IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 April 2025, accepted 4 June 2025, date of publication 9 June 2025, date of current version 16 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3577997

RESEARCH ARTICLE

Variation-Aware Bernstein-Based Upper Confidence Reinforcement Learning for Environment With Endogenous and Exogenous Uncertainty

RUOQI WEN AND RONGPENG LI, (Senior Member, IEEE)
College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China

Corresponding author: Rongpeng Li (lirongpeng@zju.edu.cn)

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LR23F010005, in part by the National Key Research and Development Program of China under Grant 2024YFE0200600, in part by the National Key Laboratory of Wireless Communications Foundation under Grant 2023KP01601, and in part by the Big Data and Intelligent Computing Key Laboratory of Chongqing University of Posts and Telecommunications (CQUPT) under Grant BDIC-2023-B-001.

ABSTRACT Online Reinforcement Learning (RL) has yielded remarkable performance in dynamic wireless communication and networks by interacting with the environment and gradually improving the effectiveness of its policy. As it is normal to witness much uncertainty in such an environment due to the intrinsic randomness of channels and service demands, designing a sample-efficient RL with bounded regrets has significant merits. In this paper, we focus on general Markov Decision Processes (MDPs) with time-evolving rewards and state transition probability unknown a priori and develop a Variation-aware Bernstein-based Upper Confidence Reinforcement Learning (VB-UCRL). In particular, we allow for restarting VB-UCRL according to a variation-aware schedule. We successfully overcome the challenges due to both endogenous and exogenous uncertainty and establish a regret bound of saving at most $\sqrt{S}$ or $S^{\frac{1}{6}}T^{\frac{1}{12}}$ compared with the latest results in the literature, where S denotes the size of the state space of the MDP and T indicates the iteration index of learning time-steps. Finally, we show via simulation that our algorithm VB-UCRL significantly outperforms the existing algorithms in the literature.

INDEX TERMS Reinforcement learning, regret bound, Markov decision process, endogenous, exogenous uncertainty.

I. INTRODUCTION

Online Reinforcement Learning (RL) has yielded remarkable performance in dynamic wireless communication and networks [1], [2], [3], [4], [5], [6], [7], [8] by interacting with the environment and gradually improving the effectiveness of its policy. Typically, on top of a formulated Markov Decision Process (MDP), an RL agent tries to maximize the cumulative rewards (or minimize the cumulative loss), by observing the environment as a state and taking an action through an "economic" perspective [9]. For example, [6] aims to optimize long-term utility by formulating a risk-sensitive MDP, where the state captures the environment dynamics and the reward is transformed through a utility function. An RL agent is then trained to make economically rational decisions under uncertainty, such as in portfolio management or recommendation systems. Meanwhile, [7] applies Deep RL (DRL) algorithms to optimize dynamic resource allocation in wireless networks and shows that DRL significantly outperforms traditional methods in adapting to learning rates and scheduling strategies. Note-worthily, in these scenarios, online RL emerges as a popular option.

The associate editor coordinating the review of this manuscript and approving it for publication was Usama Mir.

Therefore, it naturally raises a question what is the regret bound of online RL-based solutions regardless of the specific RL applications?

The difficulty in knowing this bound mainly lies in the *endogenous* and *exogenous* uncertainty in the MDP. Specifically, in the classical time-homogeneous MDP settings, only endogenous uncertainty is considered. In other words, at each time-step, the reward and the subsequent state follow a reward distribution and a state transition distribution, respectively, which solely depend on the current state and action and remain fixed along with the temporal variations. Unfortunately, in realistic environments like wireless networks with fluctuating channel conditions and dynamic service demands [8], both the reward functions and transition probabilities vary significantly over time. Therefore, the exogenous uncertainty has to be taken into account. Typically, in order to unveil the uncertainty in the MDP, the RL agent has to explore the MDP to accumulate the related knowledge of those poorly visited states and actions. As any decision of RL affects the subsequent observations, more exploration usually produces long-term impact yet affects short-term exploitation efficiency, which is also termed as the *exploration-exploitation dilemma* [10] originally discussed in the literature of Multi-Arm Bandit (MAB) [11].

There has been intense research interest in understanding the regret bound of online RL-based solutions for a time-homogeneous MDP. For example, [12] talks about the performance guarantees of a learned policy with polynomial scaling in the size of the state and action spaces, while Jaksch et al. give the regret bound of an RL algorithm during the learning [13], which is more meaningful for online RL in information processing scenarios. Specifically, Jaksch et al. propose a UCRL2 algorithm (*upper confidence bound for reinforcement learning*) for undiscounted reinforcement learning in communicating MDPs. In other words, UCRL2 implements the paradigm of ''optimism in the face of uncertainty'' by constructing plausible MDPs in confidence interval based on the Hoeffding inequality [14] and proves that the total regret of an RL algorithm concerning an optimal policy could be bounded by $\tilde{O}(DS\sqrt{AT})$, where $\tilde{\mathcal{O}}(\cdot)$ hides the logarithmic factors, S and A denote the size of the state space and action space of the MDP, respectively. D is the diameter of the communicating MDP, indicating the minimal expected number of time steps from each state to another state in the MDP. Besides, T denotes the iteration of learning time steps. Based on UCRL2, many variants have been proposed to generate tighter bounds. References [9] and [15] introduce UCRL2B, a variant of UCRL2 that refines the confidence bounds in the extended MDP using Bernstein's inequality, resulting in an improved regret bound of $\tilde{O}(\sqrt{DS\Gamma AT})$, indicating that the minimax lower bound is fairly tight, where Γ denotes the maximal number of reachable states for any state-action pair in the MDP. Reference [16] proposes a non-parametric and data-dependent algorithm based on the multiplier bootstrap for MAB. Later, [17] focuses on an infinite-horizon undiscounted setting and uses an exploration bonus to achieve the same regret bound as UCCRL. Reference [18] introduces the UCRL3 algorithm, an enhancement of UCRL2 that achieves improved regret bounds of $\mathcal{O}\left((D+\sqrt{\sum_{s,a}(D_s^2 L_{s,a}\vee 1)})\sqrt{T\log(T/\delta)}\right)$ by incorporating state-of-the-art concentration inequalities and adaptive exploration techniques, with $L_{s,a}$ representing the local effective support of $p(\cdot|s,a)$. In addition to UCRL-based variants, there are many studies related to regret bound analyses. For example, [19] proposes the first model-free and simulator-free algorithm for constrained MDPs that achieves sublinear regret bounds of $\tilde{O}(\sqrt{d^3H^4K})$, where d, H, and K denote the feature dimension, episode length, and number of episodes, respectively. Reference [20] investigates regret minimization in episodic MDPs with unknown transitions and adversarially delayed feedback, and proposes policy optimization algorithms that achieve near-optimal regret bounds depending on the number of episodes and the total delay.

Until recently, little light has been shed on MDP with both endogenous and exogenous uncertainty. Reference [8] proposes PUCRL2, PUCRLB, and their extensions for unknown periods, addressing the problem of average reward maximization in periodic MDP, and derives theoretical regret bounds with respect to both the period and the time horizon. Reference [21] proposes an algorithm for the non-stationary stochastic multi-armed bandit problem, where the reward distributions may change multiple times during learning, and shows that it achieves a near-optimal dynamic regret bound of $\tilde{O}(\sqrt{KN(S+1)})$ over a time horizon N without requiring prior knowledge of the number of optimal arm switches S. Reference [22] investigates online convex optimization in non-stationary environments and proposes a two-layer collaborative framework that achieves adaptive dynamic regret using only one gradient query per iteration. Reference [23] and [24] talk about online learning for MDP in this non-stationary environment and provide the dynamic regret analysis for exogenous uncertainty only. Based on the Hoeffding inequality, [25] develops a variation-aware UCRL2 algorithm and provides performance guarantees for the regret evaluated against the optimal non-stationary policy. Besides, [26] and [27] discuss the dynamic regret more comprehensively, and derive a bound of $\tilde{O}\left((V_r^T+V_p^T)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$, where V_r^T and V_p^T are the dynamic budget (i.e., upper bound) of variations in reward and transition probability functions for T time-steps. Reference [28] also unveils a $\tilde{O}\left((V_r^T+V_p^T)^{1/3}T^{2/3}\right)$ bound by a closed-box approach under conditions that either the diameter D or the total variation is known. However, most existing work on analyzing algorithm regret bounds lacks experimental simulation evidence. Compared with the abovementioned research, this paper's contribution can be summarized as follows.

- We focus on the RL for MDP with both endogenous and exogenous uncertainty, which has significant applications in information processing scenarios. In particular, we talk about a Variation-aware Bernstein-based Upper Confidence Reinforcement Learning (VB-UCRL) algorithm, which restarts according to a schedule dependent on the variations in the MDP and leverages the empirical Bernstein inequality [29] to give a tighter bound. Notably, compared to the work [25], the empirical Bernstein inequality could additionally capture second-order statistics.
- We prove that the VB-UCRL gives a regret bound of $\tilde{O}\left((V_r^T + V_p^T)^{1/3} T^{2/3} \sqrt{\Gamma SA}\right)$, where Γ denotes the maximal number of reachable states for any state-action pair in the MDP. As discussed in Section IV-C1, for $\Gamma < A$, this bound is tighter than the classical results in [25] and [27].
- Our simulation results show that our algorithm VB-UCRL significantly outperforms other classical MDP algorithms under a non-stationary environment when the number of reachable states Γ is considerably less than the size of the state space S, and is still comparable with existing algorithms when Γ approaches S.

The remainder of the paper is organized as follows. In Section V, we introduce some fundamentals of MDPs and formulate the regret problem of RL for MDPs with both endogenous and exogenous uncertainty, in Section III and Section IV, we provide the VB-UCRL algorithm and prove its upper confidence bound. Meanwhile, the theoretical comparison with the state-of-the-art results is also presented. In Section V, we demonstrate the performance of VB-UCRL through extensive simulations. We conclude the paper in Section VI.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. SYSTEM MODEL

In a time-homogeneous MDP $M =< \mathcal{S}, \mathcal{A}, r, p, s_1 >$ with state space $\mathcal{S}$, action space $\mathcal{A}$ and the initial state s_1. For simplicity of representation, the size of state and action space is denoted as $S = |\mathcal{S}|$ and $A = |\mathcal{A}|$, respectively. In the time-homogeneous MDP, the mean rewards and transition probabilities are only relevant to the current state and the chosen action. Every state-action pair is characterized by a reward distribution with mean $r(s,a) \in [0, r_{\max}]$ over the next states, while the number of reachable states for a state-action pair (s,a) is formulated as $\Gamma(s,a) = \|p(\cdot|s,a) > 0\|_0$ (where $\|\cdot\|_0$ denotes an l_0 norm of a vector) and $\Gamma = \max_{s,a} \Gamma(s,a)$. Besides, for an *communicating* MDP, starting in $s \in \mathcal{S}$ it is possible to reach another state $s' \in \mathcal{S}$ with positive probability, choosing appropriate actions.

Compared with the time-homogeneous MDP, we consider a time-heterogeneous MDP with time-step-dependent mean rewards and transition probabilities, that is, $r_t(s,a)$ and $p_t(s'|s,a)$ respectively. Accordingly, the time-heterogeneous MDP at time-step t can be written as $M_t =< \mathcal{S}, \mathcal{A}, r_t, p_t, s_1 >$. All MDPs M_t are communicating with diameter $D_t \le D$, where D denotes a common upper bound. Furthermore, we assume that the *variations* in mean rewards and transition probabilities are bounded in the T time-steps, that is, $V_r^T \overset{\text{def}}{=} \sum_{t=1}^{T-1} \max_{s,a} |r_{t+1}(s,a) - r_t(s,a)|$, and $V_p^T \overset{\text{def}}{=} \sum_{t=1}^{T-1} \max_{s,a} \|p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a)\|_1$, where $\|\cdot\|_1$ denotes an l_1 norm of a vector.

