are:

- We develop the Sliding Window UCRL2 with Confidence Widening (`SWUCRL2-CW`) algorithm. When the variation budgets are known, we prove it attains a $\tilde{O}\left(D_{\max}(B_r + B_p)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$ dynamic regret bound via a budget-aware analysis.
- We propose the Bandit-over-Reinforcement Learning (`BORL`) algorithm that tunes the `SWUCRL2-CW` algorithm adaptively, and retains the same $\tilde{O}\left(D_{\max}(B_r + B_p)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$ dynamic regret bound without knowing the variation budgets.
- We identify an unprecedented challenge for RL in non-stationary MDPs with conventional optimistic exploration techniques: existing algorithmic frameworks for non-stationary online learning (including non-stationary bandit and RL in piecewise-stationary MDPs) (Jaksch et al., 2010; Garivier & Moulines, 2011b; Cheung et al., 2019a) typically estimate unknown parameters by averaging historical data in a "forgetting" fashion, and construct the *tightest* possible confidence regions/intervals accordingly. They then optimistically search for the most favorable model within the confidence regions, and execute the corresponding optimal policy. However, we first demonstrate in Section 4.1, and then rigorously show in Section 6 that in the context of RL in non-stationary MDPs, the diameters induced by the MDPs in the confidence regions constructed in this manner can grow wildly, and may result in unfavorable dynamic regret bound. We overcome this with our novel proposal of extra optimism via the confidence widening technique (alternatively, in (Cheung et al., 2020a), an extended version of the current paper, the authors demonstrate that one can leverage special

structures on the state transition distributions in the context of single item inventory control with fixed cost to bypass this difficulty of exploring time-varying environments). A summary of the algorithmic frameworks for stationary and non-stationary online learning settings are provided in Table 1.

2. Problem Formulation

In this section, we introduce the notations to be used throughout paper, and introduce the learning protocol for our problem of RL in non-stationary MDPs.

2.1. Notations

Throughout the paper, all vectors are column vectors, unless specified otherwise. We define $[n]$ to be the set $\{1, 2, \ldots, n\}$ for any positive integer n. We denote $\mathbf{1}[\cdot]$ as the indicator function. For $p \in [1, \infty]$, we use $\|x\|_p$ to denote the p-norm of a vector $x \in \mathbb{R}^d$. We denote $x \vee y$ and $x \wedge y$ as the maximum and minimum between $x, y \in \mathbb{R}$, respectively. We adopt the asymptotic notations $O(\cdot), \Omega(\cdot)$, and $\Theta(\cdot)$ (Cormen et al., 2009). When logarithmic factors are omitted, we use $\tilde{O}(\cdot), \tilde{\Omega}(\cdot)$, $\tilde{\Theta}(\cdot)$, respectively. With some abuse, these notations are used when we try to avoid the clutter of writing out constants explicitly.

2.2. Learning Protocol

Model Primitives: An instance of non-stationary MDP is specified by the tuple $(\mathcal{S}, \mathcal{A}, T, r, p)$. The set $\mathcal{S}$ is a finite set of states. The collection $\mathcal{A} = \{\mathcal{A}_s\}_{s \in \mathcal{S}}$ contains a finite action set $\mathcal{A}_s$ for each state $s \in \mathcal{S}$. We say that (s, a) is a state-action pair if $s \in \mathcal{S}, a \in \mathcal{A}_s$. We denote $S = |\mathcal{S}|$, $A = (\sum_{s \in \mathcal{S}} |\mathcal{A}_s|)/S$. We denote T as the total number of time steps, and denote $r = \{r_t\}_{t=1}^T$ as the sequence of mean rewards. For each t, we have $r_t = \{r_t(s, a)\}_{s \in \mathcal{S}, a \in \mathcal{A}_s}$, and $r_t(s, a) \in [0, 1]$ for each state-action pair (s, a). In addition, we denote $p = \{p_t\}_{t=1}^T$ as the sequence of state transition distributions. For each t, we have $p_t = \{p_t(\cdot|s, a)\}_{s \in \mathcal{S}, a \in \mathcal{A}_s}$, where $p_t(\cdot|s, a)$ is a probability distribution over $\mathcal{S}$ for each state-action pair (s, a).

Exogeneity: The quantities r_t's and p_t's vary across different t's in general. Following (Besbes et al., 2014), we quantify the variations on r_t's and p_t's in terms of their respective *variation budgets* B_r, B_p (> 0):

$$B_r = \sum_{t=1}^{T-1} B_{r,t}, \qquad B_p = \sum_{t=1}^{T-1} B_{p,t}, \tag{1}$$

where $B_{r,t} = \max_{s \in \mathcal{S}, a \in \mathcal{A}_s} |r_{t+1}(s, a) - r_t(s, a)|$ and $B_{p,t} = \max_{s \in \mathcal{S}, a \in \mathcal{A}_s} \|p_{t+1}(\cdot|s, a) - p_t(\cdot|s, a)\|_1$. We emphasize although B_r and B_p might be used as inputs by the DM, individual $B_{r,t}$'s and $B_{p,t}$'s are unknown to the DM throughout the current paper.

Endogeneity: The DM faces a non-stationary MDP instance $(\mathcal{S}, \mathcal{A}, T, r, p)$. She knows $\mathcal{S}, \mathcal{A}, T$, but not r, p. The DM starts at an arbitrary state $s_1 \in \mathcal{S}$. At time t, three events happen. First, the DM observes its current state s_t. Second, she takes an action $a_t \in \mathcal{A}_{s_t}$. Third, given s_t, a_t, she stochastically transits to another state s_{t+1} which is distributed as $p_t(\cdot|s_t, a_t)$, and receives a stochastic reward $R_t(s_t, a_t)$, which is 1-sub-Gaussian with mean $r_t(s_t, a_t)$. In the second event, the choice of a_t is based on a *non-anticipatory* policy Π. That is, the choice only depends on the current state s_t and the previous observations $\mathcal{H}_{t-1} := \{s_q, a_q, R_q(s_q, a_q)\}_{q=1}^{t-1}$.

Dynamic Regret: The DM aims to maximize the cumulative expected reward $\mathbb{E}[\sum_{t=1}^T r_t(s_t, a_t)]$, despite the model uncertainty on r, p and the dynamics of the learning environment. To measure the convergence to optimality, we consider an equivalent objective of minimizing the *dynamic regret* (Besbes et al., 2014; Jaksch et al., 2010)

$$\text{Dyn-Reg}_T(\Pi) = \sum_{t=1}^T \{\rho_t^* - \mathbb{E}[r_t(s_t, a_t)]\}. \tag{2}$$

In the oracle $\sum_{t=1}^T \rho_t^*$, the summand ρ_t^* is the optimal long-term average reward of the stationary MDP with state transition distribution p_t and mean reward r_t. The optimum ρ_t^* can be computed by solving linear program (9) provided in Section A.1. We note that the same oracle is used for RL in piecewise-stationary MDPs (Jaksch et al., 2010).

