Reinforcement Learning under Drift

Wang Chi Cheung
Department of Industrial Systems Engineering and Management, National University of Singapore isecwc@nus.edu.sg

David Simchi-Levi
Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, dslevi@mit.edu

Ruihao Zhu
Statistics and Data Science Center, Massachusetts Institute of Technology, Cambridge, MA 02139, rzhu@mit.edu

We propose algorithms with state-of-the-art dynamic regret bounds for un-discounted reinforcement learning under drifting non-stationarity, where both the reward functions and state transition distributions are allowed to evolve over time. Our main contributions are: 1) A tuned Sliding Window Upper-Confidence bound for Reinforcement Learning with Confidence-Widening (SWUCRL2-CW) algorithm, which attains low dynamic regret bounds against the optimal non-stationary policy in various cases. 2) The Bandit-over-Reinforcement Learning (BORL) framework that further permits us to enjoy these dynamic regret bounds in a parameter-free manner.

Key words: non-stationary reinforcement learning, Markov decision process, parameter-free algorithm

1. Introduction

Consider a discrete-time Markovian decision process (MDP) where a decision-maker (DM) interacts with a system iteratively: in each round, the DM first observes the current state of the system, and then picks an available action. Afterwards, it receives an instant random reward, and the system transits to the next state according to some state transition distribution. The reward distribution and the state transition distribution depend on the current state and the chosen action, but are independent of all the previous states and actions. The goal of the DM is to maximize its cumulative rewards under the following challenges:

- **Uncertainty:** the reward and the state transition distributions are initially unknown to the DM.
- **Non-stationarity:** the environment is non-stationary, and both of the reward distributions and the state transition distributions can evolve over time.
- **Partial feedback:** the DM can only observe the reward and state transition resulted by the current state and the chosen action in each round.

In fact, many applications, such as inventory control (Bertsekas 2017) and transportation (Zhang and Wang 2018, Qin et al. 2019), can be modeled by this general framework.

Under stationarity, this problem can be solved by the classical Upper-Confidence bound for Reinforcement Learning (UCRL2) algorithm (Jaksch et al. 2010). Unfortunately, the strategies for
the stationary setting can deteriorate in non-stationarity environments as historical data "expires". To address this shortcoming, (Jaksch et al. 2010, Gajane et al. 2018) further consider a switching MDP setting where the MDP is piecewise-stationary, and propose solutions for it. Under the special case of multi-armed bandit (MAB), where the MDP has only one state, there is a recent stream of research initiated by (Besbes et al. 2014) that studies the so-called drifting environment (Karnin and Anava 2016, Luo et al. 2018, Cheung et al. 2019, Chen et al. 2019), in which the reward of each action can change arbitrarily over time, but the total change (quantified by a suitable metric) is upper bounded by some 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.

In this paper, we generalize the concept of “drift” from MAB to RL, i.e., the reward and state transition distributions can shift gradually as long as their T rounds total changes are bounded by $B_r$ and $B_p$, respectively. We then design and analyze novel algorithms for RL in a drifting environment. Let $S$ and $A$ be the number of states and actions for the MDP, our main contributions are listed as follows.

- When the variation budgets are known, we develop a Sliding Window UCRL2 with Confidence Widening (SWUCRL2-CW) algorithm with $\tilde{O}(D_{\text{max}}(B_r + B_p + 1)^{1/3} S^{2/3} A^{1/3} T^{2/3})$ dynamic regret bound under the time-invariant variation budget assumption that is similar to, but less restive than (Yu and Mannor 2009).

- We identify a unique challenge in RL under drift: existing works for switching MDP settings (Jaksch et al. 2010, Gajane et al. 2018) estimate unknown parameters by averaging historical data in a “forgetting” fashion, and crucially exploit the piecewise stationary environment to achieve low regret. But in non-stationary settings in general, the diameter (a complexity measure to be defined in Section 3) of the MDP estimated in this manner can grow wildly, and may result in unfavorable dynamic regret bounds for drifting environments due to a lack of piecewise stationarity. This is further discussed in Section 4.4 and Section D. We overcome this with our novel confidence widening technique when the variations are uniformly bounded over time. We also show that one can bypass this difficulty for many realistic cases stemmed from inventory control or queuing systems under mild assumption.

- When the variation budgets are unknown, we propose the Bandit-over-Reinforcement Learning (BORL) framework that tunes the SWUCRL2-CW algorithm adaptively, and hence enjoys a parameter free $\tilde{O}(D_{\text{max}}(B_r + B_p + 1)^{1/3} S^{2/3} A^{1/3} T^{2/3} + D_{\text{max}} S^{1/2} A^{1/4} T^{3/4})$ dynamic regret bound.

2. Related Works

Learning un-discounted MDPs under stationarity has been studied in (Bartlett and Tewari 2009, Jaksch et al. 2010, Agrawal and Jia 2017, Fruit et al. 2018a,b) among others. For non-stationary
MDPs, the stream of works [Even-Dar et al. 2005, Nilm and Ghaoui 2005, Xu and Mannor 2006, Dick et al. 2014] consider settings with either changing reward functions or transition kernels, but not both. In contrast, [Yu and Mannor 2009] allows arbitrary changes in reward, but (globally) bounded changes in the transition kernels, and design algorithms under additional Markov chain mixing assumptions. [Jaksch et al. 2010, Gajane et al. 2018] proposes solutions for the switching setting. [Abbasi-Yadkori et al. 2013] consider learning MDPs in an adversarial environment with full information feedback.

For online learning and bandit problems where there is only one state, the works by [Auer et al. 2002, Garivier and Moulines 2011, Besbes et al. 2014, Keskin and Zeevi 2016] propose several “forgetting” strategies for different settings. More recently, the works by [Jadbabaie et al. 2015, Karnin and Anava 2016, Luo et al. 2018, Cheung et al. 2019, Chen et al. 2019] design parameter-free algorithms for non-stationary online learning.

3. Problem Formulation

An instance of non-stationary online MDP is specified by the tuple \((S, A, T, r, p)\). The set \(S\) is a finite set of states. The collection \(A = \{A_s\}_{s \in S}\) contains a finite action set \(A_s\) for each state \(s \in S\). The quantity \(T\) is the total number of time steps. The quantity \(r = \{r_t\}_{t=1}^T\) is a sequence of mean rewards. For each \(t\), we have \(r_t = \{r_t(s, a)\}_{s \in S, a \in A_s}\), where \(r_t(s, a) \in [0, 1]\) for each \(s, a\). The quantity \(p = \{p_t\}_{t=1}^T\) is a sequence of transition kernels. For each \(t\), we have \(p_t = \{p_t(\cdot|s, a)\}_{s \in S, a \in A_s}\), where \(p_t(\cdot|s, a)\) is a probability distribution over \(S\) for each \(s, a\). The quantities \(p_t, r_t\) vary across different \(t\)’s in general. We elaborate on their temporal variations in Section 4 when we provide the performance guarantee of our algorithm according to the degree of non-stationarity.

**Dynamics.** We consider a DM who faces an online non-stationary MDP instance \((S, A, T, p, r)\). The DM knows \(S, A, T\), but not \(p, r\). It starts at an arbitrary state \(s_1 \in S\). At time \(t\), three events happen. First, the DM observes its current state \(s_t\). Second, it takes an action \(a_t \in A_{s_t}\). Third, given \(s_t, a_t\), it stochastically transits to another state \(s_{t+1}\) which is distributed as \(p_t(\cdot|s_t, a_t)\), and received 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 \(\Pi\). That is, the choice only depends on the current state \(s_t\) and the previous observations \(H_{t-1} := \{s_q, a_q, R_q(s_q, a_q)\}_{q=1}^{t-1}\).

**Objective.** The DM aims to maximize the cumulative expected reward \(E[\sum_{t=1}^T r_t(s_t, a_t)]\), despite the model uncertainty on \(p_t, r_t\)’s and the non-stationarity of the learning environment. To measure the convergence of an online algorithm to optimality, we consider an equivalent objective of **dynamic regret minimization**. For each time \(t\), let \(\rho_t^*\) be the optimum long term reward for the online MDP problem with stationary transition kernel \(p_t\) and stationary mean reward function \(r_t\). To compute
\( \rho^* \), we refer to the linear programming formulation \([18]\) in Section \( A \). The DM aims to design policy \( \Pi \) to minimize the dynamic regret

\[
\text{Dyn-Reg}_T(\Pi) = \sum_{t=1}^{T} \{ \rho^*_t - E[r_t(s_t, a_t)] \}.
\]

**Remark 1.** We note that this regret notion recovers the dynamic regret definition in bandit setting when specified to bandits. Different from bandit, the expected total reward received could be much higher than the benchmark as the latter does not take into account the starting state.

We also review the relevant concepts of communicating MDPs and their diameters.

**Definition 1 (Hitting Time, Communicating MDPs and Diameters).** Consider a set of states \( \mathcal{S} \), a collection \( \mathcal{A} = \{A_s\}_{s \in \mathcal{S}} \) of action sets, and a transition kernel \( \bar{p} = \{\bar{p}(\cdot|s, a)\}_{s \in \mathcal{S}, a \in A_s} \).

For any \( s, s' \in \mathcal{S} \) and any stationary policy \( \pi \), the hitting time from \( s \) to \( s' \) under \( \pi \) is the random variable

\[
\Lambda(s'|\pi, s) := \min \{ t : s_{t+1} = s', s_1 = s, s_{t+1} \sim \bar{p}(.|s_{t+1}, \pi(s_{t+1})) \forall \tau \},
\]

which can be infinite. We say that \( (\mathcal{S}, \mathcal{A}, \bar{p}) \) constitutes a *communicating MDP* if and only if

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

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

To enable learning in a non-stationary environment, we make the following reachability assumption.

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

### 4. Sliding Window UCRL2 with Confidence Widening

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

#### 4.1. Design Overview

The SWUCRL2-CW algorithm first specifies a window length \( W \), and runs in a sequence of episodes that partitions the \( T \) time steps. Episode \( m \) starts at round \( \tau(m) \) (in particular \( \tau(1) = 1 \)), and ends at the end of round \( \tau(m+1) - 1 \). Throughout an episode \( m \), the DM follows a certain stationary policy \( \tilde{\pi}_m \). It ceases the \( m^{th} \) episode if at least one of the following two criteria is met:

- The round index \( t \) is a multiple of \( W \). This ensures that each episode is at most \( W \) time steps long, and prevents choosing actions based on out-of-date information
There exists some (state,action) pair \((s,a)\) such that \(\nu_m(s,a)\), the number of visits to them within episode \(m\), is at least as many as the total number of visits to them within the \(W\) rounds prior to \(\tau(m)\), i.e., rounds \((\tau(m) - W) \lor 1, \ldots, (\tau(m) - 1)\). 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 near-optimal policy with historical data from an appropriately sized time window. One important piece of ingredient is the construction of the policy for each episode. To allow learning under non-stationarity, the \textsc{SWUCRL2-CW} algorithm computes the policy \(\tilde{\pi}_m\) for the \(m\)th episode based on the history in the previous \(W\) rounds before the current episode \(m\), i.e., between round \((\tau(m) - W) \lor 1\) and \(\tau(m) - 1\). The construction of \(\tilde{\pi}_m\) involves the Extended Value Iteration (EVI) (Jaksch et al. 2010), which requires the confidence/uncertainty regions \(H_{r,\tau(m)}, H_{p,\tau(m)}(\eta)\) for rewards and transition kernels as the inputs, in addition to an error parameter \(\epsilon\). We note that the parameter \(\eta \geq 0\) is a certain confidence widening parameter, which contributes to ensure the output MDP of the EVI has a bounded diameter.