We primarily focus on the infinite-horizon undiscounted MDP settings and try to learn a policy π that maximizes

$$\sup_{\pi \in \Pi} \left\{ \liminf_{T \to +\infty} \mathbb{E}_\pi \left[\frac{1}{T} \sum_{t=1}^{T} r(s_t, a_t) \middle| s_1 \sim \mu_1 \right] \right\} \tag{1}$$

where μ_1 is the state probability of the starting state s_1. $\mathbb{E}_\pi$ takes an expectation over trajectories with action $a_t \sim \pi(s_t)$. For any Markov decision rule d, which maps states to distributions over actions, the transition matrix $P_d(s'|s) \overset{\text{def}}{=} \sum_{a \in \mathcal{A}_s} d(a|s) p(s'|s,a) \in \mathbb{R}^{S \times S}$ and the associated reward vector $r_d(s) \overset{\text{def}}{=} \sum_{a \in \mathcal{A}_s} d(a|s) r(s,a) \in \mathbb{R}^S$, for all $s \in \mathcal{S}$, where $d(a|s)$ is the probability to sample a in state s when using d. Furthermore, the decision rule could be repetitively updated along with the exploration progress of the environment. We can approximately obtain a stationary policy π as the limit of $d \in D^{\text{MR}}$ (i.e., $(d)^\infty$) when the decision rule $d \in D^{\text{MR}}$ under an RL algorithm $\mathfrak{A}$ involves no further improvements. Furthermore, Chapter 9 of [30] and Theorem 1 of [31] verify the existence of an optimal stationary policy satisfying the Bellman evaluation equation under some conditions.

B. PROBLEM FORMULATION

Let M_t be the true time-heterogeneous MDP. We consider the learning problem where $\mathcal{S}$, $\mathcal{A}$, and $r_{\max}$ are known, while rewards r_t and transitions p_t are required to be estimated online and time-evolving. However, we assume that the variations of rewards and transitions in the T time-steps are bounded and known as V_r^T and V_p^T, repsectively. We aim to develop a learning algorithm $\mathfrak{A}$ with appropriate policy π to minimize its cumulative regret in T time-steps as

$$\arg\min_{\mathfrak{A}} \Delta(\mathfrak{A}, T) \tag{2}$$

where $\Delta(\mathfrak{A}, T) \overset{\text{def}}{=} v^{*,T}(s_1) - \sum_{t=1}^{T} r_t(s_t, a_t)$ and $v^{*,T}(s_1)$ denotes the optimal T-time-step average reward starting from s_1.[1]

III. THE VB-UCRL ALGORITHM

For the changing MDP settings, we introduce an RL algorithm VB-UCRL, on top of UCRL2 [13]. Notably, VB-UCRL implements the paradigm of "*optimism in the face of uncertainty*" and constructs MDPs in confidence interval based on the empirical Bernstein inequality (Theorem 1, [29])

[1] Interesting readers could refer to Page 338 of [30] for the relationship between v and h.

rather than the Hoeffding inequality for UCRL2 in Theorem 2.8 of [32].

To tackle the exploration-exploitation dilemma, VB-UCRL proceeds through episodes $k = 1, 2, \cdots$, each episode consisting of multiple time-steps. Without loss of generality, t_k is the starting time of episode k. $N_k(s,a)$ is the number of visits in (s,a) before episode k. Here, consistent with the doubling criterion in [13], VB-UCRL enters into a new episode $k+1$ once there exists one state-action pair (s,a) having just been played satisfies $v_k(s,a) = N_k^+(s,a)$, where $v_k(s,a)$ (resp. $v_k(s)$) denotes the number of visits to (s,a) (resp. s) in episode k and $N_k^+(s,a) = \max\{1, N_k(s,a)\}$. For episode $k+1$, for all state-action pairs, $N_{k+1}(s,a) = N_k(s,a) + v_k(s,a)$. Besides, t_k is defined as the starting time of episode k, that is, $t_{k+1} \overset{\text{def}}{=} \inf\Big\{T \geq t > t_k : \sum_{\tau=1}^{t-1} \mathbb{1}\{(s_\tau, a_\tau) = (s_t, a_t)\} \geq \max\big\{1, 2\sum_{\tau=1}^{t_k-1} \mathbb{1}\{(s_\tau, a_\tau) = (s_t, a_t)\}\big\}\Big\}$ and $t_1 = 1$. During this episode-driven procedure, it remains essential to learn a policy π_k to correspondingly determine the taken action a_t at the state s_t for $t_k \leq t < t_{k+1}$. In this regard, VB-UCRL first constructs a set of plausible MDPs for each episode, and then derives the policy for an optimistic MDP therein via an extended value iteration (EVI).

A. THE CONSTRUCTION OF THE SET OF PLAUSIBLE MDPS

At the beginning of each episode k, VB-UCRL computes a set $\mathcal{M}$ of statistically plausible MDPs given the observations so far, that is,

$$\mathcal{M}_k \overset{\text{def}}{=} \Big\{M =< \mathcal{S}, \mathcal{A}, \tilde{r}, \tilde{p} >: \tilde{r}(s,a) \in \mathcal{B}_{r,k}(s,a), \tilde{p}(s'|s,a) \in \mathcal{B}_{p,k}(s,a,s'), \sum_{s'} \tilde{p}(s'|s,a) = 1\Big\}, \quad (3)$$

where $\mathcal{B}_{r,k}$ and $\mathcal{B}_{p,k}$ are high-probability (adapted) confidence intervals on the rewards and transition probabilities of the true MDP M. Specifically,

$$\mathcal{B}_{p,k}(s,a,s') \overset{\text{def}}{=} [0,1] \cap \Big[\hat{p}_k(s'|s,a) - \beta_{p,k}^{sas'} - \hat{V}_p, \hat{p}_k(s'|s,a) + \beta_{p,k}^{sas'} + \hat{V}_p\Big] \quad (4)$$

where $\hat{p}_k(s'|s,a)$ is set as an estimate of transitions corresponding to the sample mean of an independent identical Bernoulli random variable with mean $p(s'|s,a)$, that is,

$$\hat{p}_k(s'|s,a) = \frac{1}{N_k^+(s,a)} \sum_{t=1}^{t_k-1} \mathbb{1}\{(s_t, a_t, s_{t+1}) = (s,a,s')\} \quad (5)$$

and $\hat{V}_p \leq V_p^T$ is an estimate of the variations on the transition probabilities. Moreover, using the empirical Bernstein inequality [29], we can derive partial confidence intervals [9] for the transition probabilities of the true MDP M,

$$\beta_{p,k}^{sas'} \overset{\text{def}}{=} 2\sqrt{\frac{\hat{\sigma}_{p,k}^2(s'|s,a)}{N_k^+(s,a)} \ln\Big(\frac{6SAN_k^+(s,a)}{\delta}\Big)} + \frac{6\ln\big(\frac{6SAN_k^+(s,a)}{\delta}\big)}{N_k^+(s,a)} \quad (6)$$

where $\delta \in (0,1)$ and the transition probability's population variance $\hat{\sigma}_{p,k}^2(s'|s,a)$ can be approximately computed as $\hat{\sigma}_{p,k}^2(s'|s,a) = \hat{p}_k(s'|s,a)(1 - \hat{p}_k(s'|s,a))$.

Similarly,

$$\mathcal{B}_{r,k}(s,a) \overset{\text{def}}{=} \Big[\hat{r}_k(s,a) - \beta_{r,k}^{sa} - \hat{V}_r, \hat{r}_k(s,a) + \beta_{r,k}^{sa} + \hat{V}_r\Big] \cap [0, r_{\max}] \quad (7)$$

where $\hat{r}_k$ is the empirical average of rewards, namely

$$\hat{r}_k(s,a) = \frac{1}{N_k^+(s,a)} \sum_{t=1}^{t_k-1} \mathbb{1}\{(s_t, a_t) = (s,a)\} \cdot r_t. \quad (8)$$

and $\hat{V}_r \leq V_r^T$ is an estimate of the variations on the mean rewards.

$$\beta_{r,k}^{sa} \overset{\text{def}}{=} 2\sqrt{\frac{\hat{\sigma}_{r,k}^2(s,a)}{N_k^+(s,a)} \ln\Big(\frac{6SAN_k^+(s,a)}{\delta}\Big)} + \frac{6r_{\max}\ln\big(\frac{6SAN_k^+(s,a)}{\delta}\big)}{N_k^+(s,a)} \quad (9)$$

where the reward's population variance $\hat{\sigma}_{r,k}^2(s,a)$ can be computed recursively at the end of every episode as

$$\begin{aligned}\hat{\sigma}_{r,k+1}^2(s,a) &\overset{\text{def}}{=} \frac{1}{N_{k+1}^+(s,a)}\Big(\sum_{l=1}^{k+1} S_l(s,a)\Big) - \big(\hat{r}_{k+1}(s,a)\big)^2 \\ &= \frac{N_k(s,a)}{N_{k+1}^+(s,a)}\Big(\hat{\sigma}_{r,k}^2(s,a) + \big(\hat{r}_k(s,a)\big)^2\Big) + \frac{S_{k+1}(s,a)}{N_{k+1}^+(s,a)} - \big(\hat{r}_{k+1}(s,a)\big)^2 \end{aligned} \quad (10)$$

with $S_k(s,a) \overset{\text{def}}{=} \sum_{t=t_{k-1}+1}^{t_k-1} \mathbb{1}\{(s_t, a_t) = (s,a)\} \cdot r_t^2$.

Furthermore, as pointed out by Section 3.1.1 of [13], any bounded parameter MDP can be equivalently represented by an *extended MDP*, which is defined as

$$M_k^+ =< \mathcal{S}, \mathcal{A}, \tilde{r}, \tilde{p}, s_1 > \quad (11)$$

In other words, the extended MDP combines all plausible MDPs constructed above into a single MDP with identical state and action space. Therefore, we use the terminology *extend MDP* and *the set of plausible MDPs* interchangeably. Besides, Theorem 3.1 of [9] proves that the true MDP M falls into the set of plausible MDPs $\mathcal{M}_k$ with a high probability.

B. THE EVI-BASED POLICY CALCULATION

Beforehand, some essential operators for the EVI are provided. Recalling Theorem 9.4.5 of [30] and Theorem 7 of [13], for EVI with aperiodic transition matrices P_{d_n} ($n \geq 1$) there exists $h^* \in \mathbb{R}^S$ such that the limit of the value function $\lim_{n\to\infty} v_n = h^*$ (where v is defined using operators $L : \mathbb{R}^S \to \mathbb{R}^S$ as $Lv(s) \overset{\text{def}}{=} \max_{d\in D^{\text{MR}}}\{r_d + P_d v\}$) and $Lh^*(s) = h^*(s) + g^*(s)$, where $g^\pi(s) = \lim_{T\to+\infty} \mathbb{E}_\pi \left[\frac{1}{T}\sum_{t=1}^T r(s_t, a_t) \middle| s_1 = s\right]$ and $h(s) \overset{\text{def}}{=} \underset{T\to+\infty}{C-\lim} \mathbb{E}_\pi \left[\sum_{t=1}^T \left(r(s_t, a_t) - g^\pi(s_t)\right) \middle| s_1 = s\right]$ defines the associated *long-term average reward* (or gain) and *bias function*, respectively. Furthermore, the bias $h^\pi(s)$ measures the expected total difference between the reward and the long-term average reward in *Cesaro-limit* (denoted by $C-\lim$).