Remark 1. *When $S = 1$, (2) reduces to the definition (Besbes et al., 2014) of dynamic regret for non-stationary K-armed bandits. Nevertheless, different from the bandit case, the offline benchmark $\sum_{t=1}^T \rho_t^*$ does not equal to the expected optimum for the non-stationary MDP problem in general. We justify our choice in Proposition 1.*

Next, we review concepts of *communicating MDPs* and *diameters*, in order to stipulate an assumption that ensures learnability and justifies our offline bnechmark.

Definition 1 ((Jaksch et al., 2010) **Communicating MDPs and Diameters).** *Consider a set of states $\mathcal{S}$, a collection $\mathcal{A} = \{\mathcal{A}_s\}_{s \in \mathcal{S}}$ of action sets, and a transition kernel $\bar{p} = \{\bar{p}(\cdot|s, a)\}_{s \in \mathcal{S}, a \in \mathcal{A}_s}$. For any $s, s' \in \mathcal{S}$ and stationary policy π, the hitting time from s to s' under π is the random variable $\Lambda(s'|\pi, s) := \min\{t : s_{t+1} = s', s_1 = s, s_{\tau+1} \sim \bar{p}(\cdot|s_\tau, \pi(s_\tau))\ \forall \tau\}$, which can be infinite. We say that $(\mathcal{S}, \mathcal{A}, \bar{p})$ is a* communicating MDP *iff*

$$D := \max_{s, s' \in \mathcal{S}} \min_{\text{stationary } \pi} \mathbb{E}\left[\Lambda(s'|\pi, s)\right]$$

is finite. The quantity D is the diameter *associated with $(\mathcal{S}, \mathcal{A}, \bar{p})$.*

We make the following assumption throughout.

Assumption 1. *For each $t \in \{1, \ldots, T\}$, the tuple $(\mathcal{S}, \mathcal{A}, p_t)$ constitutes a communicating MDP with diameter at most D_t. We denote the* maximum diameter *as $D_{max} = \max_{t\in\{1,\ldots,T\}} D_t$.*

The following proposition justifies our choice of offline benchmark $\sum_{t=1}^{T} \rho_t^*$.

Proposition 1. *Consider an instance $(\mathcal{S}, \mathcal{A}, T, p, r)$ that satisfies Assumption 1 with maximum diameter D_{max}, and has variation budgets B_r, B_p for the rewards and transition kernels respectively. In addition, suppose that $T \geq B_r + 2D_{max}B_p > 0$. It holds that $\sum_{t=1}^{T} \rho_t^* \geq \max_{\Pi} \left\{ \mathbb{E}\left[\sum_{t=1}^{T} r_t(s_t^{\Pi}, a_t^{\Pi}) \right] \right\} - 4(D_{max} + 1)\sqrt{(B_r + 2D_{max}B_p)T}$. The maximum is taken over all non-anticipatory policies Π's. We denote $\{(s_t^{\Pi}, a_t^{\Pi})\}_{t=1}^{T}$ as the trajectory under policy Π, where $a_t^{\Pi} \in \mathcal{A}_{s_t^{\Pi}}$ is determined based on Π and $\mathcal{H}_{t-1} \cup \{s_t^{\Pi}\}$, and $s_{t+1}^{\Pi} \sim p_t(\cdot|s_t^{\Pi}, a_t^{\Pi})$ for each t.*

The Proposition is proved in Section A.2 of the full version (Cheung et al., 2020b). In fact, our dynamic regret bounds are larger than the error term $4(D_{\max} + 1)\sqrt{(B_r + 2D_{\max}B_p)T}$, thus justifying the choice of $\sum_{t=1}^{T} \rho_t^*$ as the offline benchmark. The offline benchmark $\sum_{t=1}^{T} \rho_t^*$ is more convenient for analysis than the expected optimum, since the former can be decomposed to summations across different intervals, unlike the latter where the summands are intertwined (since $s_{t+1}^{\Pi} \sim p_t(\cdot|s_t^{\Pi}, a_t^{\Pi})$).

3. Related Works

3.1. RL in Stationary MDPs

RL in stationary (discounted and un-discounted reward) MDPs has been widely studied in (Burnetas & Katehakis, 1997; Bartlett & Tewari, 2009; Jaksch et al., 2010; Agrawal & Jia, 2017; Fruit et al., 2018a;b; Sidford et al., 2018b;a; Wang, 2019; Zhang & Ji, 2019; Fruit et al., 2019; Wei et al., 2019). For the discounted reward setting, the authors of (Sidford et al., 2018b; Wang, 2019; Sidford et al., 2018a) proposed (nearly) optimal algorithms in terms of sample complexity. For the un-discounted reward setting, the authors of (Jaksch et al., 2010) established a minimax lower bound $\Omega(\sqrt{D_{\max}SAT})$ on the regret when both the reward and state transition distributions are time-invariant. They also designed the UCRL2 algorithm and showed that it attains a regret bound $\tilde{O}(D_{\max}S\sqrt{AT})$. The authors of (Fruit et al., 2019) proposed the UCRL2B algorithm, which is an improved version of the UCRL2 algorithm. The regret bound of the UCRL2B algorithm is $\tilde{O}(S\sqrt{D_{\max}AT} + D_{\max}^2 S^2 A)$. The minimax optimal algorithm is provided in (Zhang & Ji, 2019) although it is not computationally efficient.

3.2. RL in Non-Stationary MDPs

In a parallel work (Ortner et al., 2019), the authors considered a similar setting to ours by applying the "forgetting principle" from non-stationary bandit settings (Garivier & Moulines, 2011b; Cheung et al., 2019b) to design a learning algorithm. To achieve its dynamic regret bound, the algorithm by (Ortner et al., 2019) partitions the entire time horizon $[T]$ into time intervals $\mathcal{I} = \{I_k\}_{k=1}^{K}$, and crucially requires the access to $\sum_{t=\min I_k}^{\max I_k - 1} B_{r,t}$ and $\sum_{t=\min I_k}^{\max I_k - 1} B_{p,t}$, *i.e.*, the variations in both reward and state transition distributions of each interval $I_k \in \mathcal{I}$ (see Theorem 3 in (Ortner et al., 2019)). In contrast, the SWUCRL2-CW algorithm and the BORL algorithm require significantly less information on the variations. Specifically, the SWUCRL2-CW algorithm does not need any additional knowledge on the variations except for B_r and B_p, *i.e.*, the variation budgets over the entire time horizon as defined in eqn. (1), to achieve its dynamic regret bound (see Theorem 1). This is similar to algorithms for the non-stationary bandit settings, which only require the access to B_r (Besbes et al., 2014). More importantly, the BORL algorithm (built upon the SWUCRL2-CW algorithm) enjoys the same dynamic regret bound even without knowing either of B_r or B_p (see Theorem 2).