### 4.2. Policy Construction

For ease of discussion, we first define for each state-action pair \(s,a\) and each \(t\) belongs to episode \(m\),

\[
N_t(s,a) := \sum_{q=(\tau(m) - W) \lor 1}^{t-1} 1((s_q,a_q) = (s,a)), \quad N^+_t(s,a) := \max\{1, N_t(s,a)\}. \tag{1}
\]

1) **Confidence Region for Rewards.** For each state-action pair \(s,a\) and each \(t\) belongs to episode \(m\), we consider the empirical mean estimator

\[
\hat{\tilde{r}}_t(s,a) := \frac{1}{N^+_t(s,a)} \left( \sum_{q=(\tau(m) - W) \lor 1}^{t-1} R_q(s_q,a_q) 1(s_q = s, a_q = a) \right),
\]

which has mean

\[
\tilde{r}_t(s,a) := \frac{1}{N^+_t(s,a)} \left( \sum_{q=(\tau(m) - W) \lor 1}^{t-1} r_q(s,a) 1(s_q = s, a_q = a) \right).
\]

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

\[
H_{r,t}(s,a) := \{\hat{r} \in [0,1] : |\hat{r} - \tilde{r}_t(s,a)| \leq \text{rad-}r_t(s,a)\}, \tag{2}
\]

where \(\text{rad-}r_t(s,a) := \sqrt{2\log(4SAT/\delta)/N^+_t(s,a)}\).
Algorithm 1 SWUCRL2-CW algorithm

1: **Input**: Time horizon $T$, state space $S$, and action space $A$, window size $W$, confidence widening parameter $\eta \in [0, 1]$, initial state $s_1$.

2: Initialize $t \leftarrow 1$.

3: for episode $m = 1, 2, \ldots$ do

4: Set $\tau(m) \leftarrow t$, $\nu_m(s, a) \leftarrow 0$, and $N_{\tau(m)}(s, a)$ according to Eqn (1), for all $s, a$.

5: Compute the confidence regions $H_{r, \tau(m)}$, $H_{p, \tau(m)}(\eta)$ according to Eqns (2, 5).

6: Compute a $(1/\sqrt{\tau(m)})$-optimal optimistic policy $\tilde{\pi}_m$:

$$
\tilde{\pi}_m \leftarrow \text{EVI}(H_{r, \tau(m)}, H_{p, \tau(m)}(\eta); 1/\sqrt{\tau(m)}).
$$

7: while $t$ is not a multiple of $W$ and $\nu_m(s_t, \tilde{\pi}_m(s_t)) < N_{\tau(m)}^{p}(s_t, \tilde{\pi}_m(s_t))$ do

8: Choose action $a_t = \tilde{\pi}_m(s_t)$, observe reward $R_t(s_t, a_t)$ and the next state $s_{t+1}$.

9: Update $\nu_m(s_t, a_t) \leftarrow \nu_m(s_t, a_t) + 1$, $t \leftarrow t + 1$.

10: if $t > T$ then

11: The algorithm is terminated.

12: end if

13: end while

14: end for

2) Confidence Region and Confidence Widening for Transition Kernels. For each $s \in S, a \in A, s' \in S$ and each time $t$ in episode $m$, we also consider the empirical mean estimator

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

which has mean

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

The confidence region is $H_{p, \tau}(\eta_p) = \{H_{p, \tau}(s, a; \eta_p) \}_{s \in S, a \in A}$ with

$$
H_{p, \tau}(s, a; \eta_p) := \{ \hat{p} \in \Delta^S : ||\hat{p}(\cdot) - \hat{p}_t(\cdot|s, a)||_1 \leq \text{rad-p}_t(s, a) + \eta \},
$$

where $\text{rad-p}_t(s, a) = 2\sqrt{2S\log(8ATW/d)/N_t^+(s, a)}$.

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 Appendix A.1. EVI inputs the confidence regions $H^r, H^p(\eta)$ for the rewards and the transition kernels. It then outputs an “optimistic MDP model”, which consists
of reward vector and transition kernel under which the optimal average gain is the largest among all reward vectors and transition kernels in the supplied confidence regions. Specifically, it works as follows:

- **Input:** confidence regions $H^r$ for $r$, $H^p$ for $p$, and an arbitrarily small error parameter $\epsilon > 0$.
- **Output:** The returned policy $\tilde{\pi}$ and the auxiliary output $(\tilde{r}, \tilde{p}, \tilde{\rho}, \tilde{\gamma})$. Here, $\tilde{r}$, $\tilde{p}$, and $\tilde{\rho}$ are the selected “optimistic” reward vector, transition kernel, and the corresponding long term average reward; While $\tilde{\gamma} \in \mathbb{R}^S_{\geq 0}$ is the bias vector (Jaksch et al. 2010). Collectively, we express $\text{EVI}(H^r, H^p, \epsilon) = (\tilde{r}, \tilde{p}, \tilde{\rho}, \tilde{\gamma}, \tilde{\pi})$. Note that $\tilde{\gamma} \geq 0$ but $\tilde{\gamma}(s) = 0$ for some $s$.

The output of the EVI further satisfy the following two properties.

**Property 1** The dual variables $(\tilde{\rho}, \tilde{\gamma})$ is optimistic, i.e.,

$$\tilde{\rho} + \tilde{\gamma}(s) \geq \max_{(s,a) \in H^r(s,a)} \{\tilde{r}(s,a)\} + \sum_{s' \in S} \gamma(s') \max_{\tilde{p} \in H^p(s,a)} \{\tilde{p}(s'|s,a)\}.$$  

**Property 2** For each state $s \in S$, we have

$$\tilde{r}(s, \tilde{\pi}(s)) \geq \tilde{\rho} + \tilde{\gamma}(s) - \sum_{s' \in S} \tilde{p}(s'|s, \tilde{\pi}(s)) \tilde{\gamma}(s') - \epsilon.$$  

Consequently, $\tilde{\pi}$ is an optimal (up to an additive factor of $\epsilon$) deterministic policy for MDP $(S, A, \tilde{p}, \tilde{r})$.

With these, the formal description of the SWUCRL2-CW algorithm is shown in Algorithm 1.

### 4.3. Performance Analysis

We are now ready to analyze the performance of the SWUCRL2-CW algorithm. As we are under a changing environment, the dynamic regret bound of SWUCRL2-CW algorithm shall adapt to the underlying shifts (Besbes et al. 2014). We thus define the following variation measures $B_r, B_p$ for the drifts in rewards and transition kernels across time:

$$B_r = \sum_{t=1}^{T-1} B_{r,t}, \quad B_{r,t} \geq \max_{s \in S, a \in A} \{r_{t+1}(s,a) - r_t(s,a)\},$$  

$$B_p = \sum_{t=1}^{T-1} B_{p,t}, \quad B_{p,t} \geq \max_{q \in [T-1], s \in S, a \in A} \{||p_{q+1}(\cdot|s,a) - p_q(\cdot|s,a)||_1\}.$$  

(6)  

(7)

One can interpret the quantities $B_{r,t}, B_{p,t}$ as an upper bound on the variations in rewards and transition kernels between round $t$ and $t + 1$, and the quantities $B_r, B_p$ are the total allowable variations for rewards and transition kernels in $T$ rounds. Similar to (but less restrictive than) (Yu and Mannor 2009), we further make the following assumption to imposes the transition kernel changes slowly at a steady rate across time.
Assumption 2. The transition kernel’s variation budgets are uniform over time, i.e.,

\[ B_{p,1} = \ldots = B_{p,T-1}. \]

Together with the confidence widening technique, we are guaranteed the resulted MDP of the EVI has a bounded diameter. For ease of exposition, we also define the piecewise variations for each \( t \) belongs to episode \( m \),

\[
\text{var}_{r,t} := \sum_{q=\tau(m)-W}^{t-1} B_{r,q}, \quad \text{var}_{p,t} := B_{p,1} \left( t - \tau(m) + \frac{W}{2} \right).
\]

To proceed, we introduce two events \( \mathcal{E}_r, \mathcal{E}_p \), which state that the estimated reward and transition kernels lie in the confidence region.

\[
\mathcal{E}_r := \{ \bar{r}_t(s,a) \in H_{r,t}(s,a) \ \forall s,a,t \}, \quad \mathcal{E}_p := \{ \bar{p}_t(s'|s,a) \in H_{p,t}(s,a;0) \ \forall s,a,t \}.
\]

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

Lemma 1. We have \( \Pr[\mathcal{E}_r] \geq 1 - \delta/2 \), \( \Pr[\mathcal{E}_p] \geq 1 - \delta/2 \).

The proof is in Section \([A]\) of the appendix. We then bound the dynamic regret of each round.

Proposition 1. Conditioning on events \( \mathcal{E}_r, \mathcal{E}_p \) and assuming that \( \eta \geq B_p W/2T \), for every episode \( m \) and every time \( t \) in the episode (i.e. \( t \in \{ \tau(m), \ldots, \tau(m+1) - 1 \} \)) the following inequality holds for certainty:

\[
\rho^*_t - r_t(s_t,a_t) \leq \sum_{s' \in S} p_t(s'|s_t,a_t) \tilde{\gamma}_{\tau(m)}(s') - \tilde{\gamma}_{\tau(m)}(s_t) + \frac{1}{\sqrt{\tau(m)}} \left[ 2\text{var}_{r,t} + 4D_{\tau(m)}(\text{var}_{p,t} + \eta) \right] + \left[ 2\text{rad}_{\tau(m)}(s_t,a_t) + 4D_{\tau(m)} \cdot \text{rad}_{p_{\tau(m)}}(s,a) \right] .
\]

Proof Sketch. The complete proof can be found in Section \([B]\) of the appendix. Consider time \( t \), which belongs episode \( m \). In a high level, the proof goes through three steps:

1. The estimated mean reward \( \tilde{r}_{\tau(m)}(s,a) \) (used in computing \( \tilde{\pi}_m \)) and the true reward \( r_t(s,a) \) differ by at most \( \text{var}_{r,t} \).

2. With the widened confidence regions \( H_{p,\tau(m)}(\eta) \), we are guaranteed that there exists a MDP, i.e., \( (S,A,p_{\tau(m)}) \), with diameter at most \( D_{\tau(m)} \) in \( H_{p,\tau(m)}(\eta) \). By Lemma 2 and the dual formulation of optimal reward \([19]\), we have the optimistic long term reward \( \tilde{\rho}_{\tau(m)} \) induced by \( \tilde{\pi}_m \) is at least \( \rho^*_t - \text{var}_{r,t} - 2D_{\tau(m)} \text{var}_{p,t} \).

3. Finally, we relate the optimistic transition kernel \( \tilde{p}_{\tau(m)} \) and the underlying kernel \( p_t \) to account for the extra loss due to a widened confidence region. Unifying all three steps, we can conclude the statement.
Combining the above, we can conclude the statement. □

Suppose we denote \( S := |S|, A := \frac{1}{3} \sum_{s \in S} |A_s| \), and use the \( \tilde{O}(\cdot) \) notation to hide logarithmic dependence on \( D_{\text{max}}, S, A, T \) and the confidence parameter \( \delta \) for defining the confidence regions \( H_{r,t}, H_{p,t}(\eta) \), our first main result is a dynamic regret upper bound on the SWUCRL2-CW algorithm.

**Theorem 1.** Assuming \( S > 1 \), the SWUCRL2-CW algorithm with window size \( W \) and \( \eta \geq B_p W/(2T) \) satisfies the dynamic regret bound

\[
\text{Dyn-Reg}_T(\text{SWUCRL2-CW}) = \tilde{O} \left( B_r W + D_{\text{max}} \left[ B_p W + \frac{S \sqrt{AT}}{\sqrt{W}} + T\eta + \frac{SAT}{W} + \sqrt{T} \right] \right)
\]

(8)

with probability \( 1 - O(\delta) \). If we further put

\[
W = W^* = S^\frac{2}{3} A^\frac{1}{3} T^\frac{2}{3} / (B_r + B_p + 1)^{\frac{2}{3}},
\]

(9)

\[
\eta = \eta^* = \frac{(B_p + 1)W}{2T} = \frac{S^\frac{2}{3} A^\frac{1}{3} (B_p + 1)}{2T^\frac{2}{3} (B_r + B_p + 1)^{\frac{2}{3}}},
\]

(10)

this is

\[
\tilde{O} \left( D_{\text{max}} (B_r + B_p + 1)^{\frac{2}{3}} S^\frac{2}{3} A^\frac{1}{3} T^\frac{2}{3} \right).
\]

The complete proof of Theorem 1 is presented in Section C of the appendix.