Based on the definition of long-term average reward and bias functions, an optimal Bellman operator L_k for the extended MDP can be defined as

$$L_k h_k(s) \overset{\text{def}}{=} \max_{a\in\mathcal{A}_s}\{r(s,a) + \sum_{s'\in\mathcal{S}} p(s'|s,a)h_k(s')\} \tag{12}$$

where $r(s,a) \in \mathcal{B}_{r,k}(s,a)$, $p(s'|s,a) \in \mathcal{B}_{p,k}(s,a,s')$, and $\mathcal{A}_s$ denotes the action sub-space under state s. Besides, an extended decision rule $d_k \in D^{\text{MR}}$ can be simultaneously obtained during the maximization of the Bellman equation. Notably, for all $s \in \mathcal{S}$, we can further define the extended Bellman operator with aperiodic transformation (Proposition 8.5.8, [30]) as

$$\begin{aligned} &L_k^\alpha h_k(s) \\ &= \max_{a\in\mathcal{A}_s}\{r(s,a) + \alpha \sum_{s'\in\mathcal{S}} p(s'|s,a)h_k(s')\} + (1-\alpha)\cdot h_k(s) \end{aligned} \tag{13}$$

where α is the coefficient of the aperiodic transformation. One benefit of the aperiodic transformation lies in that, as shown by Prop. 8.5.8 of [30], it does not affect the gain of any stationary policy. In other words, for any $\pi \in \Pi^{\text{SR}}$, $g^{\alpha,\pi} = g^\pi$. Since the aperiodic transformed MDP M^α exactly meets the conditions in [13] and [30], the EVI in Alg. 1 is feasible while the Bellman operator is optimal. Notably, by Prop. 2.7 of [9], if we run the EVI in Alg. 1 on M_k^+ with accuracy $\epsilon_k = r_{\max}/t_k$, we have that

$$|g_k(s) - g_k^*(s)| \leq \epsilon_k/2 = \frac{r_{\max}}{2t_k} \tag{14}$$

and

$$\|L_k^\alpha h_k(s) - h_k(s) - g_k(s)\|_\infty \leq \epsilon_k = \frac{r_{\max}}{t_k} \tag{15}$$

where $(g_k, h_k, \pi_k) = \text{EVI}(L_k^\alpha, \frac{r_{\max}}{t_k}, 0, s_1)$ and $\|\cdot\|_\infty$ denotes an infinity norm of a vector.

Equivalently, VB-UCRL chooses an optimistic MDP M_k (concerning the achievable average reward) among these plausible MDPs $\mathcal{M}_k$, and executes a policy π_k which is (nearly) optimal for the optimistic MDP M_k, that is,

$$\max_{\pi\in\Pi^{\text{SD}}}\left\{\sup_{M'\in\mathcal{M}_k} g_{M'}^\pi\right\} = \sup_{M'\in\mathcal{M}_k}\left\{\max_{\pi\in\Pi^{\text{SD}}} g_{M'}^\pi\right\} = \sup_{M'\in\mathcal{M}_k} g_{M'}^* \tag{16}$$

Furthermore, $r_k(s,a)$ and $p_k(s'|s,a)$ denote the optimistic reward and state transition probability for M_k at episode k.

We first summarize the VB-UCRL without variation-aware restarts as Alg. 2. On top of that, to simultaneously tackle the endogenous and exogenous uncertainty, we can formally give VB-UCRL in Alg. 3. In particular, we restart Alg. 2 in phases by continuously tuning the confidence parameter $\frac{\delta}{2t^2}$ according to a schedule dependent on the variations.

IV. REGRET BOUNDS OF VB-UCRL

In this section, we first derive the upper regret bound of VB-UCRL without variation-aware restarts and then extend it to VB-UCRL with restarts.

A. UPPER REGRET BOUND OF VB-UCRL WITHOUT VARIATION-AWARE RESTARTS

The following theorem gives the limits of regret bound in (2) for VB-UCRL without variation-aware restarts.[2]

Theorem 1: For any communicating MDP, if $\hat{V}_p$ and $\hat{V}_r$ are set as the true values V_p^T and V_r^T, with probability at least $1-\delta$, it holds that for all initial state distributions $\nu_1 \in \Delta_s$ (Δ_s denotes a S-dimensional simplex.) and for all time horizons $T \geq SA$

$$\begin{aligned} &\Delta(\textit{VB-UCRL}, T) \\ &\leq \max(r_{\max}, Dr_{\max})\Bigg(43\sqrt{T\ln\left(\frac{T}{\delta}\right)\sum_{s,a}\Gamma(s,a)} \\ &\quad + 72S^2A\ln\left(\frac{T}{\delta}\right)\ln(T)\Bigg) + Dr_{\max}TV_p^T + 2TV_r^T \end{aligned} \tag{17}$$

where $\Gamma(s,a) \overset{def}{=} \|p(\cdot|s,a)\|_0 = \sum_{s'\in\mathcal{S}} \mathbb{1}\{p(s'|s,a) > 0\}$ and $\Gamma \overset{def}{=} \max_{s,a\in\mathcal{S}\times\mathcal{A}} \Gamma(s,a)$.

Proof: By Lemma 10 of [25], which shows that under the event where the true MDP falls into the scope of plausible MDPs ($M \in \mathcal{M}_k, \forall k$), the T-time-step reward in the changing MDP settings could be bounded by the optimistic average reward g^* (for all k and all s, $v^{*,T}(s) \leq Tg_k^*(s) + D$ where $g_k^* \overset{\text{def}}{=} \max_{\pi,M\in\mathcal{M}_k} g_k^*(M)$.). So, we have

$$\begin{aligned} \Delta(\text{VB-UCRL}, T) &= v^{*,T}(s_1) - \sum_{t=1}^T r_t(s_t, a_t) \\ &\leq \sum_{t=1}^T \left(g^*(s_t) - r_t(s_t, a_t)\right) + Dr_{\max} \end{aligned} \tag{18}$$

[2]For simplicity of representation, in this part, we slightly abuse the notations for VB-UCRL with and without variation-aware restarts.

Algorithm 1 Extended Value Iteration

Input: Operators $L : \mathbb{R}^S \to \mathbb{R}^S$, accuracy $\epsilon \in (0, r_{\max})$, value iteration (VI) record $v_0(s) \in \mathbb{R}^S$ as $v_0(s) = 0$ for all $s \in \mathcal{S}$, reference state $\bar{s} \in \mathcal{S}$,

Initialize $n = 0$, $v_1 \stackrel{\text{def}}{=} Lv_0$
while $\max_{s\in\mathcal{S}} \{v_{n+1}(s) - v_n(s)\} - \min_{s\in\mathcal{S}} \{v_{n+1}(s) - v_n(s)\} > \epsilon$ **do**
  Increment $n \leftarrow n + 1$.
  Update $v_n \leftarrow v_n - v_n(\bar{s})e$.
  For $s \in \mathcal{S}$, calculate VI record $v_{n+1}(s)$ from $Lv_n(s)$.
  Compute stationary policy $\pi(s) = \text{argmax}_{a\in\mathcal{A}_s}(Lv_n)$.
end while
Set $g \stackrel{\text{def}}{=} \frac{1}{2}(\max_{s\in\mathcal{S}} \{v_{n+1}(s) - v_n(s)\} + \min_{s\in\mathcal{S}} \{v_{n+1}(s) - v_n(s)\})$, $h \stackrel{\text{def}}{=} v_n$.

Output: Gain $g \in [0, r_{\max}]$, bias vector $h \in \mathbb{R}^S$ and stationary policy π.

Algorithm 2 VB-UCRL Without Variation-Aware Restarts

Input: Confidence $\delta \in (0, 1)$, $r_{\max}$, $\mathcal{S}$, $\mathcal{A}^+$.

Initialize $t \stackrel{\text{def}}{=} 1$, s_1 and for (s, a, s'): $N_1(s, a) = 0$, $\hat{p}_1(s'|s, a) = 0$, $\hat{r}_1(s, a) = 0$, $\hat{\sigma}^2_{p,1}(s'|s, a) = 0$, $\hat{\sigma}^2_{r,1}(s, a) = 0$.
for episodes $k = 1, 2, \ldots$ **do**
  Set $t_k \leftarrow t$ and episode counters $v_k(s, a) \leftarrow 0$.
  Compute the upper-confidence bounds (4) and (7) and the extended MDP M^+_k as in (11).
  Compute an $r_{\max}/t_k$-approximation π_k of (16) $(g_k, h_k, \pi_k) = \text{EVI}\left(L^\alpha_k, \frac{r_{\max}}{t_k}, 0, s_1\right)$.
  Sample action $a_t \sim \pi_k(\cdot|s_t)$.
  while $t_k = t$ or $v_k(s_t, a_t) \le \max\{1, N_k(s_t, a_t)\}$ **do**
    Execute a_t, obtain reward r_t, and observe s_{t+1}.
    Sample action $a_{t+1} \sim \pi_k(\cdot|s_{t+1})$.
    Set $v_k(s_t, a_t) \leftarrow v_k(s_t, a_t) + 1$ and set $t \leftarrow t + 1$.
  end while
  Set $N_{k+1}(s, a) \leftarrow N_k(s, a) + v_k(s, a)$.
  Update statistics i.e., $\left(\hat{p}_{k+1}, \hat{r}_{k+1}, \hat{\sigma}^2_{p,k+1}, \hat{\sigma}^2_{r,k+1}\right)$.
end for

Algorithm 3 VB-UCRL With Restarts

Input: State space $\mathcal{S}$, action space $\mathcal{A}$, confidence parameter δ, variation terms V^T_r and V^T_p.

Initialization: Set current time-step $\tau \stackrel{\text{def}}{=} 1$.
for phase $i = 1, 2, \ldots$ **do**
  Perform VB-UCRL in Algorithm 2 with confidence parameter $\delta/2\tau^2$ for $\theta_i \stackrel{\text{def}}{=} \lceil \frac{i^2}{(2V^T_r + V^T_p)^2} \rceil$ time-steps.
  Update $\tau \leftarrow \tau + \theta_i$.
end for

where $g^* \stackrel{\text{def}}{=} \min_k g^*_k$, and $g^*_k \stackrel{\text{def}}{=} \max_{\pi, M\in\mathcal{M}_k} g^*_k(M)$.

By Lemma 1, which can be interpreted as removing all the randomness due to the stochasticity of the observed rewards and the executed policy, at the expense of $\tilde{\mathcal{O}}(\sqrt{T})$, with a probability at least $1 - \frac{\delta}{6}$, $\Delta(\text{VB-UCRL}, T)$ could be rewritten as

$$
\begin{aligned}
&\Delta(\text{VB-UCRL}, T) \\
&\le \sum_{t=1}^{T} \left(g^*(s_t) - r_t(s_t, a_t)\right) \\
&\le \sum_{t=1}^{T} \left(g^*(s_t) - \sum_{a\in A_{s_t}} \pi_{k_t}(s_t, a) r(s_t, a)\right) \\
&\quad + 2r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} \\
&= \sum_{t=1}^{k_T} \sum_{s\in\mathcal{S}} v_k(s) \left(g^*(s) - \sum_{a\in A_s} \pi_k(a|s) r(s, a)\right) \\
&\quad + 2r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} \\
&\stackrel{(a)}{=} \sum_{t=1}^{k_T} \Delta_k + 2r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} \qquad (19)
\end{aligned}
$$

where the equation (a) comes after $\Delta_k \stackrel{\text{def}}{=} \sum_{s\in\mathcal{S}} v_k(s)\left(g^*(s) - \sum_{a\in A_s} \pi_k(a|s) r(s, a)\right)$, and $k_t \stackrel{\text{def}}{=} \sup\{k \ge 1 : t \ge t_k\}$ denotes the integer-valued random variable indexing the current episode at time-step t. By Prop. 18 of [13], $k_T \le SA \log_2\left(\frac{8T}{SA}\right)$ is bounded for $T \ge SA$.