There also exists some settings that are closely related to, but different than our setting (in terms of exogeneity and feedback). (Jaksch et al., 2010; Gajane et al., 2018) proposed solutions for the RL in piecewise-stationary MDPs setting. But as discussed before, simply applying their techniques to the general RL in non-stationary MDPs may result in undesirable dynamic regret bounds (see Section 6 for more details). In (Yu et al., 2009; Neu et al., 2010; Arora et al., 2012; Dick et al., 2014; Jin et al., 2019; Cardoso et al., 2019), the authors considered RL in MDPs with changing reward distributions but fixed transition distributions. The authors of (Even-Dar et al., 2005; Yu & Mannor, 2009; Neu et al., 2012; Abbasi-Yadkori et al., 2013; Rosenberg & Mansour, 2019; Li et al., 2019) considered RL in non-stationary MDPs with full information feedback.

3.3. Non-Stationary MAB

For online learning and bandit problems where there is only one state, the works by (Auer et al., 2002b; Garivier & Moulines, 2011b; Besbes et al., 2014; Keskin & Zeevi, 2016) proposed several "forgetting" strategies for different non-stationary MAB settings. More recently, the works by (Karnin & Anava, 2016; Luo et al., 2018; Cheung et al., 2019a;b; Chen et al., 2019b) designed parameter-free algorithms for non-stationary MAB problems. Another related but different setting is the Markovian bandit (Kim & Lim, 2016; Ma, 2018), in which the state of the chosen action evolve according to an independent time-invariant Markov chain while the states of the remaining actions stay unchanged. In (Zhou et al., 2020), the authors also considered the case when the states of all the actions are governed by the same (uncontrollable) Markov chain.

4. Sliding Window UCRL2 with Confidence Widening

In this section, we present the SWUCRL2-CW algorithm, which incorporates sliding window estimates (Garivier & Moulines, 2011a) and a novel confidence widening technique into UCRL2 (Jaksch et al., 2010).

4.1. Design Challenge: Failure of Naive Sliding Window UCRL2 Algorithm

For stationary MAB problems, the UCB algorithm (Auer et al., 2002a) suggests the DM should iteratively execute the following two steps in each time step:

1. Estimate the mean reward of each action by taking the time average of *all* observed samples.
2. Pick the action with the highest estimated mean reward plus the confidence radius, where the radius scales inversely proportional with the number of observations (Auer et al., 2002a).

The UCB algorithm has been proved to attain optimal regret bounds for various stationary MAB settings (Auer et al., 2002a; Kveton et al., 2015). For non-stationary problems, (Garivier & Moulines, 2011b; Keskin & Zeevi, 2016; Cheung et al., 2019b) shown that the DM could further leverage the forgetting principle by incorporating the sliding-window estimator (Garivier & Moulines, 2011b) into the UCB algorithms (Auer et al., 2002a; Kveton et al., 2015) to achieve optimal dynamic regret bounds for a wide variety of non-stationary MAB settings. The sliding window UCB algorithm with a window size $W \in \mathbb{R}_+$ is similar to the UCB algorithm except that the estimated mean rewards are computed by taking the time average of the W *most recent* observed samples.

As noted in Section 1, (Jaksch et al., 2010) proposed the UCRL2 algorithm, which is a UCB-alike algorithm with nearly optimal regret for RL in stationary MDPs. It is thus tempting to think that one could also integrate the forgetting principle into the UCRL2 algorithm to attain low dynamic regret bound for RL in non-stationary MDPs. In particular, one could easily design a naive sliding-window UCRL2 algorithm that follows exactly the same steps as the UCRL2 algorithm with the exception that it uses only the W most recent observed samples instead of all observed samples to estimate the mean rewards and the state transition distributions, and to compute the respective confidence radius.

Under non-stationarity and bandit feedback, however, we show in Proposition 3 of the forthcoming Section 6 that the diameter of the estimated MDP produced by the naive sliding-window UCRL2 algorithm with window size W can be as large as $\Theta(W)$, which is orders of magnitude larger than $D_{\max}$, the maximum diameter of each individual MDP encountered by the DM. Consequently, the naive sliding-window UCRL2 algorithm may result in undesirable dynamic regret bound. In what follows, we discuss in more details how our novel confidence widening technique can mitigate this issue.

4.2. Design Overview

The SWUCRL2-CW algorithm first specifies a sliding window parameter $W \in \mathbb{N}$ and a confidence widening parameter $\eta \geq 0$. Parameter W specifies the number of previous time steps to look at. Parameter η quantifies the amount of additional optimistic exploration, on top of the conventional optimistic exploration using upper confidence bounds. The former turns out to be necessary for handling the drifting non-stationarity of the transition kernel.

The algorithm runs in a sequence of episodes that partitions the T time steps. Episode m starts at time $\tau(m)$ (in particular $\tau(1) = 1$), and ends at the end of time $\tau(m+1) - 1$. Throughout an episode m, the DM follows a certain stationary policy $\tilde{\pi}_m$. The DM ceases the m^{th} episode if at least one of the following two criteria is met:

- The time index t is a multiple of W. Consequently, each episode last for at most W time steps. The criterion ensures that the DM switches the stationary policy $\tilde{\pi}_m$ frequently enough, in order to adapt to the non-stationarity of r_t's and p_t's.
- There exists some state-action pair (s, a) such that $\nu_m(s, a)$, the number of time step t's with $(s_t, a_t) = (s, a)$ within episode m, is at least as many as the total number of counts for it within the W time steps prior to $\tau(m)$, *i.e.*, from $(\tau(m) - W) \vee 1$ to $(\tau(m) - 1)$. This is similar to the doubling criterion in (Jaksch et al., 2010), which ensures that each episode is sufficiently long so that the DM can focus on learning.

The combined effect of these two criteria allows the DM to learn a low dynamic regret policy with historical data from an appropriately sized time window. One important piece of ingredient is the construction of the policy $\tilde{\pi}_m$, for each episode m. To allow learning under non-stationarity, the SWUCRL2-CW algorithm computes the policy $\tilde{\pi}_m$ based on the history in the W time steps previous to the current episode m, *i.e.*, from round $(\tau(m)-W)\vee 1$ to round $\tau(m)-1$. The construction of $\tilde{\pi}_m$ involves the Extended Value Iteration (EVI) (Jaksch et al., 2010), which requires the confidence regions $H_{r,\tau(m)}, H_{p,\tau(m)}(\eta)$ for rewards and transition kernels as the inputs, in addition to an precision parameter ϵ. The confidence widening parameter $\eta \geq 0$ is capable of ensuring the MDP output by the EVI has a bounded diameter most of the time.