**Remark 2.** When \( S = 1 \), our problem model specializes to the non-stationary bandit problem studied by (Besbes et al. 2014). In this case, we have \( D_{\text{max}} = 0 \), and we are left with the first term in (8). By choosing \( W = W^* = A^{1/3} T^{2/3} / \max\{B_r, 1\}^{2/3} \), 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).

**Remark 3.** Different from (Cheung et al. 2019), there is no straightforward way of setting \( W, \eta \) to get a non-trivial bound when \( B_p, B_r \) are not known. While (Cheung et al. 2019) provide a way to set their window size (oblivious of their variational budget \( B_r \)) so that the dynamic regret is \( \tilde{O}(B_r T^{2/3}) \) in their linear bandit setting, in our setting we still need to have \( \eta \geq B_p W/(2T) \), which is a prior not clear how to ensure when we don’t know \( B_p \).

### 4.4. Uniform Variation Budget, Confidence Widening, and Alternatives

We now pause for a while to comment on Assumption 2 and the technique of confidence widening.

In online stochastic environment, one usually take time average of observed samples to estimate a certain latent quantity, even when the sample distributions vary with time. This has been proved to work well in the non-stationary bandit settings (Garivier and Moulines 2011, Cheung et al. 2019). For online MDPs, one typically look at the time average MDP \( \hat{p}_t \) in (3), which estimate \( \hat{p}_t \) to within an additive error \( \tilde{O}(1/\sqrt{N_t(s,a)}) \) for any pair of \((s,a)\). In the case of stationary MDPs where \( \forall t \in [T] \ p_t = p \), one has \( \hat{p}_t = p \), and thus can conclude that the un-widened confidence region \( H_{p,t}(0) \) contains \( p \) with high probability. An immediate consequence is that the EVI w.r.t. \( H_{p,t}(0) \)
would return a policy with bounded difference bias vector \( \tilde{\gamma} \) as \( p \) has diameter \( D \) (Please see Section 4.3 of (Jaksch et al. 2010)). These further ensure that the optimistic long term average reward is not far away from the true long term average reward, e.g., step 2) in the proof of Proposition 1. Nevertheless, this is not always the case under non-stationarity. Although (Jaksch et al. 2010, Gajane et al. 2018) uses \( H_{p,t}(0) \) for piecewise stationary MDP setting, they crucially exploit the fact that the MDP remain unchanged between jumps, and can treat the problem as if it is stationary. For changing environments, one can only guarantee that the transition kernel \( \bar{p}_t = \{ \bar{p}_t(\cdot | s,a) \}_{s \in S, a \in A} \in H_{p,t}(0) \), with high probability, but unsure about the true \( p_t \)'s due to the drift. In Section D.1 of the appendix, we show that \( D' \), the diameter of \( (S, A, \bar{p}_t) \), can grow as \( \Omega(\sqrt{W/\log W}) \), and the EVI w.r.t. \( H_{p,t}(0) \) can only promise a policy with bias vector bounded by \( \Omega(\sqrt{W/\log W}) \), which makes the dynamic regret bound vacuous for drifting environments. By assumption 2 and the confidence widening technique, we are guaranteed that \( p_{\tau(m)} \in H_{p,\tau(m)}(\eta) \) for each episode \( m \), and can proceed as what we have done.

Alternatively, if there exists \( \zeta > 0 \) such that for any states \( s, s' \in S \), there is always an action \( a_{s,s'} \in A_s \) with \( p_t(s'|s,a_{s,s'}) \geq \zeta \forall t \in [T] \), then the MDP \( (S, A, \bar{p}_t) \) has diameter \( \leq D := 1/\zeta \), and it can shown that SWUCRL2-CW algorithm with \( \eta = 0 \) enjoys a dynamic regret bound \( \tilde{O}\left( \tilde{D}(B_r + B_p + 1)^{1/3}S^{2/3}A^{1/3}T^{2/3} \right) \) by similar techniques. As we shall shown in Section D.2 this assumption can be easily satisfied in many realistic applications.

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

As pointed out by Remark 3 in the case of unknown \( B_r \) and \( B_p \), the DM cannot implement the SWUCRL2-CW algorithm as the magnitude of confidence widening cannot be determined. To handle this case, we wish to design an online algorithm that can attain reasonable dynamic regret bound in a parameter-free manner. By Theorem 1 we are assured: under Assumption 2 a fixed pair of parameters \( (W^*, \eta^*) \) can ensure low regret. For the bandit setting, (Cheung et al. 2019) 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 with parameter-free \( \tilde{O}\left( D_{\text{max}}(B_r + B_p + 1)^{1/3}S^{2/3}A^{1/3}T^{2/3} + D_{\text{max}}S^{2/3}A^{1/3}T^{3/4} \right) \) dynamic regret here.

5.1. Design Overview

Following a similar line of reasoning as (Cheung et al. 2019), we make use of the SWUCRL2-CW algorithm as a sub-routine, and “hedge” (Bubeck and Cesa-Bianchi 2012) against the (possibly adversarial) changes of \( r_i \)'s and \( p_i \)'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 \([T/H]\) 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) parameters \((W_i, \eta_i) \in J\), 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.

One immediate challenge not presented in the bandit setting (Cheung et al. 2019) 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 bandit setting where there is only one state, but fortunately, the state is observed before the starting of a block. To this end, we use the EXP3.P algorithm for multi-armed bandit against an adaptive adversary (Auer et al. 2002, Bubeck and Cesa-Bianchi 2012) as the master algorithm.

5.2. Design Details

We are now ready to state the details of the BORL algorithm. For some fixed choice of block length \(H\) (to be determined later), we first define a couple of additional notations:

\[
\Phi = \frac{S^{2/3}A^{1/3}}{2T^{1/3}}, \Delta_W = \lfloor \ln H \rfloor, \Delta_\eta = \lfloor \ln \Phi^{-1} \rfloor, \Delta = (\Delta_W + 1)(\Delta_\eta + 1),
\]

\[
J_W = \left\{ H^0, \left\lfloor \frac{H}{\Phi_\eta^W} \right\rfloor, \ldots, H \right\}, J_\eta = \left\{ \Phi^0, \Phi_\eta^{x_W}, \ldots, \Phi \right\}, J = \{(W, \eta) : W \in J_W, \eta \in J_\eta \}.
\]

Here, \(J_W\) and \(J_\eta\) are all possible choices of window length and confidence widening parameter, respectively, and \(J\) is the Cartesian product of them with \(|J| = \Delta\). We emphasize that due to the restarting, any instance of the SWUCRL2-CW algorithm cannot last for more than \(H\) rounds. Consequently, even if the EXP3.P selects a window length \(W > H\), the effective window length is \(H\), and we make \(J \subseteq [H]\). We also let \(R(W, \eta, s)\) be the total rewards for running the SWUCRL2-CW algorithm with window length \(W\) and confidence widening parameter \(\eta\) for \(H\) rounds starting from state \(s\),

\[
\text{Figure 1 Structure of the BORL algorithm}
\]
The EXP3.P algorithm \cite{BubeckC12} treats each element of \( J \) as an arm. It begins by initializing
\[
\alpha = 0.95 \sqrt{\frac{\ln \Delta}{\Delta [T/H]}}, \beta = \sqrt{\frac{\ln \Delta}{\Delta [T/H]}}, \gamma = 1.05 \sqrt{\frac{\Delta \ln \Delta}{[T/H]}}, q_{(j,k),1} = 0 \quad \forall (j,k) \in M, \tag{12}
\]
where \( M = \{(j',k') : j' \in \{0,1,\ldots,\Delta_W\}, k' \in \{0,1,\ldots,\Delta_\eta\}\}. \) At the beginning of each block \( i \in [[T/H]] \), the BORL algorithm first sees the state \( s_{(i-1)H+1} \), and computes
\[
\forall (j,k) \in M, \quad u_{(j,k),i} = (1-\gamma) \frac{\exp(\alpha q_{(j,k),i})}{\sum_{(j',k') \in M} \exp(\alpha q_{(j',k'),i})} + \frac{\gamma}{\Delta}. \tag{13}
\]
Then it sets \((j_i,k_i) = (j,k)\) with probability \( u_{(j,k),i} \), \( \forall (j,k) \in M \). The selected pair of parameters are thus \( W_i \in \left[ \frac{H^{j_i}/\Delta_W}{1} \right] \) and \( \eta_i \in \left[ \frac{\Phi^{k_i}/\Delta_\eta}{1} \right] \). Afterwards, the BORL algorithm starts from state \( s_{(i-1)H+1} \), selects actions by running the SWUCRL2-CW algorithm with window length \( W_i \) and confidence widening parameter \( \eta_i \) for each round \( t \) in block \( i \). At the end of the block, the BORL algorithm observes the total rewards \( R(W_i,\eta_i,s_{(i-1)H+1}) \). As a last step, it rescales \( R(W_i,\eta_i,s_{(i-1)H+1}) \) by dividing it by \( H \) so that it is within \([0,1]\), and updates
\[
\forall (j,k) \in M \quad q_{(j,k),i+1} = q_{(j,k),i} + \beta + \mathbf{1}_{(j,k) = (j_i,k_i)} \cdot R(W_i,\eta_i,s_{(i-1)H+1}) / H, \tag{14}
\]
The formal description of the BORL algorithm (with \( H \) defined in the next subsection) is shown in Algorithm \ref{alg:BORL}.

5.3. Performance Analysis
To analyze the performance of the BORL algorithm, we consider the following regret decomposition, for any choice of \((W^\dagger,\eta^\dagger) \in J\), we have
\[
\text{Dyn-Reg}(\text{BORL},T) = \sum_{i=1}^{[T/H]} \mathbb{E} \left[ \sum_{t=(i-1)H+1}^{iH/T} \left( \rho_t^* - R(W_t,\eta_i,s_{(i-1)H+1}) \right) \right]
= \sum_{i=1}^{[T/H]} \mathbb{E} \left[ \sum_{t=(i-1)H+1}^{iH/T} \left( \rho_t^* - R(W^\dagger,\eta^\dagger,s_{(i-1)H+1}) \right) \right]
+ \sum_{i=1}^{[T/H]} \mathbb{E} \left[ \sum_{i=1}^{[T/H]} \left( R(W^\dagger,\eta^\dagger,s_{(i-1)H+1}) - R(W_t,\eta_i,s_{(i-1)H+1}) \right) \right]. \tag{15}
\]
For the first term in eq. (15), we can apply the results from Theorem \ref{thm:swucrl} to each block \( i \in [T/H] \), \( i.e., \)
\[
\sum_{t=(i-1)H+1}^{iH/T} \left( \rho_t^* - R(W^\dagger,\eta^\dagger,s_{(i-1)H+1}) \right) = \tilde{O} \left( B_r(i) W^\dagger + D_{\max} \left[ B_p(i) W^\dagger + S \sqrt{A H / \sqrt{W^\dagger + H} \eta^\dagger + S A H / W^\dagger + \sqrt{H}} \right] \right), \tag{16}
\]
Algorithm 2 BORL algorithm

1: **Input:** Time horizon $T$, state space $S$, and action space $A$, initial state $s_1$.

2: Initialize $H \leftarrow S \sqrt{AT}$, $\Phi, \Delta_w, \Delta_q, \Delta, J_w, J_q$ according to eq. (11), and $\alpha, \beta, \gamma$ according to eq. (12).