Next, we derive the bound for Δ_k with a high probability. By Lemma 2 in Appendix, if $M \in \mathcal{M}_k, \forall k$, Δ_k could be upper bounded by

$$
\Delta_k \le \Delta^p_k + \Delta^r_k + \frac{3\epsilon_k}{2} \sum_{s\in\mathcal{S}} v_k(s) \qquad (20)
$$

where $\Delta^p_k \stackrel{\text{def}}{=} \alpha \sum_{s\in\mathcal{S}} v_k(s) \left(\sum_{\substack{a\in\mathcal{A}_s \\ s'\in\mathcal{S}}} \pi_k(a|s) p_k(s'|s, a) h_k(s') - h_k(s) \right)$ and $\Delta^r_k \stackrel{\text{def}}{=} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} v_k(s) \pi_k(a|s) \left(r_k(s, a) - r(s, a)\right)$.

We further decompose Δ^p_k into two parts $\Delta^p_k = \Delta^{p1}_k + \Delta^{p2}_k$, where $\Delta^{p1}_k \stackrel{\text{def}}{=} \alpha \sum_{s,a,s'} v_k(s) \pi_k(a|s) \left(p_k(s'|s, a) - p(s'|s, a) \right) h_k(s')$ and $\Delta^{p2}_k \stackrel{\text{def}}{=} \alpha \sum_s v_k(s) \left(\sum_{a,s'} \pi_k(a|s) p(s'|s, a) \cdot \right.$

$h_k(s') - h_k(s)\Big)$, and bound them in Lemma 3, Lemma 4, and Lemma 5 of Appendix. Accordingly, with a probability $1-\frac{\delta}{3}$,

$$\begin{aligned}\sum_{k=1}^{k_T}\Delta_k^p &\le Dr_{\max}\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)(\beta_{p,k}^{sa}+V_p^T)\\ &\quad+6Dr_{\max}\sqrt{T\ln\left(\frac{6T}{\delta}\right)}+k_TDr_{\max}\\ &\le Dr_{\max}\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa}+Dr_{\max}TV_p^T\\ &\quad+6Dr_{\max}\sqrt{T\ln\left(\frac{6T}{\delta}\right)}+k_TDr_{\max}\end{aligned}\tag{21}$$

where $\beta_{p,k}^{sa}\overset{\text{def}}{=}\sum_{s'}\beta_{p,k}^{sas'}$. Similarly, by Lemma 6 of Appendix, with probability at least $1-\frac{\delta}{6}$, we have

$$\begin{aligned}&\sum_{k=1}^{k_T}\Delta_k^r\\ &\le 4\,r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)}+2\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa}+2TV_r^T\end{aligned}\tag{22}$$

As proved in Theorem 3.1 of [9], the event that $M\in\mathcal{M}_k,\forall k$ occurs with a probability at least $1-\frac{\delta}{3}$. Merging (20), (21), (22) into (19), with a probability at least $1-\frac{5\delta}{6}$, for all $T\ge SA$, we have (23), as shown at the bottom of the next page.

As for the last three terms (i.e., ψ_1, ψ_2 and ψ_3) in (23), we have

- Since $t_k\ge N_k^+(s,a)$ for all (s,a),

$$\begin{aligned}\psi_1&=r_{\max}\sum_{k=1}^{k_T}\frac{3}{2t_k}\sum_s\nu_k(s)=\frac{3r_{\max}}{2}\sum_{s,a}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{t_k}\\ &\le\frac{3r_{\max}}{2}\sum_{s,a}\sum_{k=1}^{k_T}\frac{\nu_k(s,a)}{N_k^+(s,a)}\\ &\overset{(a)}{\le}\frac{3r_{\max}}{2}\sum_{s,a}2+2\ln\left(N_{k_T+1}^+(s,a)\right)\\ &\overset{(b)}{\le}\frac{3r_{\max}}{2}SA\left(2+2\ln\left(\frac{\sum_{s,a}N_{k_T+1}^+(s,a)}{SA}\right)\right)\\ &\overset{(c)}{\le}\frac{3r_{\max}SA}{2}\left(2+2\ln\left(\frac{T}{SA}\right)\right)\\ &\le 3r_{\max}SA\left(1+\ln T\right)\end{aligned}\tag{24}$$

  where the equation (a) comes from Prop. 1 in Appendix, while the inequality (b) leverages the concavity of a logarithmic function and the Jensen inequality. The equation (c) is due to that $\sum_{s,a}N_{k+1}^+(s,a)\le T$.
- Taking account of the definition of $\beta_{r,k}^{sa}$,

$$\begin{aligned}\psi_2&=2\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{r,k}^{sa}\\ &=4\sum_{k=1}^{k_T}\sum_{s,a}\left[\nu_k(s,a)\sqrt{\frac{\hat\sigma_{r,k}^2(s,a)}{N_k^+(s,a)}\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)}\right.\\ &\quad\left.+3r_{\max}\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)\frac{\nu_k(s,a)}{N_k^+(s,a)}\right]\\ &\overset{(a)}{\le}4\,r_{\max}\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sum_{k=1}^{k_T}\sum_{s,a}\left[\frac{\nu_k(s,a)}{\sqrt{N_k^+(s,a)}}\right]\\ &\quad+12r_{\max}\ln\left(\frac{6SAT}{\delta}\right)\sum_{k=1}^{k_T}\sum_{s,a}\frac{\nu_k(s,a)}{N_k^+(s,a)}\\ &\overset{(b)}{\le}4\,r_{\max}\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sum_{s,a}3\sqrt{N_{k_T+1}^+(s,a)}\\ &\quad+12r_{\max}\ln\left(\frac{6SAT}{\delta}\right)\sum_{s,a}\left[2+2\ln\left(N_{k_T+1}^+(s,a)\right)\right]\\ &\overset{(c)}{\le}12\,r_{\max}\sqrt{SAT\ln\left(\frac{6SAT}{\delta}\right)}\\ &\quad+24\,r_{\max}SA\ln\left(\frac{6SAT}{\delta}\right)(1+\ln T)\end{aligned}$$

  where the equation (a) comes from $\hat\sigma_{r,k}^2(s,a)\le r_{\max}^2$ and $\ln\left(\frac{6SAN_k^+(s,a)}{\delta}\right)\le\ln\left(\frac{6SAT}{\delta}\right)$, the inequality (b) comes from Prop. 1, and the inequality (c) comes from similar deduction as the previous term.
- Similarly, in terms of the definition of $\beta_{p,k}^{sa}$, by applying Prop. 2 and Cauchy-Schwartz inequality, we have

$$\begin{aligned}\psi_3&=Dr_{\max}\sum_{k=1}^{k_T}\sum_{s,a}\nu_k(s,a)\beta_{p,k}^{sa}\\ &\le 6Dr_{\max}\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sqrt{\left(\sum_{s,a}\Gamma(s,a)\right)T}\\ &\quad+12Dr_{\max}S^2A\ln\left(\frac{6SAT}{\delta}\right)(1+\ln T)\end{aligned}$$

In summary, (23) could be written as (25), as shown at the bottom of the next page. By Prop. 3, (25) could be further simplified as the conclusion in (17).

□

B. UPPER REGRET BOUND OF VB-UCRL

Theorem 2: After any T time-steps, the regret of VB-UCRL with restarting in Algorithm 3 is bounded by

$$\begin{aligned}&163\max(r_{\max},Dr_{\max})(V_r^T+V_p^T)^{1/3}T^{2/3}\\ &\cdot\sqrt{\ln\left(\frac{2T^3}{\delta}\right)\sum_{s,a}\Gamma(s,a)}\\ &+72\max(r_{\max},Dr_{\max})S^2A\ln\left(\frac{2T^3}{\delta}\right)\ln(2T^3)\end{aligned}\tag{26}$$

Proof: Inspired by the proof of Theorem 2 of [25], we write $V_r^{(i)}$ and $V_p^{(i)}$ for the variation of rewards and

transition probabilities in Phase i and abbreviate $V^{(i)} \stackrel{\text{def}}{=} 2V_r^{(i)} + V_p^{(i)}$, $V \stackrel{\text{def}}{=} 2V_r^T + V_p^T$ and $\theta_i \stackrel{\text{def}}{=} \lceil \frac{i^2}{V^2} \rceil$.

If the number of phases up to T is N. We have

$$\sum_{i=1}^{N-1} \lceil \frac{i^2}{V^2} \rceil < T \le \sum_{i=1}^{N} \lceil \frac{i^2}{V^2} \rceil \tag{27}$$

Recalling that $\sum_{i=1}^{N} i^2 = \frac{1}{6}N(N+1)(2N+1) > \frac{1}{3}N^3$, we have

$$T > \sum_{i=1}^{N-1} \lceil \frac{i^2}{V^2} \rceil > \sum_{i=1}^{N-1} \frac{i^2}{V^2} > \frac{(N-1)^3}{3V^2} \tag{28}$$

In other words, $N < 1 + \sqrt[3]{3V^2T}$.

Denoting τ_i as the initial time-step of phase i and s_{τ_i} as the state visited by the optimal T-time-step policy at time-step τ_i, we can decompose the regret as

$$\begin{aligned}\Delta(\text{VB-UCRL}, T) &= v_T^*(s_1) - \sum_{t=1}^{T} r_t(s_t, a_t) \\ &= \sum_{i=1}^{N} \left(\mathbb{E}\left[v_{\theta_i}^*(s_{\tau_i})\right] - \sum_{t=\tau_i}^{\tau_i - 1} r_t(s_t, a_t) \right)\end{aligned} \tag{29}$$

By Theorem 1 and a union bound over all possible values for state s_{τ_i}, the i-th summand ($i = 1, \cdots, N$) in (17) with probability $1 - \frac{\delta}{2(\tau^i)^2}$ is bounded by

$$\begin{aligned}&\max(r_{\max}, Dr_{\max}) \Bigg(43 \sqrt{\ln\left(\frac{2T^3}{\delta}\right) \sum_{s,a} \Gamma(s,a)} \cdot \sqrt{\theta_i} \\ &\quad + 72S^2A \ln\left(\frac{2T^3}{\delta}\right) \ln(2T^3) \Bigg) + Dr_{\max} V^{(i)} \theta_i\end{aligned}$$

If $\sqrt[3]{3V^2T} < 1$, we have $3V^2T < 1$ and hence $3V^2T^2 < T$ and $VT < \sqrt{3}VT < \sqrt{T}$. Furthermore, in this case $N = 1$ with $\theta_1 = T$ and $V^{(1)} = V$, so the regret bound is obtained as (30), as shown at the bottom of the next page, which is upper bounded by the claimed regret bound.