4.3. Policy Construction

To describe SWUCRL2-CW algorithm, we define for each state action pair (s, a) and each round t in episode m,

$$N_t(s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \mathbf{1}((s_q,a_q)=(s,a)),$$
$$N_t^+(s,a) = \max\{1, N_t(s,a)\}. \tag{3}$$

4.3.1. Confidence Region for Rewards.

For each state action pair (s, a) and each time step t in episode m, we consider the empirical mean estimator

$$\hat{r}_t(s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \frac{R_q\,(s,a)\,\mathbf{1}(s_q=s,a_q=a)}{N_t^+(s,a)},$$

which serves to estimate the average reward

$$\bar{r}_t(s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \frac{r_q(s,a)\mathbf{1}(s_q=s,a_q=a)}{N_t^+(s,a)}.$$

The confidence region $H_{r,t} = \{H_{r,t}(s,a)\}_{s\in\mathcal{S},a\in\mathcal{A}_s}$ is defined as

$$H_{r,t}(s,a) = \{\dot{r}\in[0,1] : |\dot{r}-\hat{r}_t(s,a)| \le \mathsf{rad}_{r,t}(s,a)\}, \tag{4}$$

with confidence radius

$$\mathsf{rad}_{r,t}(s,a) = 2\sqrt{2\log(SAT/\delta)/N_t^+(s,a)}.$$

4.3.2. Confidence Widening for Transition Kernels.

For each state action pair s, a and each time step t in episode m, consider the empirical estimator

$$\hat{p}_t(s'|s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \frac{\mathbf{1}(s_q=s,a_q=a,s_{q+1}=s')}{N_t^+(s,a)},$$

which serves to estimate the average transition probability

$$\bar{p}_t(s'|s,a) = \sum_{q=(\tau(m)-W)\vee 1}^{t-1} \frac{p_q(s'|s,a)\mathbf{1}(s_q=s,a_q=a)}{N_t^+(s,a)}.$$

Different from the case of estimating reward, the confidence region $H_{p,t}(\eta) = \{H_{p,t}(s,a;\eta)\}_{s\in\mathcal{S},a\in\mathcal{A}_s}$ for the transition probability involves a widening parameter $\eta \ge 0$:

$$\begin{aligned} &H_{p,t}(s,a;\eta) \\ =&\{\dot{p}\in\Delta^{\mathcal{S}} : \|\dot{p}(\cdot|s,a)-\hat{p}_t(\cdot|s,a)\|_1 \le \mathsf{rad}_{p,t}(s,a)+\eta\}, \end{aligned} \tag{5}$$

with confidence radius

$$\mathsf{rad}_{p,t}(s,a) = 2\sqrt{2S\log\left(SAT/\delta\right)/N_t^+(s,a)}.$$

In a nutshell, the incorporation of $\eta > 0$ provides an additional source of optimism, and the DM can explore transition kernels that further deviate from the sample average. This turns out to be crucial for learning MDPs under drifting non-stationarity. We treat η as a hyper-parameter at the moment, and provide a suitable choice of η when we discuss our main results.

4.3.3. Extended Value Iteration (EVI) (Jaksch et al., 2010).

The SWUCRL2-CW algorithm relies on the EVI, which solves MDPs with optimistic exploration to near-optimality. We extract and rephrase a description of EVI in Section A.3 of the full version (Cheung et al., 2020b). EVI inputs the confidence regions H_r, H_p for the rewards and the transition kernels. The algorithm outputs an "optimistic MDP model", which consists of reward vector $\tilde{r}$ and transition kernel $\tilde{p}$ under which the optimal average gain $\tilde{\rho}$ is the largest among all $\dot{r}\in H_r, \dot{p}\in H_p$:

- **Input:** Confidence regions H_r for r, H_p for p, and an error parameter $\epsilon > 0$.
- **Output:** The returned policy $\tilde{\pi}$ and the auxiliary output $(\tilde{r},\tilde{p},\tilde{\rho},\tilde{\gamma})$. In the latter, $\tilde{r}$, $\tilde{p}$, and $\tilde{\rho}$ are the selected "optimistic" reward vector, transition kernel, and the corresponding long term average reward. The output $\tilde{\gamma}\in\mathbf{R}^{\mathcal{S}}_{\ge 0}$ is a *bias vector* (Jaksch et al., 2010). For each $s\in\mathcal{S}$, the quantity $\tilde{\gamma}(s)$ is indicative of the short term reward when the DM starts at state s and follows the optimal policy. By the design of EVI, for the output $\tilde{\gamma}$, there exists $s\in\mathcal{S}$ such that $\tilde{\gamma}(s)=0$. Altogether, we express

$$\text{EVI}(H_r, H_p;\epsilon) \to (\tilde{\pi},\tilde{r},\tilde{p},\tilde{\rho},\tilde{\gamma}).$$

Combining the three components, a formal description of the SWUCRL2-CW algorithm is shown in Algorithm 1.

4.4. Performance Analysis: The Blessing of More Optimism

We now analyze the performance of the SWUCRL2-CW algorithm. First, we introduce two events $\mathcal{E}_r, \mathcal{E}_p$, which state that the estimated reward and transition kernels lie in the respective confidence regions.

$$\mathcal{E}_r = \{\bar{r}_t(s,a)\in H_{r,t}(s,a)\ \forall s,a,t\},$$
$$\mathcal{E}_p = \{\bar{p}_t(\cdot|s,a)\in H_{p,t}(s,a;0)\ \forall s,a,t\}.$$

We prove that $\mathcal{E}_r, \mathcal{E}_p$ hold with high probability.

Lemma 1. *We have* $\Pr[\mathcal{E}_r]\ge 1-\delta/2$, $\Pr[\mathcal{E}_p]\ge 1-\delta/2$.

The proof is provided in Section B of the full version (Cheung et al., 2020b). In defining $\mathcal{E}_p$, the widening parameter

Algorithm 1 `SWUCRL2-CW` algorithm

```
Input: Time horizon $T$, state space $\mathcal{S}$, action space $\mathcal{A}$, window size $W$, widening parameter $\eta$.
Initialize $t \leftarrow 1$, initial state $s_1$.
for episode $m = 1, 2, \ldots$ do
    Set $\tau(m) \leftarrow t$, $\nu_m(s,a) \leftarrow 0$, and $N_{\tau(m)}(s,a)$ according to Eqn (3), for all $s, a$.
    Compute the confidence regions $H_{r,\tau(m)}$, $H_{p,\tau(m)}(\eta)$ according to Eqns (4, 5).
    Compute a $1/\sqrt{\tau(m)}$-optimal optimistic policy $\tilde{\pi}_m$ : $\mathsf{EVI}(H_{r,\tau(m)}, H_{p,\tau(m)}(\eta); 1/\sqrt{\tau(m)}) \rightarrow (\tilde{\pi}_m, \tilde{r}_m, \tilde{p}_m, \tilde{\rho}_m, \tilde{\gamma}_m)$.
    while $t$ is not a multiple of $W$ and $\nu_m(s_t, \tilde{\pi}_m(s_t)) < N^+_{\tau(m)}(s_t, \tilde{\pi}_m(s_t))$ do
        Choose action $a_t = \tilde{\pi}_m(s_t)$, observe reward $R_t(s_t, a_t)$ and the next state $s_{t+1}$.
        Update $\nu_m(s_t, a_t) \leftarrow \nu_m(s_t, a_t) + 1, t \leftarrow t + 1$.
        if $t > T$ then
            The algorithm is terminated.
        end if
    end while
end for
```

η is set to be 0, since we are only concerned with the estimation error on p. Next, we bound the dynamic regret of each time step, under certain assumptions on $H_{p,t}(\eta)$. To facilitate our discussion, we define the following variation measure for each t in an episode m:

$$\mathsf{var}_{r,t} = \sum_{q=\tau(m)-W}^{t-1} B_{r,q}, \quad \mathsf{var}_{p,t} = \sum_{q=\tau(m)-W}^{t-1} B_{p,q}.$$