3: $M \leftarrow \{(j', k') : j' \in \{0, 1, \ldots, \Delta_w\}, k' \in \{0, 1, \ldots, \Delta_q\}\}, q_{(j,k),1} \leftarrow 0 \forall (j, k) \in M.$

4: for $i = 1, 2, \ldots, \ceil{T/H}$ do

5: Define distribution $(u_{(j,k),i})_{(j,k) \in M}$ according to eq. (13), and set $(j_i, k_i) \leftarrow (j, k)$ with probability $u_{(j,k),i}$.

6: $W_i \leftarrow \floor{H j_i/\Delta_w}, \eta_i \leftarrow \floor{\Phi k_i/\Delta_q}$.

7: for $t = (i - 1)H + 1, \ldots, i \cdot H \land T$ do

8: Run the SWUCRL2-CW algorithm with window length $W_i$ and blow up parameter $\eta_i$, and observe the total rewards $R(W_i, \eta_i, s_{(i-1)H+1})$.

9: end for

10: Update $q_{(j,k),i+1}$ according to eq. (14).

11: end for

where we have defined

$$B_r(i) := \left( \sum_{t=(i-1)H+1}^{i \cdot H + T} B_{r,t} \right), \quad B_p(i) = \left( \sum_{t=(i-1)H+1}^{i \cdot H + T} B_{p,t} \right)$$

for brevity. For the second term, it captures the additional rewards of the DM were it uses the fixed parameters $(W^\dagger, \eta^\dagger)$ throughout w.r.t. the trajectory on the starting states of each block by the BORL algorithm, i.e., $s_1, \ldots, s_{(i-1)H+1}, \ldots, s_{(\ceil{T/H}-1)H+1}$, and this is exactly the regret of the EXP3.P algorithm when it is applied to a $\Delta$-arm adaptive adversarial bandit problem with reward from $[0, H]$. Therefore, for any choice of $(W^\dagger, \eta^\dagger)$, we can upper bound this by

$$\tilde{O}\left(H \sqrt{\Delta T/H} \right) = \tilde{O}\left(\sqrt{TH} \right)$$

as $\Delta = O\left(\ln^2 T \right).$ Summing these two, the regret of the BORL algorithm is

$$\text{Dyn-Reg}_T(\text{BORL}) = \tilde{O}\left(B_r W^\dagger + D_{\max} \left[ B_p W^\dagger + \frac{S \sqrt{AT}}{\sqrt{W^\dagger}} + T \eta^\dagger + \frac{S A T}{W^\dagger} + \sqrt{TH} \right] \right).$$

(17)

We now point out a key trade-off in the choice of $H$:

- On one hand, $H$ should be small enough so that the regret bound in eq. (17) is small.
- On the others, $H$ should be large to allow $W^\dagger$ to get close to $W^*$ even when $B_r + B_p$ is small.

To this end, we pick

$$H = \ceil{S \sqrt{AT}}.$$
We also justify the choice of $H$, and formalize the dynamic regret bound of the BORL algorithm as follows.

**Theorem 2.** Assume that $S > 1$, the dynamic regret bound of the BORL algorithm is

$$\bar{O} \left( D_{\max} (B_r + B_p + 1)^{\frac{1}{2}} S^{\frac{3}{2}} A^{\frac{1}{2}} T^{\frac{3}{2}} + D_{\max}^{\frac{1}{2}} A^{\frac{1}{2}} T^{\frac{3}{2}} \right)$$

with probability $1 - O(\delta)$.

The complete proof can be found in Section E of the Appendix.
References

Abbasi-Yadkori, Yasin, Peter L Bartlett, Varun Kanade, Yevgeny Seldin, Csaba Szepesvári. 2013. Online learning in markov decision processes with adversarially chosen transition probability distributions. *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS).*

Agrawal, Shipra, Randy Jia. 2017. Optimistic posterior sampling for reinforcement learning: worst-case regret bounds. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30.* Curran Associates, Inc., 1184–1194.

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

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

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

Besbes, Omar, Yonatan Gur, Assaf Zeevi. 2014. Stochastic multi-armed bandit with non-stationary rewards. *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS).*

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

Chen, Yifang, Chung-Wei Lee, Haipeng Luo, Chen-Yu Wei. 2019. A new algorithm for non-stationary contextual bandits: Efficient, optimal, and parameter-free. *Proceedings of Conference on Learning Theory (COLT).*

Cheung, Wang Chi, David Simchi-Levi, Ruihao Zhu. 2019. Learning to optimize under non-stationarity. *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS).*

Dick, Travis, András György, Csaba Szepesvári. 2014. Online learning in markov decision processes with changing cost sequences. *Proceedings of the International Conference on Machine Learning (ICML).*

Even-Dar, Eyal, Sham M Kakade, , Yishay Mansour. 2005. Experts in a markov decision process. *Proceedings of the 19th Annual Conference on Neural Information Processing Systems (NIPS).*

Fruit, Ronan, Matteo Pirotta, Alessandro Lazaric. 2018a. Near optimal exploration-exploitation in non-communicating markov decision processes. S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett, eds., *Advances in Neural Information Processing Systems 31.* Curran Associates, Inc., 2998–3008.

Fruit, Ronan, Matteo Pirotta, Alessandro Lazaric, Ronald Ortner. 2018b. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. Jennifer Dy, Andreas Krause, eds., *Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research,* vol. 80. PMLR, Stockholmsmssan, Stockholm Sweden, 1578–1586.
Gajane, Pratik, Ronald Ortner, Peter Auer. 2018. A sliding-window algorithm for markov decision processes with arbitrarily changing rewards and transitions. CoRR abs/1805.10066. URL http://arxiv.org/abs/1805.10066.

Garivier, A., E. Moulines. 2011. On upper-confidence bound policies for switching bandit problems. Proceedings of International Conference on Algorithmic Learning Theory (ALT).

Hoeffding, Wassily. 1963. Probability inequalities for sums of bounded random variables. Journal of the American statistical association 58(301) 13–30.

Jadbabaie, Ali, Alexander Rakhlin, Shahin Shahrampour, Karthik Sridharan. 2015. Online optimization : Competing with dynamic comparators. Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS).

Jaksch, Thomas, Ronald Ortner, Peter Auer. 2010. Near-optimal regret bounds for reinforcement learning. J. Mach. Learn. Res. 11 1563–1600.

Karnin, Z., O. Anava. 2016. Multi-armed bandits: Competing with optimal sequences. Proceeding of Annual Conference on Neural Information Processing Systems (NIPS).

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

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

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

Nilim, Arnab, Laurent El Ghaoui. 2005. Robust control of markov decision processes with uncertain transition matrices. Operations Research.

Qin, Zhiwei (Tony), Jian Tang, Jieping Ye. 2019. Deep reinforcement learning with applications in transportation. Tutorial of the 33rd AAAI Conference on Artificial Intelligence (AAAI-19).

Weissman, Tsachy, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, , Marco L. Weinberger. 2003. Inequalities for the l1 deviation of the empirical distribution. Technical Report HPL-2003-97, HP Laboratories Palo Alto: www.hpl.hp.com/techreports/2003/HPL-2003-97R1.

Xu, Huan, Shie Mannor. 2006. The robustness-performance tradeoff in markov decision processes. Proceedings of the 20th Annual Conference on Neural Information Processing Systems (NIPS).

Yu, Jia Yuan, Shie Mannor. 2009. Online learning in markov decision processes with arbitrarily changing rewards and transitions. Proceedings of the International Conference on Game Theory for Networks.

Zhang, Anru, Mengdi Wang. 2018. Spectral state compression of markov processes. https://arxiv.org/abs/1802.02920.
Appendix. Supplementary

A. Supplementary Details about MDPs

The optimal long term reward \( \rho^*_t \) is equal to the optimal value of the linear program \( P(r_t, p_t) \). For a reward vector \( r \) and a transition kernel \( p \), we define

\[
P(r, p) \max_{s \in S, a \in A_s} r(s, a) x(s, a) \tag{18}
\]

subject to

\[
\sum_{a \in A_s} x(s, a) = \sum_{s' \in S, a' \in A_{s'}} p(s' | s, a') x(s', a') \quad \forall s \in S
\]

\[
\sum_{s \in S, a \in A_s} x(s, a) = 1
\]

\[
x(s, a) \geq 0 \quad \forall s \in S, a \in A_s
\]

Throughout our analysis, it is useful to consider the following dual formulation \( D(r, p) \) of the optimization problem \( P(r, p) \):

\[
D(r, p) \min \rho \tag{19}
\]

subject to

\[
\rho + \gamma(s) \geq r(s, a) + \sum_{s' \in S} p(s' | s, a) \gamma(s') \quad \forall s \in S, a \in A_s
\]

\[
\phi, \gamma(s) \text{ free} \quad \forall s \in S.
\]

The following Lemma shows that any feasible solution to \( D(r, p) \) is essentially bounded if the underlying MDP is communicating, which will be crucial in the subsequent analysis.

**Lemma 2.** Let \( (\rho, \gamma) \) be a feasible solution to the dual problem \( D(r, p) \), where \( (S, A, p) \) constitute a communicating MDP with diameter \( D \). We have

\[
\max_{s, s' \in S} \{ \gamma(s) - \gamma(s') \} \leq 2D.
\]

The Lemma is extracted from Section 4.3.1 of [Jaksch et al. 2010], and it is more general than [Lattimore and Szepesvári 2018], which requires \( (\rho, \gamma) \) to be optimal instead of just feasible.

A.1. Extended Value Iteration by [Jaksch et al. 2010]

We provide the pseudo-codes of the extended value iteration proposed by [Jaksch et al. 2010], displayed in Algorithm 3. By [Jaksch et al. 2010], the algorithm converges in finite time when \( H^p \) contains a transition kernel \( p \) such that \( (S, A, p) \) constitutes a communicating MDP.

Appendix. Proofs

A. Proof of Lemma 1

We divide the proof into two parts.
we first prove the following proposition. Following a covering argument from (Lattimore and Szepesvári 2018), the assumption of (Weissman et al. 2003), and we have to appeal to other approaches to analyze the deviation

\[ \Pr[ A.2. Proving \quad (\text{for any fixed } s, a, t) \quad \text{Proposition 3.} \quad \text{and we can further conclude the statement by a simple union bound over all possible choices of } (s, a, t) \in (S, A, [T]). \]

A.1. Proving \( \Pr[E_s] \geq 1 - \delta/2 \)

The proof makes use of the well known Hoeffding’s inequality (Hoeffding 1963) as stated below.

**Proposition 2.** Let \( \{X_t\}_{t=1}^T \in [0, 1]^T \) be a martingale difference sequence adapted to a filtration \( \{\mathcal{F}_t\}_{t=1}^T \). That is, \( X_t \) is \( \mathcal{F}_t \)-measurable, and \( E[X_t | \mathcal{F}_{t-1}] = 0 \). For any \( \delta \in (0, 1) \), we have

\[ \Pr \left[ \left\| \frac{1}{T} \sum_{t=1}^T X_t \right\| \leq \sqrt{\frac{2\log(2/\delta)}{T}} \right] \geq 1 - \delta. \]

Now, by Hoeffding’s inequality

\[ \Pr [\|\hat{r}_t(s, a) - \bar{r}_t(s, a)\| \leq \text{rad-}\bar{r}_t(s, a)] \geq 1 - \frac{\delta}{2\text{SAT}}, \]

and we can further conclude the statement by a simple union bound over all possible choices of \( (s, a, t) \in (S, A, [T]). \)

A.2. Proving \( \Pr[E_p] \geq 1 - \delta/2 \)

Unlike the stationary setting, the data in the non-stationary environment is no longer i.i.d., which violates the assumption of (Weissman et al. 2003), and we have to appeal to other approaches to analyze the deviation of the estimated transition kernels. Following a covering argument from (Lattimore and Szepesvári 2018), we first prove the following proposition.