On the other hand, if $\sqrt[3]{3V^2T} \ge 1$, then $N < 2\sqrt[3]{3V^2T}$ and summing over all N phases yields from (27) that with a probability $\sum_i \frac{\delta}{2(\tau^{(i)})^2} < \sum_t \frac{\delta}{2t^2} < \delta$, the regret is bounded by

$$\begin{aligned}&\max(r_{\max}, Dr_{\max}) \Bigg(43 \sqrt{\ln\left(\frac{2T^3}{\delta}\right) \sum_{s,a} \Gamma(s,a)} \cdot \sum_{i=1}^{N} \sqrt{\theta_i} \\ &\quad + 72S^2A \ln\left(\frac{2T^3}{\delta}\right) \ln(2T^3) \Bigg) + Dr_{\max} \sum_{i=1}^{N} V^{(i)} \left(\frac{i^2}{V^2} + 1\right)\end{aligned}$$

Noting that using Jensen's inequality $\sum_{i=1}^{N} \sqrt{\theta_i} \le \sqrt{NT} \le 1.7 \cdot V^{1/3}T^{2/3}$, $\sum_{i=1}^{N} V^{(i)}(\frac{i^2}{V^2} + 1) \le \sum_{i=1}^{N} V^{(i)}(\frac{N^2}{V^2} + 1) \le \frac{N^2}{V} + V < 8.33\, V^{1/3}T^{2/3} + V$, and $V \le 2(V_r^T + V_p^T)$, we have the bound. □

$$\begin{aligned}&\Delta(\text{VB-UCRL}, T) \\ &\le 2r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} + \sum_{k=1}^{k_T} \frac{3\epsilon_k}{2} \sum_s \nu_k(s) + Dr_{\max} \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{p,k}^{sa} + Dr_{\max} T V_p^T + 6Dr_{\max}\sqrt{T \ln\left(\frac{6T}{\delta}\right)} \\ &\quad + k_T Dr_{\max} + 4\, r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} + 2\sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa} + 2TV_r^T \\ &\le 6r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} + 6Dr_{\max}\sqrt{T \ln\left(\frac{6T}{\delta}\right)} + Dr_{\max} SA \log_2\left(\frac{T}{SA}\right) + Dr_{\max} T V_p^T + 2TV_r^T \\ &\quad + \underbrace{r_{\max} \sum_{k=1}^{k_T} \frac{3}{2t_k} \sum_s \nu_k(s)}_{\psi_1} + \underbrace{2\sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{r,k}^{sa}}_{\psi_2} + \underbrace{Dr_{\max} \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a) \beta_{p,k}^{sa}}_{\psi_3}\end{aligned} \tag{23}$$

$$\begin{aligned}&\Delta(\text{VB-UCRL}, T) \\ &\le 6r_{\max}\sqrt{T \ln\left(\frac{4T}{\delta}\right)} + 6Dr_{\max}\sqrt{T \ln\left(\frac{6T}{\delta}\right)} + Dr_{\max} SA \log_2\left(\frac{8T}{SA}\right) + Dr_{\max} T V_p^T + 2TV_r^T \\ &\quad + 3r_{\max} SA\left(1 + \ln T\right) + 12\, r_{\max}\sqrt{SAT \ln\left(\frac{6SAT}{\delta}\right)} + 24\, r_{\max} SA \ln\left(\frac{6SAT}{\delta}\right)(1 + \ln T) \\ &\quad + 6Dr_{\max}\sqrt{\ln\left(\frac{6SAT}{\delta}\right)}\sqrt{\left(\sum_{s,a} \Gamma(s,a)\right) T} + 12Dr_{\max} S^2A \ln\left(\frac{6SAT}{\delta}\right)(1 + \ln T)\end{aligned} \tag{25}$$

C. DISCUSSIONS

1) REGRET BOUND ANALYSIS

Taking account of $\sum_{s,a}\Gamma(s,a) \leq \Gamma SA$, the regret bound of VB-UCRL could be $\tilde{O}\left(Dr_{\max}(V_r^T + V_p^T)^{1/3}T^{2/3}\sqrt{\Gamma SA}\right)$ as in Theorem 2 based on knowing the variations of each episode. Meanwhile, [25] and [27] give the closest regret bound of RL in MDP with both endogenous and exogenous uncertainty. In particular, [25] shows that if we ignore logarithmic terms (i.e., regarding the logarithmic terms as a constant), up to a multiplicative numerical constant, the regret bound of variation-aware UCRL in [25] is bounded by $Dr_{\max}(V_r^T + V_p^T)^{1/3}T^{2/3}S\sqrt{A}$ on the same basis as ours. Meanwhile, [27] shows that a regret bound of $\tilde{O}\left(Dr_{\max}(V_r^T + V_p^T)^{1/4}S^{2/3}A^{1/2}T^{3/4}\right)$ if we know the total variation, slightly relaxing the assumption of [25]. Since by definition $\Gamma \leq S$, the regret bound of VB-UCRL is no greater than that in [25]. However, as Γ is usually equal to $\mathcal{O}(1)$ and significantly smaller than S, our bound is superior than [25] and [27]. In particular, it can save at most $\sqrt{S}$ than [25] and $S^{\frac{1}{6}}T^{\frac{1}{12}}$ than [27], respectively.

On the other hand, [28] also unveils a $\tilde{O}\left((V_r^T + V_p^T)^{1/3}T^{2/3}\right)$ bound by a closed-box approach under conditions that either the diameter D or the total variation is known. We believe the Bernstein inequality-based performance improvement is also applicable there.

2) COMPUTATION COMPLEXITY ANALYSIS

Here, we briefly analyze the computational complexity for the three proposed algorithms.

- **Algorithm 1 (Extended Value Iteration):** This algorithm iteratively updates the value function for all state-action pairs until convergence to an ϵ-optimal policy. Each iteration has a cost of $\mathcal{O}(SA)$, and the number of iterations required to reach an ϵ-accurate solution is $\mathcal{O}(\frac{r_{\max}}{\epsilon})$. Thus, the overall computational complexity is $\mathcal{O}\left(\frac{r_{\max}SA}{\epsilon}\right)$.
- **Algorithm 2 (VB-UCRL without Variation-Aware Restarts):** In each episode, the algorithm (i) constructs confidence intervals for the rewards and transitions, and (ii) solves an optimistic MDP via EVI. Constructing the confidence intervals across all state-action pairs requires $\mathcal{O}(SA)$ operations, while solving the optimistic MDP costs $\mathcal{O}(\frac{r_{\max}SA}{\epsilon})$. Based on the doubling trick, the total number of episodes up to time T is $\mathcal{O}(SA\log T)$. Hence, the overall complexity is $\mathcal{O}\left(SA\log T\cdot\left(SA+\frac{r_{\max}SA}{\epsilon}\right)\right) = \mathcal{O}\left(\frac{r_{\max}S^2A^2\log T}{\epsilon}\right)$.
- **Algorithm 3 (VB-UCRL with Restarts):** This algorithm repeatedly invokes VB-UCRL (Algorithm 2) over a sequence of phases, each with a distinct confidence level and length determined by the known variation parameters V_r^T and V_p^T. Let θ_i denote the duration of the i-th phase. Within each phase, VB-UCRL is executed for θ_i steps with accuracy $\epsilon_i = \frac{r_{\max}}{\theta_i}$. The computational cost of VB-UCRL in each phase consists of computing confidence intervals and solving an optimistic MDP via EVI. The complexity per phase is therefore given by: $\mathcal{O}\left(SA\log\theta_i\cdot\left(SA+\frac{r_{\max}SA}{\epsilon_i}\right)\right) = \mathcal{O}\left(S^2A^2\log\theta_i(1+\theta_i)\right)$. Summing over all phases yields the total complexity: $\sum_{i=1}\mathcal{O}\left(S^2A^2\log\theta_i(1+\theta_i)\right) \leq \mathcal{O}\left(S^2A^2\log T\cdot\sum_{i=1}(1+\theta_i)\right) = \mathcal{O}\left(S^2A^2T\log T\right)$, where we have used the fact that the total time steps $\sum_i\theta_i \leq T$ and $\log\theta_i \leq \log T$ for all i. Hence, the overall computational complexity of VB-UCRL with restarts is: $\mathcal{O}(S^2A^2T\log T)$.

V. NUMERICAL EXPERIMENTS

A. EXPERIMENTAL SETTINGS

In this section, we evaluate the performance of VB-UCRL and provide the comparison with several classical methods, including Q-learning with ϵ-greedy ($\epsilon = 0.1$), UCRL2 [13], and variation-aware UCRL [25]. Meanwhile, to stand consistent with [24], we intentionally construct an initial status of the underlying non-stationary MDP with S states and A actions. In particular, the rewards and state transition probabilities satisfy the following rules: (1) The initial reward r_1 for each state-action pair is randomly generated from an independent and identically distributed uniform distribution bounded in $[0, 1]$ (i.e., Unif$[0, 1]$). Meanwhile, the initial state transition probabilities p_1 for all state-action pairs form a sparse diagonal matrix, since the number of reachable states $\Gamma \leq S$. In other words, regardless of the current state, following any action, the environment can only be transferred to a limited set of states. In addition, we adopt the cumulative reward as the primary performance metric in our experiments. Specifically, the cumulative reward is defined as the total sum of rewards collected over all time steps across all episodes during training or evaluation, that is,

$$R_{\text{cum}} = \sum_{t=1}^{T} r_t.$$

$$\begin{aligned}&\max(r_{\max}, Dr_{\max})\left(43\sqrt{\ln\left(\frac{2T^3}{\delta}\right)\sum_{s,a}\Gamma(s,a)}\cdot\sqrt{T} + 72S^2A\ln\left(\frac{2T^3}{\delta}\right)\ln(2T^3)\right) + Dr_{\max}VT\\ &\leq 43\max(r_{\max}, Dr_{\max})\sqrt{\ln\left(\frac{2T^3}{\delta}\right)\sum_{s,a}\Gamma(s,a)}\cdot\sqrt{T} + Dr_{\max}\sqrt{T} + 72\max(r_{\max}, Dr_{\max})S^2A\ln\left(\frac{2T^3}{\delta}\right)\ln(2T^3) \qquad (30)\end{aligned}$$

This metric quantifies the agent's overall performance by measuring the total reward accumulated throughout its entire interaction with the environment. (2) We introduce the implementation methods of a drifting environment (i.e., variational rewards and transition probabilities), to simulate the MDP with endogenous uncertainty. Specifically, the reward r_{t+1} at time-step $t+1$ equals the reward r_t at time-step t plus a random drifting term ξ_r, that is, $r_{t+1} = r_t + \xi_r$, with ξ_r denoting a random value sampled from Unif$[-0.02, 0.02]$ in default. Similarly, $p_{t+1} = p_t + \xi_p$, with $|\xi_p| \leq 0.02$.

Afterwards, the agent randomly selects a state as the starting state s_1 in such a non-stationary MDP. At each time-step, the agent can select an action and receive a reward, while the environment is correspondingly transferred to the next state following the transition probability matrix. We believe such an MDP could manifest the challenging exploration problem and effectively demonstrate the performance of different reinforcement learning algorithms.

B. NUMERICAL RESULTS

Firstly, we evaluate the average cumulative rewards over 5 independent runs under a non-stationary environment with state space size $S = 10$, maximal reachable states $\Gamma = 3$, and action space size $A = 3$. The results are presented in Fig. 1, where the X-axis represents the number of time steps, and the Y-axis denotes the cumulative rewards averaged over the 5 runs. The legend identifies the different algorithms under comparison, including our proposed VB-UCRL and several representative baselines. As observed from Fig. 1, VB-UCRL consistently achieves higher cumulative rewards than the baselines throughout the learning process. This indicates its superior adaptability to non-stationary environments, and is aligned with the tighter upper regret bound established in our theoretical analysis.

To further demonstrate the scalability of VB-UCRL, we conduct comparative experiments in an environment with a larger state and action space ($S = 20, A = 10$), as shown in Fig. 2, where we perform five independent runs and evaluate the cumulative rewards over 100,000 time steps. The results show that ε-greedy Q-learning and the non-restarting UCRL2 exhibit noticeable performance degradation in the more complex setting, indicating limited scalability. In contrast, restart-based methods maintain stable growth, with the Bernstein-based VB-UCRL achieving the highest cumulative rewards, demonstrating superior robustness and adaptability.

Furthermore, Fig. 3 compares the cumulative rewards obtained after 100, 000 time steps under different values of Γ (i.e., the number of reachable states), with $S = 10$ and $A = 3$. In this figure, the X-axis represents different algorithms, including VB-UCRL, variation-aware UCRL, UCRL2, and ϵ-greedy. For each algorithm, two bars are shown, corresponding to different values of Γ ($\Gamma = 3$ and $\Gamma = 10$), as indicated in the legend. The Y-axis shows the cumulative rewards. It can be observed that VB-UCRL achieves the highest cumulative reward under both settings, and the performance gain is more significant when $\Gamma = 3$, i.e., when the number of reachable states is much smaller than the total number of states S. This observation supports our intuition that in structured environments with limited state transitions, the variation-aware exploration strategy employed by VB-UCRL can better capture the underlying dynamics and thus achieve superior performance.