Proposition 2. *Consider an episode m. Condition on events $\mathcal{E}_r, \mathcal{E}_p$, and suppose that there exists a transition kernel p satisfying two properties: (1) $\forall s \in \mathcal{S}\ \forall a \in \mathcal{A}_s$, we have $p(\cdot|s,a) \in H_{p,\tau(m)}(s,a;\eta)$, and (2) the diameter of $(\mathcal{S}, \mathcal{A}, p)$ at most D. Then, for every $t \in \{\tau(m), \ldots, \tau(m+1)-1\}$ in episode m, we have*

$$\rho^*_t - r_t(s_t, a_t) \leq \left[\sum_{s' \in \mathcal{S}} p_t(s'|s_t, a_t)\tilde{\gamma}_{\tau(m)}(s')\right] - \tilde{\gamma}_{\tau(m)}(s_t) \tag{6}$$

$$+ \frac{1}{\sqrt{\tau(m)}} + [2\mathsf{var}_{r,t} + 4D(\mathsf{var}_{p,t} + \eta)]$$

$$+ \left[2\mathsf{rad}_{r,\tau(m)}(s_t, a_t) + 4D \cdot \mathsf{rad}_{p,\tau(m)}(s,a)\right]. \tag{7}$$

The complete proof is in Section C of the full version (Cheung et al., 2020b). Unlike Lemma 1, the parameter η plays an important role in the Proposition. As η increases, the confidence region $H_{p,\tau(m)}(s,a;\eta)$ becomes larger for each s, a, and the assumed diameter D is expected to decrease. Our subsequent analysis shows that η can be suitably calibrated so that $D = O(D_{\text{max}})$. Next, we state our first main result, which provides a dynamic regret bound assuming the knowledge of B_r, B_p to set W, η:

Theorem 1. *Assuming $S > 1$, the `SWUCRL2-CW` algorithm with window size W and confidence widening parameter $\eta > 0$ satisfies the dynamic regret bound*

$$\tilde{O}\Big(B_pW/\eta + B_rW + \sqrt{SAT}/\sqrt{W}$$

$$+ D_{\textit{max}}\left[B_pW + S\sqrt{A}T/\sqrt{W} + T\eta + SAT/W + \sqrt{T}\right]\Big),$$

with probability $1 - O(\delta)$. Putting $W = W^* := 3S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}}/(B_r + B_p + 1)^{\frac{1}{2}}$ and $\eta = \eta^* := \sqrt{(B_p + 1)W^*/T}$, the bound specializes to

$$\tilde{O}\left(D_{\textit{max}}(B_r + B_p + 1)^{\frac{1}{4}}S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{3}{4}}\right). \tag{8}$$

Proof Sketch. The complete proof is presented in Section D of the full version (Cheung et al., 2020b). Proposition 2 states that if the confidence region $H_{p,\tau(m)}(\eta)$ contains a transition kernel that induces a MDP with bounded diameter D, the EVI supplied with $H_{p,\tau(m)}(\eta)$ can return a policy with controllable dynamic regret bound. However, as we show in Section 6, one in general cannot expect this to happen. Nevertheless, we bypass this with our novel confidence widening technique and a budget-aware analysis. We consider the first time step $\tau(m)$ of each episode m : if $p_{\tau(m)}(\cdot|s,a) \in H_{p,\tau(m)}(s,a;\eta)$ for all (s,a), then Proposition 2 can be leveraged; otherwise, the widened confidence region enforces that a considerable amount of variation budget is consumed. □

Remark 2. *When $S = \{s\}$, our problem becomes the non-stationary bandit problem studied by (Besbes et al., 2014), and we have $D_{\textit{max}} = 0$ and $B_p = 0$. By choosing $W = W^* = A^{1/3}T^{2/3}/B_r^{2/3}$, our algorithm has dynamic regret $\tilde{O}(B_r^{1/3}A^{1/3}T^{2/3})$, matching the minimax optimal dynamic regret bound by (Besbes et al., 2014) when $B_r \in [A^{-1}, A^{-1}T]$.*

Remark 3. *Similar to (Cheung et al., 2019b), if B_p, B_r are not known, we can set W and η obliviously as $W = S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{1}{2}}$, $\eta = \sqrt{W/T} = S^{\frac{2}{3}}A^{\frac{1}{2}}T^{-\frac{1}{2}}$ to obtain a dynamic regret bound $\tilde{O}\left(D_{\textit{max}}(B_r + B_p + 1)S^{\frac{2}{3}}A^{\frac{1}{2}}T^{\frac{3}{4}}\right)$.*

5. Bandit-over-Reinforcement Learning: Towards Parameter-Free

As said in Remark 3, in the case of unknown B_r and B_p, the dynamic regret of `SWUCRL2-CW` algorithm scales linearly in B_r and B_p, which leads to a $\Omega(T)$ dynamic regret when B_r or $B_p = \Omega(T^{1/4})$. In comparison, Theorem 1 assures us that by using (W^*, η^*), we can achieve a $o(T)$

dynamic regret when $B_r, B_p = o(T)$. For the bandit setting, (Cheung et al., 2019b) proposes the bandit-over-bandit framework that uses a separate copy of EXP3 algorithm to tune the window length. Inspired by it, we develop a novel Bandit-over-Reinforcement Learning (BORL) algorithm, which is parameter free and has dynamic regret bound equal to (8). Following (Cheung et al., 2019b), we view the SWUCRL2-CW algorithm as a sub-routine, and "hedge" (Bubeck & Cesa-Bianchi, 2012) against the (possibly adversarial) changes of r_t's and p_t's to identify a reasonable fixed window length and confidence widening parameter. As illustrated in Fig. 1, the BORL algorithm divides the whole time horizon into $\lceil T/H \rceil$ blocks of equal length H rounds (the length of the last block can $\leq H$), and specifies a set J from which each pair of (window length, confidence widening parameter) are drawn from. For each block $i \in [\lceil T/H \rceil]$, the BORL algorithm first calls some master algorithm to select a pair of (window length, confidence widening parameter) $(W_i, \eta_i) \in J$, and restarts the SWUCRL2-CW algorithm with the selected parameters as a sub-routine to choose actions for this block. Afterwards, the total reward of block i is fed back to the master, and the "posterior" of these parameters are updated accordingly.