**Proposition 3.** For any fixed \( s \in S, a \in A, \text{ and } t \in [T], \)

\[ \Pr [\|\tilde{p}_t(., s) - \bar{p}_t(., s)\|_1 \leq \text{rad}^p_{\gamma}(s, a)] \geq 1 - \frac{\delta}{2\text{SAT}}. \]
Proof of Proposition 3. We consider only non-trivial case where the \((s, a)\) pairs that has \(N_t(s, a) \geq 1\). By definition,
\[
\|\hat{p}_t(\cdot | s, a) - \hat{p}_t(\cdot | s, a)\|_1 = \left\| \hat{p}_t(\cdot | s, a) - \frac{\sum_{q=(r(m) - W)^{\neq 1} \cap \mathbb{Q}_t(\cdot | s, a) 1(s_q = s, a_q = a)}{N_t(s, a)} \right\|_1
\]
We follow the covering argument from [Lattimore and Szepesvári 2018]
\[
\|\hat{p}_t(\cdot | s, a) - \frac{\sum_{q=(r(m) - W)^{\neq 1} \cap \mathbb{Q}_t(\cdot | s, a) 1(s_q = s, a_q = a)}{N_t(s, a)} \right\|_1 = \sup_{x \in \mathbb{R}^2: \|x\|_\infty \leq 1} \left\langle \hat{p}_t(\cdot | s, a) - \frac{\sum_{q=(r(m) - W)^{\neq 1} \cap \mathbb{Q}_t(\cdot | s, a) 1(s_q = s, a_q = a)}{N_t(s, a)}, x \right\rangle. \tag{21}
\]
Consider the finite subset \(C\), which is the regular grid with stride \(1/(2W)\) of the set \([-1, 1]^S\), and it can be easily verified that for any \(x \in [-1, 1]^S\), there exists \(y \in C\) such that \(\|x - y\|_{\infty} \leq 1/(2W)\). By the Hoeffding's inequality (Proposition 2) and a union bound over all possible \(y \in C\) and all possible values of \(N_t(s, a) \in [W]\), we have with probability at least \(1 - \delta\), for all \(y \in C\),
\[
\left\langle \hat{p}_t(\cdot | s, a) - \frac{\sum_{q=(r(m) - W)^{\neq 1} \cap \mathbb{Q}_t(\cdot | s, a) 1(s_q = s, a_q = a)}{N_t(s, a)}, y \right\rangle \leq \frac{2 \log \left( |C| W/\delta \right)}{N_t(s, a)} \leq \frac{2 \log \left( (4W)^2 W/\delta \right)}{N_t(s, a)} \tag{22}
\]
Here, we sequentially exploit the fact that \( - \) \( y \) is the closest point to \( x^* \) in \( C \). From inequality \( 22 \), we have with probability at least \( 1 - \delta' \), \( 23 \) is upper bounded as

\[
\sqrt{\frac{2\log((4W)^8W/\delta')}{N_t(s,a)}} + 1/W
\leq \sqrt{\frac{2\log((4W)^8W/\delta')}{N_t(s,a)}} + \frac{1}{N_t(s,a)}
\leq \sqrt{\frac{2\log((4W)^8W/\delta')}{N_t(s,a)}} + \sqrt{\frac{1}{N_t(s,a)}}
\leq \sqrt{\frac{2\log((4W)^8W/\delta')}{N_t(s,a)}} + \frac{2\log((4W)^8W/\delta')}{N_t(s,a)}.
\]

Here, we sequentially exploit the fact that \( N_t(s,a) \leq W \), \( N_t(s,a) \geq 1 \), and \( 2\log((4W)^8W/\delta') \geq 1 \). From inequalities \( 23 \) and \( 24 \), we have with probability at least \( 1 - 1/(2SAT) \) (i.e., setting \( \delta' = 1/(2SAT) \))

\[
\left\| \hat{p}_t(\cdot|s,a) - \frac{\sum_{q=(r(m)-W)}^{t-1} p_q(|s,a)1(s_q = s, a_q = a)}{N_t(s,a)} \right\|_1 \leq 2\sqrt{2S\log(8ATW/\delta')/N_t(s,a)}.
\]

Now, by an union bound over all possible choices of \( (s,a,t) \in (S,A,[T]) \), we can conclude the statement.

**B. Proof of Proposition 1**

In this section, we prove Proposition 1. In doing so, we need the following auxiliary claims and lemmas

**Claim 1.** Let \( t \) be in episode \( m \), then

\[
|r_t(s,a) - \bar{r}_{r(m)}(s,a)| \leq \text{var}_{r,t}
\]

holds for every state-action pair \((s,a)\).

**Lemma 3.** Conditioned on \( \mathcal{E}_r, \mathcal{E}_p \), if \( \eta \geq B_pW/(2T) \), then

\[
\hat{p}_{r(m)} \geq \rho^*_t - \text{var}_{r,t} - 2D_{r(m)} \cdot \text{var}_{p,t}
\]

for any \( t \) in episode \( m \).

**Claim 2.** Let \( t \) be in episode \( m \). Conditioned on \( \mathcal{E}_p \) and supposing that \( \eta \geq B_{p,t}W/2 \), for every state-action pair \((s,a)\) we have

\[
\sum_{s'} p_{r(m)}(s'|s,a)\gamma_{r(m)}(s') \geq \sum_{s'} p_{t}(s'|s,a)\gamma_{r(m)}(s') - 2D_{r(m)} \left[ \text{var}_{p,t} + 2\text{rad} \cdot p_{r(m)}(s,a) + \eta \right].
\]

Claims 1, 2 and Lemma 3 are proved in Sections B.1, B.4, B.3 respectively. To prove Claim 2 and Lemma 3, we further require two auxiliary results, Claim 3 and Lemma 4, which are stated and proved in Section B.2.
B.1. Proof of Claim 1

We have

\[ |r_t(s, a) - \tilde{r}_{\tau(m)}(s, a)| \leq |r_t(s, a) - r_{\tau(m)}(s, a)| + |r_{\tau(m)}(s, a) - \tilde{r}_{\tau(m)}(s, a)| \]

\[ \leq \sum_{q=r(m)}^{t-1} |r_{q+1}(s, a) - r_q(s, a)| + \frac{1}{W} \sum_{w=1}^{W} |r_{\tau(m)} - r_{\tau(m) - w}|. \]

By the definition of \( B_{r,q} \), we have

\[ \sum_{q=r(m)}^{t-1} |r_{q+1}(s, a) - r_q(s, a)| \leq \sum_{q=r(m)}^{t-1} B_{r,q}, \]

and

\[ \frac{1}{W} \sum_{w=1}^{W} |r_{\tau(m)} - r_{\tau(m) - w}| \leq \frac{1}{W} \sum_{w=1}^{W} \sum_{i=1}^{w} |r_{\tau(m) - i+1} - r_{\tau(m) - i}| \]

\[ = \frac{1}{W} \sum_{i=1}^{W} \sum_{w=i}^{W} |r_{\tau(m) - i+1} - r_{\tau(m) - i}| \]

\[ = \sum_{i=1}^{W} \frac{W - i + 1}{W} |r_{\tau(m) - i+1} - r_{\tau(m) - i}| \leq \sum_{i=1}^{W} \frac{W - i + 1}{W} B_{r,\tau(m) - i}. \]

Altogether, the claim is shown.

B.2. Further Auxiliary Results

Claim 3 and Lemma 4 are stated as follows.

Claim 3. Let \( t \) be in episode \( m \). For every state-action pair \((s, a)\) we have

\[ \|p_t(\cdot|s, a) - \tilde{p}_{\tau(m)}(\cdot|s, a)\|_1 \leq \text{var}_{p_t}. \]

Proof of Claim 3 The proof is similar to that of Claim 1.

Lemma 4. Let \( t \) be in episode \( m \). Conditioned on \( \mathcal{E}_p \), and supposing that \( \eta \geq B_pW/(2T) \), we have 0 \leq \hat{\gamma}_{\tau(m)}(s) \leq 2D_{\tau(m)} for each \( s \in \mathcal{S} \) with certainty.

Proof of Lemma 4 First, we claim that \( p_{\tau(m)}(\cdot|s, a) \in H_{p,t}(s, a; \eta) \), conditional on \( \mathcal{E}_p \) and assuming that \( \eta \geq B_pW/(2T) \). Indeed, we know that \( \tilde{p}_{\tau(m)}(\cdot|s, a) \in H_{p,t}(s, a; 0) \) for each state-action pair. Now, by Claim 3 we also know that \( |p_{\tau(m)}(\cdot|s, a) - \tilde{p}_{\tau(m)}(\cdot|s, a)| \leq B_pW/(2T) \). Altogether, our first claim is shown. To complete the proof, by Property 1 for the output of EVI, we know that \( (\hat{\rho}_{\tau(m)}, \hat{\gamma}_{\tau(m)}) \) is feasible to the dual problem \( D(\tau_{\gamma(m)}, p_{\tau(m)}) \). Therefore, the required bounds for \( \hat{\gamma}_{\tau(m)} \) hold by applying Lemma 2.

B.3. Proof of Lemma 3

As argued in the proof of Lemma 4, we know that \( p_{\tau(m)}(\cdot|s, a) \in H_{p,t}(s, a; \eta) \), conditional on \( \mathcal{E}_p \) and assuming that \( \eta \geq B_pW/(2T) \). In addition, condition on \( \mathcal{E}_r \), we know that \( \tilde{r}_{\tau(m)}(s, a) \in H_{r,t}(s, a) \) for each state action pair \( s, a \). Consequently, by Property 1 of the output of EVI, we know that

\[ \hat{\rho}_{\tau(m)} + \hat{\gamma}_{\tau(m)}(s) \geq \tilde{r}_{\tau(m)}(s, a) + \sum_{s' \in \mathcal{S}} \hat{\gamma}_{\tau(m)}(s') p_{\tau(m)}(s'|s, a), \]  

(26)
where $\gamma_{\tau(m)} \in [0, 2D_{\tau(m)}]$ by Lemma 4.

Now, we claim that $\tilde{\rho}_{\tau(m)} + \var_{r,t} + 2D_{\tau(m)} \cdot \var_{p,t} \cdot \gamma_{\tau(m)}$ is a feasible solution to the $t$th period dual problem $D(r_t, p_t)$, which immediately implies the Lemma. To demonstrate the claim, for every state action pair $(s, a)$ we have

$$\tilde{\rho}_{\tau(m)}(s, a) \geq r_t(s, a) - \var_{r,t}$$

$$\sum_{s' \in S} \gamma_{\tau(m)}(s') p_{r_t}(s'|s, a) \geq \sum_{s' \in S} \gamma_{\tau(m)}(s') p_t(s'|s, a) - \|\gamma_{\tau(m)}\|_\infty \|p_t(\cdot|s, a) - \tilde{\rho}_{\tau(m)}(\cdot|s, a)\|_1$$

$$\geq \sum_{s' \in S} \gamma_{\tau(m)}(s') p_t(s'|s, a) - \frac{2D_{\tau(m)} B_p(t - \tau(m))}{T}$$

$$\geq \sum_{s' \in S} \gamma_{\tau(m)}(s') p_t(s'|s, a) - 2D_{\tau(m)} \cdot \var_{p,t},$$

where step (27) is by Claim 1 and step (28) is by Claim 3 and Lemma 4. Altogether, putting (27), (28) to (26), our claim is shown, and the Lemma is proved.