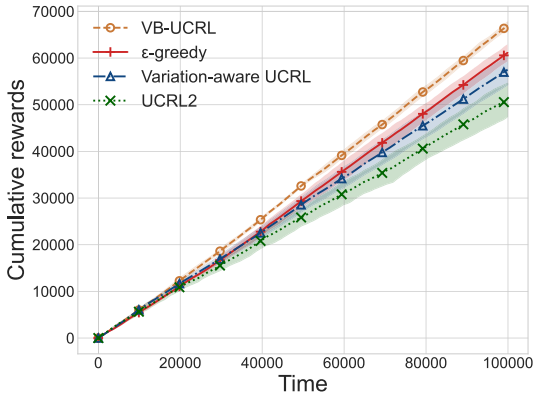


FIGURE 1. Comparison of cumulative rewards for $S = 10$ and $A = 3$.

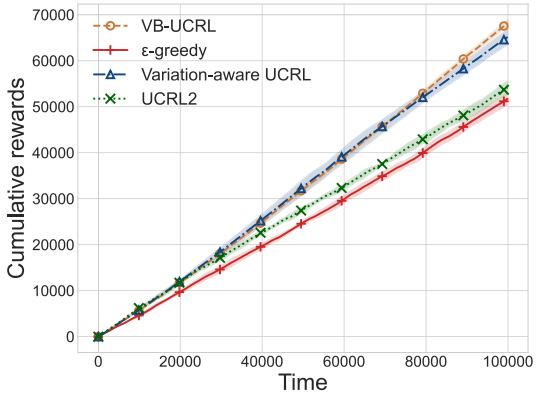


FIGURE 2. Comparison of cumulative rewards for $S = 20$ and $A = 10$.

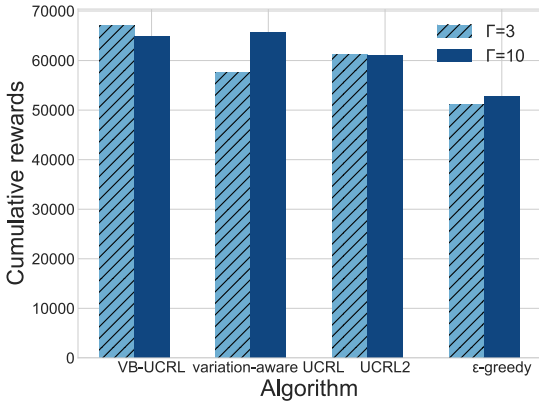


FIGURE 3. Performance comparison under different numbers Γ of reachable states for $S = 10$ and $A = 3$.

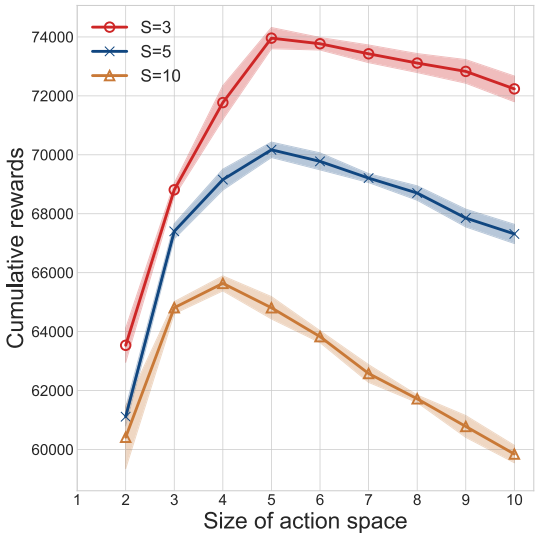


FIGURE 4. Cumulative rewards of VB-UCRL with respect to different sizes of action space A.

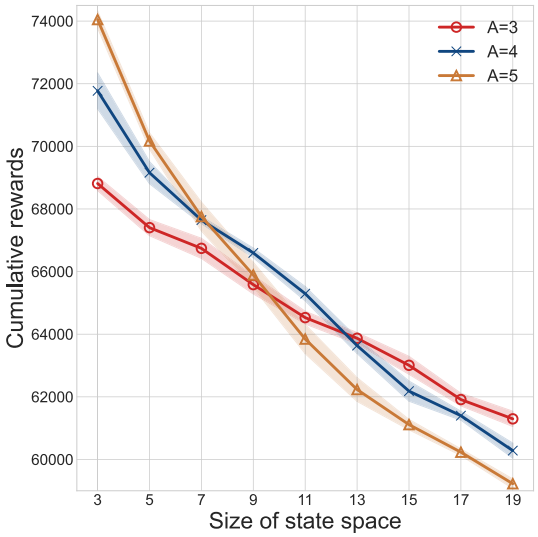


FIGURE 5. Cumulative rewards of VB-UCRL with respect to different sizes of state space S.

In Fig. 4, we present the cumulative rewards of VB-UCRL under different action space sizes A, with the X-axis denoting A and the Y-axis showing the cumulative rewards. The three curves correspond to different state space sizes $S = 3$, 5, and 10. Notably, the results exhibit a first-increase-then-decrease trend with respect to the size of the action space A, with the turning point occurring around $A = 5$. This is because, when A increases from 2 to 5, the difference in cumulative rewards between the optimal and sub-optimal actions gradually narrows. For instance, when $S = 3$, the differences for $A = 2$ to 5 are 33, 197, 24, 810, 20, 931, and 17, 096, respectively. Similar patterns are observed for $S = 5$ and $S = 10$. This narrowing gap offsets the increased difficulty in action selection, allowing performance to improve. However, as A continues to grow beyond 5, the difficulty of identifying optimal actions becomes dominant, leading to performance degradation. It is also worth noting that larger action spaces exacerbate this effect, since the number of state-action pairs grows more rapidly with increasing A in our case.

In Fig. 5, we evaluate the performance of VB-UCRL under varying sizes of the state space S, with different action space sizes $A = 3, 4, 5$. In this figure, the X-axis represents the size of the state space, and the Y-axis shows the cumulative rewards. Each curve corresponds to a different value of A, as indicated in the legend. We observe that the cumulative reward consistently decreases as S increases. This performance degradation is attributed to the growing number of state-action pairs, which makes efficient exploration more challenging. Moreover, larger action spaces (e.g., $A = 5$) exhibit a steeper decline, highlighting the compounded difficulty introduced by both large state and action spaces.

Lastly, we analyze the performance of VB-UCRL under different drifting terms ξ_r and ξ_p with $S = 10$, $\Gamma = 3$ and $A = 3$ in Fig. 6. In this 3D surface plot, the X-axis and Y-axis correspond to the values of ξ_p and ξ_r, respectively, and the Z-axis represents the cumulative rewards. Each point on the surface illustrates the performance of VB-UCRL under a specific pair of drifting terms. It can be observed that as either ξ_r or ξ_p increases, the cumulative reward decreases.This is

TABLE 1. The numerical relation between drifting terms (i.e., ξ_p and ξ_r) and the variation budgets (i.e., V_p^T and V_r^T).

Name	Value				
ξ_p	0.02	0.04	0.06	0.08	0.10
V_p^T	1,791.69	3,612.41	5,246.68	7,202.54	8,621.67
ξ_r	0.02	0.04	0.06	0.08	0.1
V_r^T	1,874.20	3,743.95	5,613.51	7,486.67	9,353.97

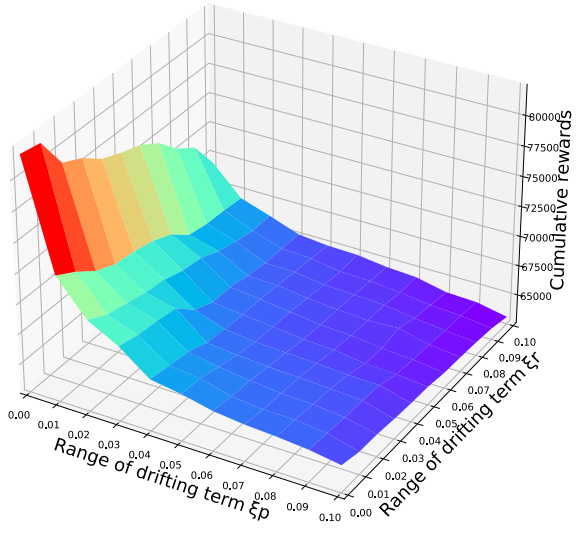


FIGURE 6. Cumulative rewards of VB-UCRL with different drifting terms.

because, as indicated in Table1, larger drifting terms result in greater variation budgets (V_r^T and V_p^T), which in turn lead to reduced learning accuracy. These empirical results fully support our theoretical analysis of VB-UCRL's sensitivity to environment dynamics.

VI. CONCLUSION

In this paper, we studied the problem of online RL for MDP with both endogenous and exogenous uncertainty, where the unknown reward and state transition distributions vary within some variation budgets. We first proposed a variation-aware Bernstein-based upper confidence reinforcement learning algorithm. In particular, we allowed UCRL to restart according to a schedule based on the variations and replaced the commonly used Hoeffding inequality with the Bernstein inequality. Our approach achieved tighter regret bounds than some of the latest works in the literature. Given the wide application of RL, our approach could contribute to the understanding of RL-based optimization performance. In the simulation, our algorithm VB-UCRL outperforms the existing algorithms in the literature.

There are many interesting future directions, e.g., how to change the method of estimating $\hat{V}_p, \hat{V}_r$ to obtain more accurate confidence intervals to improve performance. In addition to that, we will try to extend the Bernstein inequality-based performance improvement to the block-box approach in [28].

APPENDIX

Lemma 1 (Lemma 3.1 of [9]): With a probability at least $1-\frac{\delta}{6}$, $\forall T \geq 1$,

$$-\sum_{t=1}^{T} r_t \leq -\sum_{t=1}^{T}\sum_{a\in A_{s_t}} \pi_{k_t}(s_t,a) r(s_t,a) + 2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)} \tag{31}$$

Proof: The proof is a direct application of Azuma's inequality [9]. We leave it here for easier proof of the following lemmas.

Define $X_t \overset{\text{def}}{=} r_t(s_t,a_t) - \sum_{a\in A_{s_t}} \pi_{k_t}(s_t,a) r(s_t,a)\ \forall t \geq 1$. It can be observed that X_t is bounded (i.e., $|X_t| \leq r_{\max}$) and $(X_t, \mathcal{F}_t)_{t\geq 1}$ is an MDS. Thus, by applying Azuma's inequality, we have

$$\mathbb{P}\left(\sum_{t=1}^{T}\left(r_t(s_t,a_t) - \sum_{a\in A_{s_t}} \pi_{k_t}(s_t,a) r(s_t,a)\right) \leq -2r_{\max}\sqrt{T\ln\left(\frac{4T}{\delta}\right)}\right) \leq \left(\frac{\delta}{4T}\right)^2 \leq \frac{\delta}{16T^2} \tag{32}$$

Recalling that $\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}$ and taking a union bound for all $T \geq 1$, we have the probability at least $1 - \sum_{T=1}^{\infty}\frac{\delta}{16T^2} = 1 - \frac{\pi^2\delta}{96} \geq 1 - \frac{\delta}{6}$ and conclude the proof. □

Lemma 2: Under the event that $M \in \mathcal{M}_k, \forall k$, Δ_k *could be upper bounded by*

$$\Delta_k \leq \Delta_k^p + \Delta_k^r + \frac{3\epsilon_k}{2}\sum_{s\mathcal{S}} \nu_k(s) \tag{33}$$

where $\Delta_k^p \overset{def}{=} \alpha\sum_{s\in\mathcal{S}}\nu_k(s)\Big(\sum_{\substack{a\in\mathcal{A}_s\\ s'\in\mathcal{S}}}\pi_k(a|s)p_k(s'|s,a)h_k(s') - h_k(s)\Big)$ *and* $\Delta_k^r \overset{def}{=} \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}_s}\nu_k(s)\pi_k(a|s)\big(r_k(s,a)-r(s,a)\big)$.