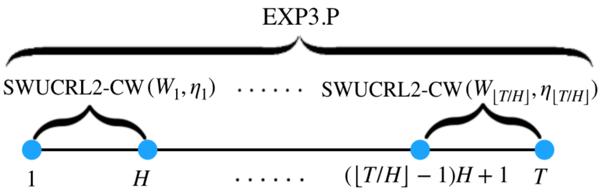
EXP3.P
SWUCRL2-CW(W_1, η_1)
......
SWUCRL2-CW$(W_{\lfloor T/H \rfloor}, \eta_{\lfloor T/H \rfloor})$
1
H
......
$(\lfloor T/H \rfloor - 1)H + 1$
T

Figure 1. Structure of the BORL algorithm

One immediate challenge not presented in the bandit setting (Cheung et al., 2019b) is that the starting state of each block is determined by previous moves of the DM. Hence, the master algorithm is not facing a simple oblivious environment as the case in (Cheung et al., 2019b), and we cannot use the EXP3 (Auer et al., 2002b) algorithm as the master. Nevertheless, the state is observed before the starting of a block. Thus, we use the EXP3.P algorithm for multi-armed bandit against an adaptive adversary (Auer et al., 2002b) as the master algorithm. Owing to its similarity to the BOB algorithm (Cheung et al., 2019b), we defer the design details and the proof of dynamic regret bound for the BORL algorithm to Sections E and F of the full version (Cheung et al., 2020b), respectively.

Theorem 2. *Assume $S > 1$, with probability $1 - O(\delta)$, the dynamic regret bound of the BORL algorithm is $\tilde{O}(D_{\max}(B_r + B_p + 1)^{\frac{1}{4}} S^{\frac{2}{3}} A^{\frac{1}{2}} T^{\frac{3}{4}})$.*

6. The Perils of Drift in Learning Markov Decision Processes

In stochastic online learning problems, one usually estimates a latent quantity by taking the time average of observed samples, even when the sample distribution varies across time. This has been proved to work well in stationary and non-stationary bandit settings (Auer et al., 2002a; Garivier & Moulines, 2011b; Cheung et al., 2019b;a). To extend to RL, it is natural to consider the sample average transition distribution $\hat{p}_t$, which uses the data in the previous W rounds to estimate the time average transition distribution $\bar{p}_t$ to within an additive error $\tilde{O}(1/\sqrt{N_t^+(s,a)})$ (see Lemma 1). In the case of stationary MDPs, where $\forall\ t \in [T]\ p_t = p$, one has $\bar{p}_t = p$. Thus, the un-widened confidence region $H_{p,t}(0)$ contains p with high probability (see Lemma 1). Consequently, the UCRL2 algorithm by (Jaksch et al., 2010), which optimistic explores $H_{p,t}(0)$, has a regret that scales linearly with the diameter of p.

The approach of optimistic exploring $H_{p,t}(0)$ is further extended to RL in *piecewise-stationary* MDPs by (Jaksch et al., 2010; Gajane et al., 2018). The latter establishes a $O(\ell^{1/3} D_{\max}^{2/3} S^{2/3} A^{1/3} T^{2/3})$ dynamic regret bounds, when there are at most ℓ changes. Their analyses involve partitioning the T-round horizon into $C \cdot T^{1/3}$ equal-length intervals, where C is a constant dependent on $D_{\max}, S, A, \ell$. At least $CT^{1/3} - \ell$ intervals enjoy stationary environments, and optimistic exploring $H_{p,t}(0)$ in these intervals yields a dynamic regret bound that scales linearly with $D_{\max}$. Bounding the dynamic regret of the remaining intervals by their lengths and tuning C yield the desired bound.

In contrast to the stationary and piecewise-stationary settings, optimistic exploration on $H_{p,t}(0)$ might lead to unfavorable dynamic regret bounds in non-stationary MDPs. In the non-stationary environment where $p_{t-W}, \ldots, p_{t-1}$ are generally distinct, we show that it is impossible to bound the diameter of $\bar{p}_t$ in terms of the maximum of the diameters of $p_{t-W}, \ldots, p_{t-1}$. More generally, we demonstrate the previous claim not only for $\bar{p}_t$, but also for every $\tilde{p} \in H_{p,t}(0)$ in the following Proposition. The Proposition showcases the unique challenge in exploring non-stationary MDPs that is absent in the piecewise-stationary MDPs, and motivates our notion of confidence widening with $\eta > 0$. To ease the notation, we put $t = W + 1$ without loss of generality.

Proposition 3. *There exists a sequence of non-stationary MDP transition distributions $p_1, \ldots, p_W$ such that 1) The diameter of $(\mathcal{S}, \mathcal{A}, p_n)$ is 1 for each $n \in [W]$. 2) The total variations in state transition distributions is $O(1)$. Nevertheless, under some deterministic policy,*

- *The empirical MDP $(\mathcal{S}, \mathcal{A}, \hat{p}_{W+1})$ has diameter $\Theta(W)$*
- *Further, for every $\tilde{p} \in H_{p,W+1}(0)$, the MDP $(\mathcal{S}, \mathcal{A}, \tilde{p})$ has diameter $\Omega(\sqrt{W/\log W})$*

Proof. The sequence $p_1, \ldots, p_W$ alternates between the following 2 instances p^1, p^2. Now, define the common state space $\mathcal{S} = \{1, 2\}$ and action collection $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2\}$, where $\mathcal{A}_1 = \{a_1, a_2\}$, $\{\mathcal{A}_2\} = \{b_1, b_2\}$. We assume all the state transitions are deterministic, and a graphical illustration is presented in Fig. 2. Clearly, we see that both instances have diameter 1.

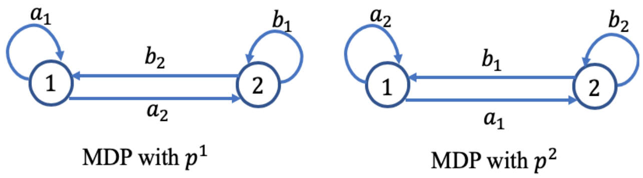

Figure 2. Example MDPs (with deterministic transitions).

Now, consider the following two deterministic and stationary policies π^1 and π^2 : $\pi^1(1) = a_1, \pi^1(2) = b_2, \pi^2(1) = a_2, \pi^2(2) = b_1$. Since the MDP is deterministic, we have $\hat{p}_{W+1} = \bar{p}_{W+1}$.

In the following, we construct a trajectory where the DM alternates between policies π^1, π^2 during time $\{1, \ldots, W\}$ while the underlying transition kernel alternates between p^1, p^2. In the construction, the DM is almost always at the self-loop at state 1 (or 2) throughout the horizon, no matter what action a_1, a_2 (or b_1, b_2) she takes. Consequently, it will trick the DM into thinking that $\hat{p}_{W+1}(1|1, a_i) \approx 1$ for each $i \in \{1, 2\}$, and likewise $\hat{p}_{W+1}(2|2, b_i) \approx 1$ for each $i \in \{1, 2\}$. Altogether, this will lead the DM to conclude that $(\mathcal{S}, \mathcal{A}, \hat{p}_{W+1})$ constitute a high diameter MDP, since the probability of transiting from state 1 to 2 (and 2 to 1) are close to 0.

The construction is detailed as follows. Let $W = 4\tau$. In addition, let the state transition kernels be p^1 from time 1 to τ and from time step $2\tau + 1$ to 3τ and be p^2 for the remaining time steps. The DM starts at state 1. She follows policy π^1 from time 1 to time 2τ, and policy π^2 from $2\tau + 1$ to 4τ. Under the specified instance and policies, it can be readily verified that the DM takes

- action a_1 from time 1 to $\tau + 1$,
- action b_2 from time $\tau + 2$ to 2τ,
- action b_1 from time $2\tau + 1$ to $3\tau + 1$,
- action a_2 from time $3\tau + 2$ to 4τ.