B.4. Proof of Claim 2

Claim 2. Let $t$ be in episode $m$. Conditioned on $E_p$ and supposing that $\eta \geq B_{p,t} W/2$, for every state-action pair $(s, a)$ we have

$$\sum_{s' \in S} \tilde{\rho}_{\tau(m)}(s'|s, a) \gamma_{\tau(m)}(s') \geq \sum_{s' \in S} p_t(s'|s, a) \gamma_{\tau(m)}(s') - 2D_{\tau(m)} \left[ \var_{p,t} + 2\text{rad} - \rho_{\tau(m)}(s, a) + \eta \right].$$

Proof of Claim 2. We have

$$\left| \sum_{s' \in S} \left[ \tilde{\rho}_{\tau(m)}(s'|s, a) - p_t(s'|s, a) \right] \gamma_{\tau(m)}(s') \right|$$

$$\leq \left\| \gamma_{\tau(m)} \right\|_\infty \left( \left\| \tilde{\rho}_{\tau(m)}(\cdot|s, a) - \rho_{\tau(m)}(\cdot|s, a) \right\|_{(a)} + \left\| \tilde{\rho}_{\tau(m)}(\cdot|s, a) - p_t(\cdot|s, a) \right\|_{(b)} + \left\| p_t(\cdot|s, a) - p_t(\cdot|s, a) \right\|_{(c)} \right).$$

In step (29), we know that

- (a) $\leq 2D_{\tau(m)}$ by Lemma 4
- (b) $\leq 2\text{rad} - \rho_{\tau(m)}(s, a) + \eta$, by the facts that $\tilde{\rho}_{\tau(m)}(\cdot|s, a) \in H_{p,\tau(m)}(s, a; \eta)$ and $\rho_{\tau(m)}(\cdot|s, a) \in H_{p,\tau(m)}(s, a; 0),$
- (c) $\leq \var_{p,t}$ by Claim 3

Altogether, the Claim is proved.

B.5. Finalizing the Proof

Now, we have

$$r_t(s_t, a_t)$$

$$\geq \gamma_{\tau(m)}(s_t, a_t) - \var_{r,t}$$

$$\geq \gamma_{\tau(m)}(s_t, a_t) - \var_{r,t} - 2 \cdot \text{rad} - \rho_{\tau(m)}(s_t, a_t)$$

$$\geq \tilde{\rho}_{\tau(m)} + \gamma_{\tau(m)}(s_t) - \sum_{s' \in S} \tilde{\rho}_{\tau(m)}(s'|s_t, a_t) \gamma_{\tau(m)}(s') - \frac{1}{\sqrt{\tau(m)}}$$
We accomplish the promised dynamic regret bound by the following three Lemmas that bound the dynamic regret as follows: denoting $\rho_t \equiv \rho_{r(m)}(s_t, a_t)$, we have

$$- \text{var}_{r,t} - 2 \cdot \text{rad} - r_{r(m)}(s_t, a_t)$$

$$\geq \rho_t + \hat{\gamma}_{r(m)}(s_t) - \sum_{s' \in S} p_t(s'|s_t, a_t)\hat{\gamma}_{r(m)}(s') - \frac{1}{\sqrt{\tau(m)}}$$

$$- 2\left[\text{var}_{r,t} + \text{rad} - r_{r(m)}(s_t, a_t)\right] - 2D_{r(m)} \left[2 \cdot \text{var}_{p,t} + 2 \cdot \text{rad} - p_{r(m)}(s, a) + \eta\right].$$

Step (30) is by Claim 1. Step (31) is by conditioning that event $\mathcal{E}_r$ holds. Step (32) is by Property 2 for the output of EVI. In step (33), we upper bound $\rho_t$ by Lemma 3 and we upper bound $\sum_{s' \in S} \hat{p}_{r(m)}(s'|s_t, a_t)\hat{\gamma}_{r(m)}(s')$ by Claim 2. Rearranging gives the Proposition.

## C. Proof of Theorem 1

In this section, we present the complete proof of Theorem 1. Using Proposition 1, we can bound the dynamic regret as follows: denoting $M(T)$ as the total number of episodes, and by abusing the notation we let $\tau(M(T) + 1) - 1 = T$. Episode $M(T)$ is interrupted by the break in Line 11 and the algorithm is forced to terminate as the end of time $T$ is reached.

$$\text{Dyn-Reg}_T(\text{SWUCRL2-CW}) = \sum_{t=1}^{T} \rho_t - r_t(s_t, a_t)$$

$$\leq \sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m+1)-1} \left\{ \left[ \sum_{s' \in S} p_t(s'|s_t, a_t)\hat{\gamma}_{r(m)}(s') - \hat{\gamma}_{r(m)}(s_t) \right] + \frac{1}{\sqrt{\tau(m)}} \right\}$$

$$+ \sum_{t=1}^{T} \left\{ 2\text{var}_{r,t} + 4D_{\max} \cdot \text{var}_{p,t} + 2D_{\max} \eta \right\}$$

$$+ \sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m+1)-1} \left\{ 2\text{rad} - r_{r(m)}(s_t, a_t) + 4D_{\max} \cdot \text{rad} - p_{r(m)}(s_t, a_t) \right\}.$$  

We accomplish the promised dynamic regret bound by the following three Lemmas that bound the dynamic regret terms $\bigcirc \bigcirc \bigcirc$.

### C.1. Bounding $\bigcirc$

**Lemma 5.** Conditioning on events $\mathcal{E}_r, \mathcal{E}_p$ and assuming that $\eta \geq B_{p,1}W/2$, with probability $1 - O(\delta)$ we have $\bigcirc = \tilde{O}(D_{\max}[M(T) + \sqrt{T}]).$ Under the same assumption, with certainty we have $M(T) \leq SA(2 + \log_2 W)/T/W = \tilde{O}(SAT/W)$.

First, to demonstrate the bound for $M(T)$, it suffices to show that there are at most $SA(2 + \log_2 W)$ many episodes in each of the following cases: (1) between time steps 1 and $W$, (2) between time steps $jW$ and $(j + 1)W$, for any $j \in \{1, \ldots, \lceil T/W \rceil - 1\}$, (3) between time steps $\lceil T/W \rceil \cdot W$ and $T$. We focus on case (2), and the edge cases (1, 3) can be analyzed similarly.

Between time steps $jW$ and $(j + 1)W$, a new episode $m + 1$ is started when the second criterion $\nu_{m}(s_t, \pi_{m}(s_t)) < N^{+}_{r(m)}(s_t, \pi_{m}(s_t))$ is violated during the current episode $m$. We first illustrate how the second criterion is invoked for a fixed state-action pair $(s, a)$, and then bound the number of invocations due to $(s, a)$. Now, let’s denote $m^1, \ldots, m^L$ as the episode indexes, where $jW \leq \tau(m^1) < \tau(m^2) < \ldots < \tau(m^L) < (j + 1)W$, and the second criterion for $(s, a)$ is invoked during $m^\ell$ for $1 \leq \ell \leq L$. That is, for each $\ell \in \{1, \ldots, L\}$, the DM breaks the while loop for episode $m^\ell$ because $\nu_{m^\ell}(s, a) = N^{+}_{r(m^\ell)}(s, a)$, leading to the new episode $m^\ell + 1$. 
To demonstrate our claimed bound for $M(T)$, we show that $L \leq 2 + \log_2 W$ as follows. To ease the notation, let’s denote $\psi^d = \nu_m(t, s, a)$. We first observe that $\psi^1 = N_{r(m)}^+(s, a) \geq 1$. Next, we note that for $\ell \in \{2, \ldots, L\}$, we have $\psi^\ell \geq \sum_{k=1}^{\ell-1} \psi^k$. Indeed, we know that for each $\ell$ we have $(\tau^\ell + 1) - (m^1) \leq W$, by our assumption on $m^1, \ldots, m^\ell$. Consequently, the counting sum in $N_{r(m)}(s, a)$, which counts the occurrences of $(s, a)$ in the previous $W$ time steps, must have count those occurrences corresponding to $\psi^1, \ldots, \psi^{\ell-1}$.

Altogether, during the $T$ time steps, there are at most $(SAT(2 + \log_2 W))/W$ episodes due to the second criterion and $T/W$ due to the first, leading to our desired bound on $M(T)$.

Next, we establish the bound for $\psi$. Now, by virtue of confidence widening and Lemma 2, we know that $\hat{\gamma}_{\tau(m)}(s) \in [0, 2D_{\max}]$ for all $m$ and $s$ by Lemma $\Box$. For each episode $m$, we have

$$
\sum_{t=\tau_1}^{\tau(m)+1} \sum_{s' \in S} p_t(s'|s_t, a_t) \hat{\gamma}_{\tau(m)}(s') - \hat{\gamma}_{\tau(m)}(s_t) \leq 2D_{\max} \cdot M(T) + \sum_{t=1}^{T} Y_t.
$$

Summing (34) over $m = 1, \ldots, M(T)$ we get

$$
\sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m)+1} \sum_{s' \in S} p_t(s'|s_t, a_t) \hat{\gamma}_{\tau(m)}(s') - \hat{\gamma}_{\tau(m)}(s_t) \leq 2D_{\max} \cdot M(T) + \sum_{t=1}^{T} Y_t.
$$

Define the filtration $\mathcal{H}_{t-1} = \{(s_q, a_q, R_q(s_q, a_q))\}_{q=1}^t$. Now, we know that $\mathbb{E}[Y_t|\mathcal{H}_{t-1}] = 0$, $Y_t$ is $\sigma(\mathcal{H}_t)$-measurable, and $|Y_t| \leq 2D_{\max}$. Therefore, we can apply the Hoeffding inequality (see Proposition 2) to show that

$$
\sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m)+1} \sum_{s' \in S} p_t(s'|s_t, a_t) \hat{\gamma}_{\tau(m)}(s') - \hat{\gamma}_{\tau(m)}(s_t) = O(D_{\max} \cdot M(T) + D_{\max} \sqrt{T \log \frac{1}{\delta}})
$$

with probability $1 - O(\delta)$. Finally, note that

$$
\sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m)+1} \frac{1}{\sqrt{\tau(m)}} \leq \sum_{i=1}^{\lceil T/W \rceil} \frac{W}{\sqrt{iW}} = O(\sqrt{T}).
$$

Hence, the Lemma is proved.

**C.2. Bounding Φ**

**Lemma 6.** With certainty, $\Phi = O((B_r + D_{\max}B_p)W + D_{\max}T\eta)$.

Proof of Lemma 6. We first bound $\sum_{t=1}^{T} \text{var}_{r,t}$. Now, recall the definition that, for time $t$ in episode $m$, we have defined $\text{var}_{r,t} = \sum_{q=\tau(m)-W}^{t-1} B_{r,q}$. Clearly, for $iW \leq q < (i+1)W$, the summand $B_{r,q}$ only appears in $\text{var}_{r,t}$ for $iW \leq q < t \leq (i+2)W$, since each episode is contained in $\{iW, \ldots, (i+1)W\}$ by our first criterion in Line 7 in Algorithm 1. Altogether, we have

$$
2 \sum_{t=1}^{T} \text{var}_{r,t} \leq 2 \sum_{t=1}^{T-1} B_{r,t} W = 2B_r W.
$$

(36)
Next, we bound $\sum_{t=1}^{T} \text{var}_{p,t}$. Now, we know that $\tau(m + 1) - \tau(m) \leq W$ by the first criterion in Line 7 in Algorithm 1. Consequently,

$$4D_{\max} \sum_{t=1}^{T} \text{var}_{p,t} \leq 4D_{\max} \sum_{t=1}^{T-1} \frac{2B_{\mu}}{T} = 8D_{\max}B_{\mu}W. \quad (37)$$

Finally, combining $(36, 37)$ with $2D_{\max} \sum_{t=1}^{T} \eta$, the Lemma is established.

### C.3. Bounding $\bigtriangledown$

**Lemma 7.** With certainty, we have $\bigtriangledown = \tilde{O}((\sqrt{SAT} + D_{\max}S\sqrt{AT})/\sqrt{W})$.

**Proof of Lemma 7.** We first show that, with certainty,

$$M(T) \tau(m+1)-1 \sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m+1)-1} \text{rad}_r(m)(s_t, a_t) = M(T) \tau(m+1)-1 \sum_{m=1}^{M(T)} \sum_{t=\tau(m)}^{\tau(m+1)-1} \text{rad}_p(m)(s_t, a_t) = O\left(\sqrt{S\Sigma W}\right). \quad (38)$$

We analyze by considering the dynamics of the algorithm in each consecutive block of $W$ time steps, in a way similar to the proof of Lemma 5. Consider the episodes indexes $m_0, m_1, \ldots, m_{\lfloor T/W \rfloor}, m_{\lfloor T/W \rfloor+1}$, where $\tau(m_0) = 1$, and $\tau(m_j) = jW$ for $j \in \{1, \ldots, \lfloor T/W \rfloor\}$, and $m_{\lfloor T/W \rfloor+1} = m(T) + 1$ (so that $\tau(m_{\lfloor T/W \rfloor+1})$ is the time index for the last episode in the horizon).