Proof: The proof is analogous to that in Section 3.5.2 of [9].

The event that $M \in \mathcal{M}_k$ implies $g^* \leq g_k^*$. By Prop. 2.7 of [9], for an EVI produced by an optimal Bellman operator with aperiodic transformation, $|g_k - g_k^*| \leq \frac{\epsilon_k}{2}$. Therefore, $g_k \geq g^* - \frac{\epsilon_k}{2}$. Hence, we have

$$\Delta_k \leq \sum_{s\in\mathcal{S}}\nu_k(s)\left(g_k(s) - \sum_{a\in A_s}\pi_k(a|s)r(s,a) + \frac{\epsilon_k}{2}\right) \tag{34}$$

Given the extended optimal Bellman operator defined in (13), we have $\forall s \in \mathcal{S}$

$$L_k^\alpha h_k(s) = \sum_{a\in\mathcal{A}_s}\pi_k(a|s)\left\{r_k(s,a) + \alpha\sum_{s'}p_k(s'|s,a)h_k(s')\right\} + (1-\alpha)\cdot h_k(s) \tag{35}$$

By Prop. 2.7 of [9], $\|L_k^\alpha h_k - h_k - g_k\|_\infty \leq \epsilon_k$. Therefore,

$$\sum_{a\in\mathcal{A}_s}\pi_k(a|s)\left\{r_k(s,a) + \alpha\sum_{s'}p_k(s'|s,a)h_k(s')\right\} - \alpha\cdot h_k(s) - g_k(s) \geq -\epsilon_k \tag{36}$$

It can be rewritten as

$$\left(g_k(s) - \sum_{a\in\mathcal{A}_s}\pi_k(a|s)r_k(s,a)\right) \leq \alpha\left(\sum_{a\in\mathcal{A}_s}\pi_k(a|s)\sum_{s'}p_k(s'|s,a)h_k(s') - h_k(s)\right) + \epsilon_k \tag{37}$$

So,

$$\begin{aligned}\Delta_k &\leq \sum_{s\in\mathcal{S}}\nu_k(s)\left(g_k(s) - \sum_{a\in\mathcal{A}_s}\pi_k(a|s)r_k(s,a) + \frac{\epsilon_k}{2}\right) + \sum_{s\in\mathcal{S}}\nu_k(s)\sum_{a\in\mathcal{A}_s}\pi_k(a|s)\left(r_k(s,a) - \sum_{a\in A_s}r(s,a)\right)\\ &\leq \alpha\sum_{s\in\mathcal{S}}\nu_k(s)\left(\sum_{a\in\mathcal{A}_s}\pi_k(a|s)\sum_{s'}p_k(s'|s,a)h_k(s') - h_k(s)\right) + \sum_{s\in\mathcal{S}}\nu_k(s)\sum_{a\in\mathcal{A}_s}\pi_k(a|s)\left(r_k(s,a) - \sum_{a\in A_s}r(s,a)\right)\\ &\quad + \frac{3\epsilon_k}{2}\sum_{s\in\mathcal{S}}\nu_k(s) = \Delta_k^p + \Delta_k^r + \frac{3\epsilon_k}{2}\sum_{s\mathcal{S}}\nu_k(s)\end{aligned} \tag{38}$$

which concludes the proof. □

Recalling that $\Delta_k^{p1} \stackrel{\text{def}}{=} \alpha \sum_{s,a,s'} v_k(s)\pi_k(a|s)\big(p_k(s'|s,a) - p(s'|s,a)\big)h_k(s')$. If we define $\Delta_k^{p3} \stackrel{\text{def}}{=} \alpha \sum_{s,a,s'} v_k(s,a)\Big(p_k\big(s'|s,a\big) - p\big(s'|s,a\big)\Big)h_k\big(s'\big), p_k(s'|s) \stackrel{\text{def}}{=} \sum_a \pi_k(a|s)p_k(s'|s,a), \bar{p}_k(s'|s) \stackrel{\text{def}}{=} \sum_a \pi_k(a|s)p(s'|s,a)$, and $\bar{p}_k(s'|s) \stackrel{\text{def}}{=} \sum_a \pi_k(a|s)p(s'|s,a)$, we can have the following lemma.

Lemma 3: Under the case that $M \in \mathcal{M}_k, \forall k$, with probability at least $1 - \frac{\delta}{6}$, we have $\sum_{k=1}^{k_T} \Delta_k^{p1} \le \sum_{k=1}^{k_T} \Delta_k^{p3} + 4Dr_{\max}\sqrt{T \ln\big(\frac{6T}{\delta}\big)}$.

Proof: The proof is similar to that of Lemma 1. Thus, only a proof sketch is given here.

For a defined stochastic process $X_t \stackrel{\text{def}}{=} \alpha \sum_{a,s'} \pi_{k_t}(a|s_t) \cdot p_{k_t}(s'|s_t,a)h_{k_t}(s') - \alpha \sum_{s'} p_{k_t}(s'|s_t,a_t)h_{k_t}(s')$, since a given s_t, $\sum_{a,s'} \pi_{k_t}(a|s_t)p_{k_t}(s'|s_t,a) = 1$ and $\sum_{s'} p_{k_t}(s'|s_t,a_t) = 1$, we have

$$X_t = \alpha \sum_{a,s'} \pi_{k_t}(a|s_t)p_{k_t}(s'|s_t,a)w_t(s') - \alpha \sum_{s'} p_{k_t}(s'|s_t,a_t)w_t(s'),$$

where $w_t \stackrel{\text{def}}{=} h_{k_t} + \lambda_t e$ with λ_t being any constant and e being an all-one vector.

Next, we show that X_t is bounded. Theorem 2.1 of [9] shows that the aperiodic transformation could update the diameter of a communicating MDP as $D_k^\alpha = \frac{D}{\alpha} \le D$. Furthermore, Theorem 4 of [14] proves that for a communicating MDP with non-negative rewards, under a stationary policy π, the solution (g^*, h^*) of the Bellman optimality equation (i.e., $Lh^* = h^* + g^*$) satisfies $h^*(s') - h^*(s) \le \max_s (g^*(s))\, \mathbb{E}_\pi[\tau(s \to s')|s]$. Combining these observations, together with $g_k^{\alpha *} = g^* \le r_{\max}$, $\text{sp}(h_k^\alpha) \le \frac{Dr_{\max}}{\alpha}$. Therefore, if we take $\lambda_t \stackrel{\text{def}}{=} -\frac{1}{2}(\min h_{k_t} + \max h_{k_t})$, $\|w_t(s')\|_\infty \le \frac{\text{sp}(h_{k_t}^\alpha)}{2} = \frac{Dr_{\max}}{2\alpha}$, we have $|X_t| \le 2\alpha\|w_t(s')\|_\infty \le Dr_{\max}$ for all t.

On the other hand

$$\sum_{t=1}^{T} X_t = \alpha \sum_{k=1}^{k_T} \sum_{s,a,s'} (v_k(s)\pi_k(a|s) - v_k(s,a))\, p_k(s'|s,a)h_k(s') \tag{39}$$

Therefore, similar to the proof of Lemma 1, by applying Azuma's inequality and taking a union bound, we have the lemma. □

Lemma 4: Under the case $M \in \mathcal{M}_k, \forall k$, if $\hat{V}_p$ is set as the true value V_p^T, $\Delta_k^{p3} \le Dr_{\max} \sum_{s,a} v_k(s,a)(\beta_{p,k}^{sa} + V_p^T)$, where $\beta_{p,k}^{sa} \stackrel{def}{=} \sum_{s'} \beta_{p,k}^{sas'}$.

Proof: We bound Δ_k^{p3} as

$$\begin{aligned}
\Delta_k^{p3} &= \alpha \sum_{s,a,s'} v_k(s,a)\Big(p_k\big(s'|s,a\big) - p\big(s'|s,a\big)\Big)h_k\big(s'\big) \\
&\stackrel{(a)}{=} \alpha \sum_{s,a,s'} v_k(s,a)\big(p_k(s'|s,a) - p(s'|s,a)\big)w_k^s\big(s'\big) \\
&\stackrel{(b)}{\le} \alpha \sum_{s,a} v_k(s,a)\|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \|w_k^s(\cdot)\|_\infty \\
&\stackrel{(c)}{\le} \frac{Dr_{\max}}{2} \sum_{s,a} v_k(s,a)\|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \\
&\stackrel{(d)}{\le} Dr_{\max} \sum_{s,a} v_k(s,a)(\beta_{p,k}^{sa} + V_p^T)
\end{aligned}$$

where the equation (a) comes by applying a constant shift same as in Proof of Lemma 3 with $\lambda_k \stackrel{\text{def}}{=} -\frac{1}{2}(\min h_{k_t} + \max h_{k_t})$ and $w_k \stackrel{\text{def}}{=} h_k + \lambda_k e$. The inequality (b) comes from the Hölder inequality and the inequality (c) is due to $\|w_k\|_\infty \le \frac{Dr_{\max}}{2}$. Based on the following inequality, we have the inequality (d).

$$\begin{aligned}
&\|p_k(\cdot|s,a) - p(\cdot|s,a)\|_1 \\
&\stackrel{(e)}{\le} \|p_k(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 + \|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 \\
&\stackrel{(f)}{\le} 2\beta_{p,k}^{sa} + \hat{V}_p + V_p^T = 2\beta_{p,k}^{sa} + 2V_p^T
\end{aligned} \tag{40}$$

where the inequality (e) comes from the triangle inequality. Meanwhile, by construction $p_k(\cdot|s,a) \in \mathcal{B}_{p,k}$, for any $s' \in \mathcal{S}$, $|p_k(s'|s,a) - \hat{p}_k(s'|s,a)| < \beta_{p,k}^{sas'}$. Hence, $\|p_k(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 < \beta_{p,k}^{sa} + \hat{V}_p$. On the other hand, since $M \in \mathcal{M}_k, \forall k, p(\cdot|s,a) \in \mathcal{B}_{p,k}$, we have $\|p(\cdot|s,a) - \hat{p}_k(\cdot|s,a)\|_1 < \beta_{p,k}^{sa} + V_p^T$. Therefore, if $\hat{V}_p$ is set as the true value V_p^T, we obtain the inequality (f).

The conclusion comes. □

Lemma 5: Under the case $M \in \mathcal{M}_k, \forall k$, with probability at least $1 - \frac{\delta}{6}$, we have

$$\sum_{k=1}^{k_T} \Delta_k^{p2} \le 2Dr_{\max}\sqrt{T \ln\Big(\frac{6T}{\delta}\Big)} + k_T Dr_{\max}$$

Proof:

$$\begin{aligned}
&\sum_{k=1}^{k_T} \Delta_k^{p2} \\
&= \alpha \sum_{k=1}^{k_T} \sum_s v_k(s)\Big(\sum_{a,s'} \pi_k(a|s)p\big(s'|s,a\big)h_k\big(s'\big) - h_k(s)\Big) \\
&\stackrel{(a)}{=} \alpha \sum_{k=1}^{k_T} \sum_s v_k(s)\Big(\sum_{a,s'} \pi_k(a|s)p\big(s'|s,a\big)w_k\big(s'\big) - w_k(s)\Big) \\
&= \alpha \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \Big(\sum_{a,s'} \pi_k(a|s)p\big(s'|s,a\big)w_k\big(s'\big) - w_k(s_t)\Big) \\
&= \alpha \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \Big(\sum_{a,s'} \pi_k(a|s)p\big(s'|s,a\big)w_k\big(s'\big) - w_k(s_{t+1})\Big) \\
&\quad + \alpha \sum_{k=1}^{k_T} \big(w_k(s_{t_{k+1}}) - w_k(s_t)\big)
\end{aligned} \tag{41}$$

where the equation (a) comes by applying a constant shift same as in Proof of Lemma 3 with $\lambda_k \overset{\text{def}}{=} -\frac{1}{2}(\min h_{k_t} + \max h_{k_t})$ and $w_k \overset{\text{def}}{=} h_k + \lambda_k e$.