As a result, the DM is at state 1 from time 1 to $\tau + 1$, and time $3\tau + 2$ to 4τ; while she is at state 2 from time $\tau + 2$ to $3\tau + 1$, as depicted in Fig. 3. We have:

$$\hat{p}_{W+1}(1|1, a_1) = \frac{\tau}{\tau + 1}, \quad \hat{p}_{W+1}(2|1, a_1) = \frac{1}{\tau + 1}$$
$$\hat{p}_{W+1}(1|1, a_2) = 1, \quad \hat{p}_{W+1}(2|1, a_2) = 0$$
$$\hat{p}_{W+1}(2|2, b_1) = \frac{\tau}{\tau + 1}, \quad \hat{p}_{W+1}(1|2, b_1) = \frac{1}{\tau + 1}$$
$$\hat{p}_{W+1}(2|2, b_2) = 1, \quad \hat{p}_{W+1}(1|2, b_2) = 0.$$

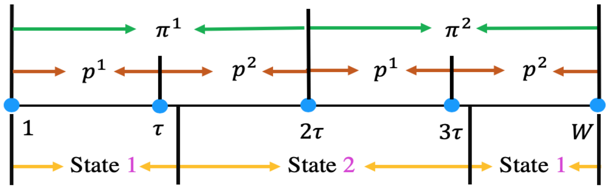

Figure 3. (From top to bottom) Underlying policies, transition kernels, time steps, and state visits.

Finally, for the confidence region $H_{p,W+1}(0) = \{H_{p,W+1}(s, a; 0)\}_{s,a}$ constructed without confidence widening, for any $\tilde{p} \in H_{p,W+1}(0)$ we have $\tilde{p}(2|1, a_1) = \tilde{p}(1|2, b_1) = O\left(\sqrt{\frac{\log W}{\tau+1}}\right)$ and $\tilde{p}(2|1, a_2) = \tilde{p}(1|2, b_2) = O\left(\sqrt{\frac{\log W}{\tau-1}}\right)$ respectively, since the stochastic confidence radii $\Theta\left(\sqrt{\frac{\log W}{\tau+1}}\right)$ and $\Theta\left(\sqrt{\frac{\log W}{\tau-1}}\right)$ dominate the sample mean $\frac{1}{\tau+1}$ and 0. Therefore, for any $\tilde{p} \in H_{p,W+1}(0)$, the diameter of the MDP constructed by $(\mathcal{S}, \mathcal{A}, \tilde{p})$ is at least $\Omega\left(\sqrt{\frac{W}{\log W}}\right)$. □

Remark 4. *Inspecting the prevalent OFU guided approach for stochastic MAB and RL in MDPs settings (Auer et al., 2002a; Abbasi-Yadkori et al., 2011; Jaksch et al., 2010; Bubeck & Cesa-Bianchi, 2012; Lattimore & Szepesvári, 2018), one usually concludes that a tighter design of confidence region can result in a lower (dynamic) regret bound. In (Abernethy et al., 2016), this insights has been formalized in stochastic K-armed bandit settings via a potential function type argument. Nevertheless, Proposition 3 (together with Theorem 1) demonstrates that using the tightest confidence region in learning algorithm design may not be enough to ensure low dynamic regret bound for RL in non-stationary MDPs.*

7. Conclusion

In this paper, we studied the problem of non-stationary reinforcement learning where the unknown reward and state transition distributions can be different from time to time as long as the total changes are bounded by some variation budgets, respectively. We first incorporated the sliding window estimator and the novel confidence widening technique into the UCRL2 algorithm to propose the `SWUCRL2-CW` algorithm with low dynamic regret when the variation budgets are known. We then designed the parameter-free `BORL` algorithm that allows us to enjoy this dynamic regret bound without knowing the variation budgets. The main ingredient of the proposed algorithms is the novel confidence widening technique, which injects extra optimism into the design of learning algorithms. This is in contrast to the widely held believe that optimistic exploration algorithms for (stationary and non-stationary) stochastic online learning settings should employ the lowest possible level of optimism.

Acknowledgements

The work of the second and the third authors was supported in part by the MIT Data Science Lab. The authors would like to express sincere gratitude to Dylan Foster, Negin Golrezaei, and Mengdi Wang, as well as various seminar attendees for helpful discussions and comments.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances Neural Information Processing Systems 25 (NIPS)*, 2011.

Abbasi-Yadkori, Y., Bartlett, P. L., Kanade, V., Seldin, Y., and Szepesvári, C. Online learning in markov decision processes with adversarially chosen transition probability distributions. In *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013.

Abernethy, J., Amin, K., and Zhu, R. Threshold bandits, with and without censored feedback. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Agrawal, S. and Jia, R. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pp. 1184–1194. Curran Associates, Inc., 2017.

Agrawal, S. and Jia, R. Learning in structured mdps with convex cost functions: Improved regret bounds for inventory management. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2019.

Arora, R., Dekel, O., and Tewari, A. Deterministic mdps with adversarial rewards and bandit feedback. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pp. 93–101, 2012.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. In *Machine learning, 47, 235–256*, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. The nonstochastic multiarmed bandit problem. In *SIAM Journal on Computing, 2002, Vol. 32, No. 1 : pp. 48–77*, 2002b.

Balseiro, S. and Gur, Y. Learning in repeated auctions with budgets: Regret minimization and equilibrium. In *Management Science*, 2019.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*, pp. 35–42, 2009.

Bertsekas, D. *Dynamic Programming and Optimal Control*. Athena Scientific, 2017.

Besbes, O., Gur, Y., and Zeevi, A. Stochastic multi-armed bandit with non-stationary rewards. In *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

Bimpikis, K., Candogan, O., and Saban, D. Spatial pricing in ride-sharing networks. In *Operations Research*, 2019.

Bubeck, S. and Cesa-Bianchi, N. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning, 2012, Vol. 5, No. 1: pp. 1–122, 2012.

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for markov decision processes. In *Mathematics of Operations Research*, volume 22, pp. 222–255, 1997.

Cai, H., Ren, K., Zhang, W., Malialis, K., Wang, J., Yu, Y., and Guo, D. Real-time bidding by reinforcement learning in display advertising. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, 2017.

Cardoso, A. R., Wang, H., and Xu, H. Large scale markov decision processes with changing rewards. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Chen, W., Shi, C., and Duenyas, I. Optimal learning algorithms for stochastic inventory systems with random capacities. In *SSRN Preprint 3287560*, 2019a.

Chen, Y., Lee, C.-W., Luo, H., and Wei, C.-Y. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. In *Proceedings of Conference on Learning Theory (COLT)*, 2019b.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Learning to optimize under non-stationarity. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019a.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Hedging the drift: Learning to optimize under non-stationarity. In *arXiv:1903.01461*, 2019b. URL https://arxiv.org/abs/1903.01461.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Non-stationary reinforcement learning: The blessing of (more) optimism. In *arXiv:1906.02922v4 [cs.LG]*, 2020a. URL https://arxiv.org/abs/1906.02922.