To prove $(38)$, it suffices to show that, for each $j \in \{0, 1, \ldots, \lfloor T/W \rfloor\}$, we have

$$\sum_{m=m_j}^{m_{j+1}-1} \sum_{t=\tau(m)}^{\tau(m+1)-1} \frac{1}{\sqrt{N_{\tau(m)}^+(s_t, a_t)}} = O\left(\sqrt{S\Sigma W}\right). \quad (40)$$

Without loss of generality, we assume that $j \in \{1, \ldots, \lfloor T/W \rfloor - 1\}$, and the edge cases of $j = 0, \lfloor T/W \rfloor$ can be analyzed similarly.

Now, we fix a state-action pair $(s, a)$ and focus on the summands in $(40)$:

$$\sum_{m=m_j}^{m_{j+1}-1} \sum_{t=\tau(m)}^{\tau(m+1)-1} 1((s_t, a_t) = (s, a)) = \sum_{m=m_j}^{m_{j+1}-1} \frac{\nu_m(s, a)}{\sqrt{N_{\tau(m)}^+(s, a)}} \quad (41)$$

It is important to observe that:

1. It holds that $\nu_{m_j}(s, a) \leq N_{\tau(m_j)}(s, a)$, by the first criterion in Line 7 in Algorithm 1.

2. For $m \in \{m_j, m_{j+1}-1\}$, we have $\sum_{m=m_j}^{m_{j+1}-1} \nu_{m'}(s, a) \leq N_{\tau(m)}(s, a)$ . Indeed, we know that episodes $m_j, m_{j+1}-1$ are inside the time interval $\{jW, \ldots, (j+1)W\}$. Consequently, the counts of “$(s_t, a_t) = (s, a)$” associated with $\{\nu_{m'}(s, a)\}_{m'=m_j}^{m_{j+1}-1}$ are contained in the $W$ time steps preceding $\tau(m)$, hence the desired inequality.

With these two observations, we have

$$\frac{\nu_{m_j}(s, a)}{\max\{\nu_{m_j}(s, a), 1\}} + \sum_{m=m_j}^{m_{j+1}-1} \frac{\nu_m(s, a)}{\max\{\sum_{m'=m_j}^{m_{j+1}-1} \nu_{m'}(s, a), 1\}} \leq \sqrt{\max\{\nu_{m_j}(s, a), 1\} + (\sqrt{2} + 1) \max\left\{ \sum_{m=m_j}^{m_{j+1}-1} \nu_m(s, a), 1 \right\}} \quad (42)$$
\[
\leq (\sqrt{2} + 2) \sqrt{\max \left\{ \sum_{t=jW}^{(j+1)W-1} \mathbb{1}(s_t, a_t) = (s, a), 1 \right\}}.
\]

Step (42) is by Lemma 19 in [Jaksch et al. 2010], which bounds the sum in the previous line. Step (43) is by the fact that episodes \(m_j, \ldots, m_{j+1} - 1\) partitions the time interval \(jW, \ldots, (j+1)W - 1\), and \(\nu_m(s, a)\) counts the occurrences of \((s_t, a_t) = (s, a)\) in episode \(m\). Finally, observe that (41) = 0 if \(\nu_m(s, a) = 0\) for all \(m \in \{m_j, \ldots, m_{j+1} - 1\}\). Thus, we can refine the bound in (43) to

\[
\leq (\sqrt{2} + 2) \sqrt{\max \left\{ \sum_{t=jW}^{(j+1)W-1} \mathbb{1}(s_t, a_t) = (s, a), 1 \right\}}.
\]

The required inequality (40) is finally proved by summing (44) over \(s \in S, a \in A\) and applying Cauchy Schwartz:

\[
\leq (\sqrt{2} + 2) \sqrt{\sum_{s \in S, a \in A} \left( \frac{1}{\sqrt{N^+_{\tau(r_m)}(s, a)}} \sum_{s \in S, a \in A} \sum_{m=m_j}^{m_{j+1}-1} \nu_m(s, a) \sqrt{N^+_{\tau(r_m)}(s, a)} \right)}
\leq (\sqrt{2} + 2) \sqrt{\sum_{s \in S, a \in A} \left( \frac{1}{\sqrt{N^+_{\tau(r_m)}(s, a)}} \sum_{s \in S, a \in A} \sum_{m=m_j}^{m_{j+1}-1} \nu_m(s, a) \sqrt{N^+_{\tau(r_m)}(s, a)} \right)}
\leq (\sqrt{2} + 2) \sqrt{SAW}.
\]

Altogether, the Lemma is proved.

\section*{D. More on Uniform Variation Budget and Confidence Widening}

\subsection*{D.1. The Peril of Averaging Time Varying MDPs}

Following the discussion in Section 4.4, we show that the MDP constituted by \(\hat{p}_N\) is not “similar” to the MDP constitute by \(p_{t'}\) for any \(t' \in [T]\) (in the sense shown by [Jaksch et al. 2010]) by demonstrating that non-stationary MDPs present a difficult environment for learning directly from time average.

\textbf{Lemma 8.} There exists a sequence of non-stationary MDP transition kernels \(p_1, \ldots, p_N\) (which have the same \(S, A\)) such that

1. The diameter of \((S, A, p_n)\) is 1 for each \(n \in [N]\);
2. The total variation budget is \(B_p = O(1)\);
3. Under some deterministic policy, the empirical MDP constituted by \((S, A, \tilde{p}_N)\) has diameter \(\Omega(\sqrt{N/\log N})\) for every \(\tilde{p}_N \in H_{p,N}(0)\).

The discrepancy between \((S, A, \tilde{p}_N)\) and \((S, A, p_1), \ldots, (S, A, p_N)\) is due to the fact that, for different \((s, a)\), the time average \(\tilde{p}_N(\cdot|s, a)\) involve the transition probability from different time steps. The Lemma demonstrates that the optimistic MDP derived from the time-averaged transition kernel \(\tilde{p}_N\) can be very different from each of the MDPs constituted by \((S, A, p_1), \ldots, (S, A, p_N)\).

In fact, even if the DM can construct, for each \(s, a\), the desired time average \(\tilde{p}(\cdot|s, a) = \sum_{n=1}^{N} p_n(\cdot|s, a)\), the resulting MDP constituted by \((S, A, \tilde{p})\) can still be very different from the MDPs constituted by \(\{(S, A, p_n)\}_{n=1}^{N}\).
Lemma 9. There exists \( p_1, \ldots, p_N \) such that

1. \((\mathcal{S}, \mathcal{A}, p_n)\) constitutes an MDP with diameter at most \( S \).
2. \( B_p = O(1) \).
3. \((\mathcal{S}, \mathcal{A}, p)\) constitutes an MDPs with diameter at least \( \Omega(S^2) \)

Before going to the proofs of Lemmas 8, 9, we remark that there is no contradiction between Lemma 8 and (Jaksch et al. 2010, Gajane et al. 2018), who demonstrate a \( O(\ell^{1/3} T^{2/3}) \) dynamic regret in a piece-wise stationary environment with at most \( \ell \) changes (The work by Gajane et al. 2018 improves upon the work by Jaksch et al. 2010) in the dependence on \( D, S, A \). Throughout this paragraph, we hide the dependence on \( D, S, A \) in the \( O(\cdot), \Theta(\cdot) \) notations, and only show the dependence on \( \ell, T \). Indeed, they partition the whole horizon into \( T \) in \( \Theta(\ell^{2/3} T^{1/3}) \) time intervals of length \( \Theta((T/\ell)^{2/3}) \) each, and run UCRL2 on each. The main argument is that at least \( \Theta(\ell^{2/3} T^{1/3}) - \ell \) intervals enjoy stationary MDP environments. In the interval with non-stationary MDP, which are at most \( \ell \) many, they bound the dynamic regret directly as \( \ell \) times the length of each interval, which is equal to \( O(\ell^{2/3} T^{1/3}) \). The latter linear-in-time-horizon bound does not contradict our previous claim that diameter is unbounded and the application of UCRL2 can be ruined.

Proof of Lemma 8

The sequence \( p_1, \ldots, p_N \) alternates between the following two instances \( p^1, p^2 \). Now, define the common state space \( \mathcal{S} = \{1, 2\} \) and action collection \( \mathcal{A} = \{A_1, A_2\} \), where \( A_1 = \{a_1, a_2\}, \{A_2\} = \{b_1, b_2\} \).

For \( p^1 \), we have
\[
p^1(1|1, a_1) = p^1(2|2, b_1) = 1, p^1(2|1, a_2) = p^1(1|2, b_2) = 1.
\]

For \( p^2 \), we have
\[
p^2(1|1, a_2) = p^2(2|2, b_2) = 1, p^2(2|1, a_1) = p^2(1|2, b_1) = 1.
\]

Please refer to Fig. 2 for a graphical illustration.

![Diagram](image_url)

**Figure 2** Example MDPs. Since the transitions are deterministic, the probabilities are omitted.

Clearly, we see that both instances have diameter 1. Now, consider the following two deterministic and stationary policy \( \pi^1 \):
\[
\pi^1(1) = a_1, \pi^1(2) = b_2,
\]
and \( \pi^2 \):
\[
\pi^2(1) = a_2, \pi^2(2) = b_1.
\]
Since the MDP is deterministic, we have $\hat{p}_N = \overline{p}_N$.

In the following, we construct a trajectory where the DM alternates between policies $\pi^1, \pi^2$ during time $\{1, \ldots, N\}$ 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 $\hat{p}_N(1|1, a_i) \approx 1$ for each $i \in \{1, 2\}$, and likewise $\hat{p}_N(2|2, b_i) \approx 1$ for each $i \in \{1, 2\}$. Altogether, this will lead the DM to conclude that $(S, A, \hat{p}_N)$ constitute a high diameter MDP, since the probability of transiting from state 1 to 2 (or 2 to 1) would appear very low.

The construction is detailed as follows. Let $N = 4\tau + 2$. In addition, let
- $p_1 = p_2 = \ldots = p_\tau = p^1$,
- $p_{\tau+1} = p_{\tau+2} = \ldots = p_{2\tau+1} = p^2$,
- $p_{2\tau+3} = p_{2\tau+4} = \ldots = p_{3\tau+1} = p^1$,
- $p_{3\tau+2} = p_{3\tau+3} = \ldots = p_{4\tau+2} = p^2$.

The DM starts at state 1. It follows
- policy $\pi^1$ from time 1 to time $2\tau + 1$,
- policy $\pi^2$ from $2\tau + 2$ to $4\tau + 2$.

Under the specified MDP models 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 $2\tau + 1$,
- action $b_1$ from time $2\tau + 2$ to $3\tau + 2$,
- action $a_2$ from time $3\tau + 3$ to $4\tau + 2$.

As a result, the DM is at
- state 1 from time 1 to $\tau + 1$,
- state 2 from time $\tau + 2$ to $2\tau + 1$,
- state 2 from time $2\tau + 2$ to $3\tau + 3$,
- state 1 from time $3\tau + 3$ to $4\tau + 2$.