Cheung, W. C., Simchi-Levi, D., and Zhu, R. Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. In *arXiv:2006.14389*, 2020b. URL https://arxiv.org/abs/2006.14389.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. Introduction to algorithms. In *MIT Press*, 2009.

Dick, T., György, A., and Szepesvári, C. Online learning in markov decision processes with changing cost sequences. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Even-Dar, E., Kakade, S. M., , and Mansour, Y. Experts in a markov decision process. In *Advances in Neural Information Processing Systems 18 (NIPS)*, 2005.

Flajolet, A. and Jaillet, P. Real-time bidding with side information. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating markov decision processes. In *Advances in Neural Information Processing Systems 31*, pp. 2998–3008. Curran Associates, Inc., 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1578–1586. PMLR, 10–15 Jul 2018b.

Fruit, R., Pirotta, M., and Lazaric, A. Improved analysis of ucrl2b. In *https://rlgammazero.github.io/*, 2019.

Gajane, P., Ortner, R., and Auer, P. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. 2018.

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. In *Proceedings of Interna-*

tional Conferenc on Algorithmic Learning Theory (ALT), 2011a.

Garivier, A. and Moulines, E. On upper-confidence bound policies for switching bandit problems. In *Algorithmic Learning Theory*, pp. 174–188. Springer Berlin Heidelberg, 2011b.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Gurvich, I., Lariviere, M., and Moreno, A. Operations in the on-demand economy: Staffing services with self-scheduling capacity. In *Sharing Economy: Making Supply Meet Demand*, 2018.

Han, Y., Zhou, Z., and Weissman, T. Optimal no-regret learning in repeated first-price auctions. In *arXiv:2003.09795*, 2020.

Hoeffding, W. Probability inequalities for sums of bounded random variables. In *Journal of the American statistical association*, volume 58, pp. 13–30. Taylor & Francis Group, 1963.

Huh, W. T. and Rusmevichientong, P. A nonparametric asymptotic analysis of inventory planning with censored demand. In *Mathematics of Operations Research*, 2009.

Jaksch, T., Ortner, R., and Auer, P. Near-optimal regret bounds for reinforcement learning. In *J. Mach. Learn. Res.*, volume 11, pp. 1563–1600. JMLR.org, August 2010.

Jin, C., Jin, T., Luo, H., Sra, S., and Yu, T. Learning adversarial markov decision processes with bandit feedback and unknown transition. 2019.

Kanoria, Y. and Qian, P. Blind dynamic resource allocation in closed networks via mirror backpressure. In *arXiv:1903.02764*, 2019.

Karnin, Z. and Anava, O. Multi-armed bandits: Competing with optimal sequences. In *Advances in Neural Information Processing Systems 29 (NIPS)*, 2016.

Keskin, N. and Zeevi, A. Chasing demand: Learning and earning in a changing environments. In *Mathematics of Operations Research, 2016, 42(2), 277–307*, 2016.

Kim, M. J. and Lim, A. E. Robust multiarmed bandit problems. In *Management Science*, 2016.

Kiss, I. Z., Miller, J. C., and Simon, P. L. Mathematics of epidemics on networks. In *Springer*, 2017.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. Tight regret bounds for stochastic combinatorial semi-bandits. In *AISTATS*, 2015.

Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cambridge University Press, 2018.

Li, Y., Zhong, A., Qu, G., and Li, N. Online markov decision processes with time-varying transition probabilities and rewards. In *ICML workshop on Real-world Sequential Decision Making*, 2019.

Luo, H., Wei, C., Agarwal, A., and Langford, J. Efficient contextual bandits in non-stationary worlds. In *Proceedings of Conference on Learning Theory (COLT)*, 2018.

Ma, W. Improvements and generalizations of stochastic knapsack and markovian bandits approximation algorithms. In *Mathematics of Operations Research*, 2018.

Neu, G., Antos, A., György, A., and Szepesvári, C. Online markov decision processes under bandit feedback. In *Advances in Neural Information Processing Systems 23 (NIPS)*, pp. 1804–1812. 2010.

Neu, G., Gyorgy, A., and Szepesvari, C. The adversarial stochastic shortest path problem with unknown transition probabilities. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22, pp. 805–813. PMLR, 21–23 Apr 2012.

Nowzari, C., Preciado, V. M., and Pappas, G. J. Analysis and control of epidemics: A survey of spreading processes on complex networks. In *IEEE Control Systems Magazine*, 2016.

Ortner, R., Gajane, P., and Auer, P. Variational regret bounds for reinforcement learning. In *Proceedings of Conference on Uncertainty in Artificial Intelligence (UAI)*, 2019.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

Qin, Z. T., Tang, J., and Ye, J. Deep reinforcement learning with applications in transportation. In *Tutorial of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

Rosenberg, A. and Mansour, Y. Online convex optimization in adversarial Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5478–5486. PMLR, 2019.

Shortreed, S., Laber, E., Lizotte, D., Stroup, S., and Murphy, J. P. S. Informing sequential clinical decision-making through reinforcement learning: an empirical study. In *Machine Learning*, 2010.

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2018b.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.

Taylor, T. On-demand service platforms. In *Manufacturing & Service Operations Management*, 2018.

Tewari, A. and Bartlett, P. L. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems 20 (NIPS)*, pp. 1505–1512. 2008.

Wang, M. Randomized linear programming solves the markov decision problem in nearly-linear (sometimes sublinear) running time. In *Mathematics of Operations Research*, 2019.

Wei, C.-Y., Jafarnia-Jahromi, M., Luo, H., Sharma, H., and Jain, R. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. In *arXiv:1910.07072*, 2019.

Xu, K. and Yun, S.-Y. Reinforcement with fading memories. In *Mathematics of Operations Research*, 2020.

Yu, J. Y. and Mannor, S. Online learning in markov decision processes with arbitrarily changing rewards and transitions. In *Proceedings of the International Conference on Game Theory for Networks*, 2009.

Yu, J. Y., Mannor, S., and Shimkin, N. Markov decision processes with arbitrary reward processes. In *Mathematics of Operations Research*, volume 34, pp. 737–757, 2009.

Zhang, A. and Wang, M. Spectral state compression of markov processes. In *arXiv:1802.02920*, 2018.

Zhang, H., Chao, X., and Shi, C. Closing the gap: A learning algorithm for the lost-sales inventory system with lead times. In *Management Science*, 2018.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

Zhou, X., Chen, N., Gao, X., and Xiong, Y. Regime switching bandits. In *arXiv:2001.09390*, 2020.

Zhou, Z. and Bambos, N. Wireless communications games in fixed and random environments. In *IEEE Conference on Decision and Control (CDC)*, 2015.

Zhou, Z., Glynn, P., and Bambos, N. Repeated games for power control in wireless communications: Equilibrium and regret. In *IEEE Conference on Decision and Control (CDC)*, 2016.