From the dynamics above, we can see that

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

Finally, for the confidence region $H_{p,N+1}(0) = \{H_{p,N+1}(s, a; 0)\}_{s, a}$ constructed without confidence widening, we know that for any $\hat{p} \in H_{p,N+1}$ we have $\hat{p}(2|1, a_1), \hat{p}(2|1, a_2), \hat{p}(1|2, b_1), \hat{p}(1|2, b_2) = O\left(\sqrt{\frac{\log N}{\tau + 1}}\right)$, since the stochastic confidence radius $\Theta\left(\sqrt{\frac{\log N}{\tau + 1}}\right)$ dominates the sample mean $\frac{1}{\tau + 1}$. Therefore, for any $\hat{p} \in H_{p,N+1}$, the diameter of the MDP constructed by $(S, A, \hat{p})$ is at least $\Omega(\sqrt{N/\log N})$. 

Proof of Lemma 9] Denote $S = \{1, \ldots, S\}$. For each $1 \leq n \leq N/2$ and each $s \in S$, we have the set $A_s = \{a\}$ being a singleton, with $p_n(s+1|s,a) = 1$ when $1 \leq s < S$, and $p_n(1|s,a) = 1$. That is, the MDP is just a Markov Chain on a directed cycle. For each $N/2 + 1 \leq n \leq N$ and each $s \in S$, we have the reverse: $p_n(s-1|s,a) = 1$ when $1 < s \leq S$, and $p_n(1|s,a) = 1$. That is, when $N/2 + 1 \leq n \leq N$, the directed cycle is reversed. It is clear that each $(S, A, p_n)$ constitutes an MDP with diameter at most $S$. However, the resulting time average transition kernel $\bar{p}$ has property that $\bar{p}(s-1|s,a) = \bar{p}(s+1|s,a) = 1/2$ for $s = 2, \ldots, S-1$, and $\bar{p}(S-1|S,a) = \bar{p}(1|S,a) = 1/2$, and $\bar{p}(S|1,a) = \bar{p}(2|1,a) = 1/2$. It can be easily verified that the resulting MDP formed by $(S, A, \bar{p})$ has diameter at least $\Omega(S^2)$ as follows.

We follow the notations in Definition 4 to use $\Lambda(s'|\pi, s)$ to denote the hitting time from state $s$ to $s'$. Since we only have one action, we suppress $\pi$ for brevity. Note that $\forall s \in S \setminus \{\frac{S}{2}\}$

$$
\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) \right] = 1 + \frac{1}{2} \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| (s+1) \mod S \right. \right) \right] + \frac{1}{2} \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| (s-1) \mod S \right. \right) \right],
$$

(45)

which indicate

$$
\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) \right] = (S-1) + \mathbb{E} \left[ \sum_{s \in S} \Lambda \left( \frac{S}{2} \left| s \right. \right) - \frac{1}{2} \Lambda \left( \frac{S}{2} \left| (s+1) \mod S \right. \right) - \frac{1}{2} \Lambda \left( \frac{S}{2} \left| (s-1) \mod S \right. \right) \right]
$$

$$
\Rightarrow \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) \right] = \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) \right] = (S-1),
$$

(46)

where we have made use of the fact that $\Lambda \left( \frac{S}{2} \left| \frac{S}{2} \right. \right) = 0$ and $\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| \frac{S}{2} + 1 \right. \right) \right] = \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| \frac{S}{2} - 1 \right. \right) \right]$. Rearranging terms in eq. (45), we also have $\forall s \in S \setminus \{\frac{S}{2}\}$

$$
\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) - \Lambda \left( \frac{S}{2} \left| (s-1) \mod S \right. \right) \right] = 2 + \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| (s+1) \mod S \right. \right) - \Lambda \left( \frac{S}{2} \left| s \right. \right) \right].
$$

(47)

That is to say $\forall s < \frac{S}{2}$,

$$
\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| (s+1) \mod S \right. \right) - \Lambda \left( \frac{S}{2} \left| s \right. \right) \right] = -(S-1) + 2 \left( \frac{S}{2} - s + 1 \right)
$$

(48)

and

$$
\mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| 1 \right. \right) \right] = \sum_{s=1}^{S/2-1} \mathbb{E} \left[ \Lambda \left( \frac{S}{2} \left| s \right. \right) - \Lambda \left( \frac{S}{2} \left| s+1 \right. \right) \right] = \sum_{s=1}^{S/2-1} (2s-3) = \frac{(S-6)(S-2)}{4}.
$$

(49)

D.2. Alternatives for Uniform Budget

We next propose a way to deviate from Assumption 2 by examining the case of inventory control (similar arguments can be made for queueing systems and beyond). In inventory control Bertsekas (2017), a seller decides the stock level (state) by ordering a quantity of a certain item (action) in each of $T$ rounds. For each round, a demand is realized (state transition), and unfulfilled demand incurs a shortage cost while excess inventory leads to a holding cost (reward). The seller’s objective is to minimize the cumulative shortage+holding cost. With some abuse of notations, we interpret state $s$ as the level of stock, i.e., in state $s$, the level of stock is $s$. Then as long as the demand distribution has at least $\zeta$ mass on each of the support $\{0,1,\ldots,S\}$, the DM could always choose to order $S-s$ quantity of the item so that with probability at least $\zeta$, the demand realization is $S-s'$. That is to say, $p_t(s'|s,S-s) \geq \zeta$ for any $t \in [T]$. Now it is easy to see that for any $t \in [T]$, we have the diameter of the time average MDP $(S, A, \bar{p}_t)$ is at most $\bar{D} = 1/\zeta$. In short, the system presented here can transition from states to states relatively easily. We formalize this observation as follows.
**Assumption 3.** There exists $\zeta > 0$ (not necessarily known to the DM), such that for any states $s, s' \in S$, there is always an action $a_{s,s'} \in A_s$ that satisfies $p_t(s'|s, a_{s,s'}) \geq \zeta$ for all $t \in [T]$.

Under this very mild assumption, the DM can then implement the SWUCRL2-CW algorithm with $\eta = 0$ as $\hat{p}_t \in H_{p,t}(0)$ for all $t \in [T]$ with high probability by Lemma [1]. To analyze its performance, we update Proposition [1] as follows.

**Proposition 4.** Conditioning on events $E_r, E_p$, for every episode $m$ and every time $t$ in the episode (i.e. $t \in \{\tau(m), \ldots, \tau(m+1) - 1\}$) the following inequality holds for certainty:

$$\rho_t^* - r_t(s_t, a_t) \leq \sum_{s' \in S} p_t(s'|s_t, a_t)\gamma_{r_t}(s_t) - \gamma_{r_t}(s_t) + \frac{1}{\sqrt{T}}$$

$$+ \left[2\text{var}_{r,t} + 4\bar{D}\text{var}_{p,t} + 2\text{rad}_{r_t}(s_t, a_t) + 4\bar{D} \cdot \text{rad}_{r_t}(s_t, a_t)\right].$$

The proof is very similar to that of Proposition [1] with $D_{r_t}$ and $\eta$ replaced by $\bar{D}$ and 0, respectively.

**Theorem 3.** Assuming $S > 1$, the SWUCRL2-CW algorithm with window size $W$ and $\eta = 0$ satisfies the dynamic regret bound

$$\text{Dyn-Reg}_r(\text{SWUCRL2-CW}) = \tilde{O}\left(B_r W + \bar{D}\left[B_p W + \frac{S \sqrt{A^1 T}}{W^1} + \frac{SAT}{W^2} + \sqrt{T}\right]\right)$$

Here, we set $\delta = 1/T$. If we further put $W = W^* = S^{2/3}A^{1/3}T^{2/3}/(B_r + B_p + 1)^{2/3}$ this is $\tilde{O}\left(\bar{D}(B_r + B_p + 1)^{1/3}S^{2/3}A^{1/3}T^{2/3}\right)$.

The proof is very similar to that of Theorem [1] and is thus omitted.

**E. Proof of Theorem 2**

By virtue of the EXP3.P, the BORL algorithm is able to adapt to any choice of $(W^1, \eta^1) \in J$. Specifically, we have from eq. [17] that for any choice of $(W^1, \eta^1) \in J$, the dynamic regret of the BORL algorithm is

$$\text{Dyn-Reg}_r(\text{BORL}) = \tilde{O}\left(B_r W^1 + D_{\max}\left[B_p W^1 + \frac{S \sqrt{A^1 T}}{W^1} + T\eta^1 + \frac{SAT}{W^2} + \sqrt{T\bar{H}}\right]\right)$$

$$= \tilde{O}\left(B_r W^1 + D_{\max}\left[B_p W^1 + \frac{S \sqrt{A^1 T}}{W^1} + T\eta^1 + \frac{SAT}{W^2} + S^{1/2}A^{1/2}T^{1/2}\right]\right).$$

Also note that

$$1 \geq \eta^1 = \frac{S^{2/3}A^{1/3}(B_r + B_p + 1)^{2/3}}{2T^{1/2}(B_r + B_p + 1)^{1/2}} \geq \frac{S^{2/3}A^{1/3}(B_r + B_p + 1)^{1/2}}{2T^{1/2}} = \Phi.$$ (51)

Therefore, there must exists some $k^1$ such that

$$\Phi^{k^1/\Delta_n} \geq \eta^1 \geq \Phi^{(k^1+1)/\Delta_n}.$$ (52)

By adapting $\eta^1$ to $\Phi^{k^1/\Delta_n}$, we further upper bound eq. [50] as

$$\text{Dyn-Reg}_r(\text{BORL})$$

$$= \tilde{O}\left(B_r W^1 + D_{\max}\left[B_p W^1 + \frac{S \sqrt{A^1 T}}{W^1} + T\eta^1 \Phi^{k^1/\Delta_n} + \frac{SAT}{W^2} + S^{1/2}A^{1/2}T^{1/2}\right]\right)$$

$$= \tilde{O}\left(B_r W^1 + D_{\max}\left[B_p W^1 + \frac{S \sqrt{A^1 T}}{W^1} + \frac{(B_r + B_p + 1)S^{2/3}A^{1/3}T^{1/2}}{W^1} + \frac{SAT}{W^2} + S^{1/2}A^{1/2}T^{1/2}\right]\right).$$ (53)

where we have made use of the fact that $\Phi^{k^1/\Delta_n} = \exp(1)$ in the last step.

We then turn to the choice of $W^1$ by distinguishing the following two cases:
\( W^* \leq H \) or \( (B_r + B_p + 1) \geq T^{1/4} / (S^{1/2} A^{1/4}) \): under this case, the BORL algorithm is able to adapt \( W^\dagger \) to some \( H^\dagger/\Delta W \) such that

\[
H^\dagger/\Delta W \leq W^* \leq H^{(\dagger+1)/\Delta W}.
\]

(54)

Note again that \( H^\dagger/\Delta W = O(1) \), the dynamic regret of the BORL algorithm is

\[
\text{Dyn-Reg}_T(\text{BORL}) = \tilde{O} \left( B_r W^* + D_{\max} \left[ B_p W^* + S\sqrt{AT} \sqrt{W^*} + (B_p + 1) S A^{1/4} T^{3/4} + SAT + S^{1/2} A^{1/4} T^{3/4} / W^* \right] \right) \\
= \tilde{O} \left( D_{\max} (B_r + B_p + 1)^{1/4} S A^{1/4} T^{3/4} + D_{\max} S A^{1/4} T^{3/4} \right).
\]

(55)

\( W^* > H \) or \( (B_r + B_p + 1) < T^{1/4} / (S^{1/2} A^{1/4}) \): under this case, the BORL algorithm is only able to adapt \( W^\dagger = H \) the dynamic regret of the BORL algorithm is

\[
\text{Dyn-Reg}_{ST}(\text{BORL}) = \tilde{O} \left( B_r H + D_{\max} \left[ B_p H + S\sqrt{AT} \sqrt{H} + (B_p + 1) S A^{1/4} T^{3/4} / H + SAT + S^{1/2} A^{1/4} T^{3/4} \right] \right) \\
= \tilde{O} \left( D_{\max} S A^{1/4} T^{3/4} \right). \]

(56)

Combining the above two cases, we conclude the proof.