Applying similar methodology in Lemma 1 and 3, we have $\alpha \sum_{k=1}^{k_T} \sum_{t=t_k}^{t_{k+1}-1} \left(\sum_{a,s'} \pi_k(a|s) p(s'|s,a) w_k(s') - w_k(s_{t+1}) \right) \leq 2Dr_{\max} \sqrt{T \ln\left(\frac{6T}{\delta}\right)}$ with a probability at least $1 - \frac{\delta}{6}$. On the other hand, $w_k(s_{t_{k+1}}) - w_k(s_t) \leq \frac{Dr_{\max}}{\alpha}$.

The conclusion comes. □

Lemma 6: Under the case where $M \in \mathcal{M}_k, \forall k$, if $\hat{V}_r$ is set as the true value V_r^T, with probability at least $1 - \frac{\delta}{6}$, we have $\sum_{k=1}^{k_T} \Delta_k^r \leq 4\, r_{\max} \sqrt{T \ln\left(\frac{4T}{\delta}\right)} + 2 \sum_{k=1}^{k_T} \sum_{s,a} \nu_k(s,a)(\beta_{r,k}^{sa} + V_r^T)$.

Proof: The proof is similar to that of Lemma 3.

Denote $\Delta_k^{r1} \overset{\text{def}}{=} \sum_{s \in \mathcal{S}, a \in \mathcal{A}_s} \nu_k(s,a)\big(r_k(s,a) - r(s,a)\big)$. By Azuma's inequality [9], $\sum_{k=1}^{k_T} \Delta_k^r \leq \sum_{k=1}^{k_T} \Delta_k^{r1} + 4\, r_{\max} \sqrt{T \ln\left(\frac{4T}{\delta}\right)}$. Meanwhile, analogously to Lemma 4, $\Delta_k^{r1} \leq 2 \sum_{s,a} \nu_k(s,a)(\beta_{r,k}^{sa} + V_r^T)$. □

Proposition 1: For a sequence $z_1, \cdots, z_i, \cdots$ with $0 \leq z_i \leq Z_{i-1} \overset{def}{=} \max\{1, \sum_{i=1}^{k-1} z_i\}$, we have for $n \geq 1$,

$$\sum_{i=1}^{n} \frac{z_i}{Z_i} \leq 2 + 2\ln(Z_{n+1}) \tag{42}$$

and

$$\sum_{i=1}^{n} \frac{z_i}{\sqrt{Z_i}} \leq 3\sqrt{Z_{n+1}} \tag{43}$$

Proof: The proposition can be easily proven by induction. □

Proposition 2: For all $(s,a,s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$:

$$\sum_{s' \in \mathcal{S}} \sqrt{\hat{p}(s'|s,a)(1 - \hat{p}(s'|s,a))} \leq \sqrt{\Gamma(s,a) - 1} \tag{44}$$

Proof: The proposition can be proven by applying the Cauchy-Schwarz inequality and taking the fact that $\Gamma(s,a) - 1 = \sum_{s' \in \mathcal{S}} (1 - \hat{p}(s'|s,a))$. □

Proposition 3: (25) could be further simplified as

$$\begin{aligned} &\Delta(\textit{VB-UCRL}, T) \\ &\leq \max(r_{\max}, Dr_{\max}) \Bigg(43 \sqrt{T \ln\left(\frac{T}{\delta}\right) \sum_{s,a} \Gamma(s,a)} \\ &\quad + 72 S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T) \Bigg) + Dr_{\max} T V_p^T + 2 T V_r^T \end{aligned} \tag{45}$$

Proof: For $T < 6SA$, we can directly have

$$\begin{aligned} \Delta(\text{VB-UCRL}, T) &\leq r_{\max} T = r_{\max} \sqrt{T} \sqrt{T} \\ &\leq r_{\max} \sqrt{6SAT} \leq \sqrt{6T \sum_{s,a} \Gamma(s,a)} \end{aligned} \tag{46}$$

Also, if $1 \leq T \leq 43^2 A \log\left(\frac{T}{\delta}\right)$, we have $T^2 \leq 43^2 AT \log\left(\frac{T}{\delta}\right)$ (or $T \leq \sqrt{AT \log\left(\frac{T}{\delta}\right)}$), thus $\Delta(\text{VB-UCRL}, T) \leq r_{\max} \sqrt{AT \log\left(\frac{T}{\delta}\right)}$.

For $T \geq 6SA$, we have $6SAT \leq T^2$, thus $\ln\left(\frac{6SAT}{\delta}\right) \leq \ln\left(\frac{T^2}{\delta}\right) \leq 2\ln\left(\frac{T}{\delta}\right)$. Similarly, if $T \geq 43^2 A \log\left(\frac{T}{\delta}\right)$, we have $A \leq \frac{\sqrt{AT \log\left(\frac{T}{\delta}\right)}}{43 \log\left(\frac{T}{\delta}\right)}$. Together with $\log(8T) \leq 2\log(T)$, $\log\left(\frac{8T}{SA}\right) \leq \frac{2}{43} \sqrt{AT \log\left(\frac{T}{\delta}\right)}$

Thus, we can simplify (25) as

$$\begin{aligned} &\Delta(\text{VB-UCRL}, T) \\ &\leq \max(r_{\max}, Dr_{\max}) \Bigg(\sqrt{T \ln\left(\frac{T}{\delta}\right) \sum_{s,a} \Gamma(s,a)} \\ &\quad \times (6\sqrt{2} + 6\sqrt{2} + 12\sqrt{2} + \frac{2}{43}) \\ &\quad + S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T)(24 + 48) \Bigg) + Dr_{\max} T V_p^T + 2 T V_r^T \\ &\leq \max(r_{\max}, Dr_{\max}) \Bigg(43 \sqrt{T \ln\left(\frac{T}{\delta}\right) \sum_{s,a} \Gamma(s,a)} \\ &\quad + 72 S^2 A \ln\left(\frac{T}{\delta}\right) \ln(T) \Bigg) + Dr_{\max} T V_p^T + 2 T V_r^T \end{aligned} \tag{47}$$

Replacing $\delta' = \frac{5}{6}\delta$, and taking account of $\log\left(\frac{T}{\delta'}\right) = \log\left(\frac{6T}{5\delta}\right) < 2\log\left(\frac{T}{\delta}\right)$, we have the conclusion. □

REFERENCES

[1] K. Qi, Q. Wu, P. Fan, N. Cheng, W. Chen, J. Wang, and K. B. Letaief, "Deep-reinforcement-learning-based AoI-aware resource allocation for RIS-aided IoV networks," *IEEE Trans. Veh. Technol.*, vol. 74, no. 1, pp. 1365–1378, Jan. 2025.

[2] M. M. Salah, R. S. Saad, R. M. Zaki, K. Rabie, and B. M. ElHalawany, "Multi-armed bandits for resource allocation in UAV-assisted LoRa networks," *IEEE Internet Things Mag.*, vol. 8, no. 2, pp. 40–45, Mar. 2025.

[3] A. Alhammadi, I. Shayea, A. A. El-Saleh, M. H. Azmi, Z. H. Ismail, L. Kouhalvandi, and S. A. Saad, "Artificial intelligence in 6G wireless networks: Opportunities, applications, and challenges," *Int. J. Intell. Syst.*, vol. 2024, no. 1, Mar. 2024, Art. no. 8845070.

[4] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1226–1252, 2nd Quart., 2021.

[5] M. Banafaa, I. Shayea, J. Din, M. H. Azmi, A. Alashbi, Y. I. Daradkeh, and A. Alhammadi, "6G mobile communication technology: Requirements, targets, applications, challenges, advantages, and opportunities," *Alexandria Eng. J.*, vol. 64, pp. 245–274, Feb. 2023.

[6] Z. Wu and R. Xu, "Risk-sensitive Markov decision process and learning under general utility functions," 2023, *arXiv:2311.13589*.

[7] S. Malhotra, F. Yashu, M. Saqib, D. Mehta, J. Jangid, and S. Dixit, "Deep reinforcement learning for dynamic resource allocation in wireless networks," 2025, *arXiv:2502.01129*.

[8] A. Aniket and A. Chattopadhyay, "Online reinforcement learning in periodic MDP," 2023, *arXiv:2303.09629*.

[9] R. Fruit, "Exploration-exploitation dilemma in reinforcement learning under various form of prior knowledge," Ph.D. dissertation, Dept. Sciences et Technologie, Université de Lille, Lille, France, 2019.

[10] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[11] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

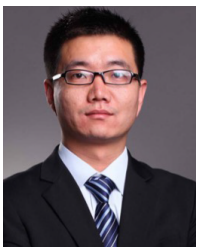
[12] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Mach. Learn.*, vol. 49, no. 2, pp. 209–232, Nov. 2002.

[13] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *J. Mach. Learn. Res.*, vol. 11, no. 51, pp. 1563–1600, Mar. 2010.

[14] P. L. Bartlett and A. Tewari, "REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs," in *Proc. UAI*, Montreal, QC, Canada, Jan. 2012, pp. 35–42.

[15] R. Fruit, M. Pirotta, and A. Lazaric, "Improved analysis of UCRL2 with empirical Bernstein inequality," 2020, *arXiv:2007.05456*.
[16] B. Hao, Y. A. Yadkori, Z. Wen, and G. Cheng, "Bootstrapping upper confidence bound," in *Proc. NIPS*, Vancouver, BC, Canada, Dec. 2019, pp. 2442–2452.
[17] J. Qian, R. Fruit, M. Pirotta, and A. Lazaric, "Exploration bonus for regret minimization in discrete and continuous average reward MDPs," in *Proc. NIPS*, vol. 32, Vancouver, BC, Canada, Sep. 2019, pp. 4890–4899.
[18] H. Bourel, O.-A. Maillard, and M. S. Talebi, "Tightening exploration in upper confidence reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jan. 2020, pp. 1093–1103.
[19] A. Ghosh, X. Zhou, and N. Shroff, "Towards achieving sub-linear regret and hard constraint violation in model-free RL," in *Proc. AISTATS*, Valencia, Spain, May 2024, pp. 1334–1356.
[20] T. Lancewicki, A. Rosenberg, and Y. Mansour, "Learning adversarial Markov decision processes with delayed feedback," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, Jun. 2022, pp. 7281–7289.
[21] Y. Abbasi-Yadkori, A. György, and N. Lazic, "A new look at dynamic regret for non-stationary stochastic bandits," *J. Mach. Learn. Res.*, vol. 24, pp. 288:1–288:37, Jan. 2022.
[22] P. Zhao, Y. Zhang, L. Zhang, and Z. Zhou, "Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization," *J. Mach. Learn. Res.*, vol. 25, no. 98, pp. 1–52, 2024.
[23] Y. Li and N. Li, "Online Markov decision processes with time-varying transition probabilities and rewards," in *Proc. ICML*, Long Beach, CA, USA, Jun. 2019, pp. 3924–3933.
[24] Y. Li and N. Li, "Online learning for Markov decision processes in nonstationary environments: A dynamic regret analysis," in *Proc. Amer. Control Conf. (ACC)*, Philadelphia, PA, USA, Jul. 2019, pp. 1232–1237.
[25] P. Gajane, R. Ortner, and P. Auer, "Variational regret bounds for reinforcement learning," in *Proc. UAI*, Tel Aviv-Yafo, Israel, Jan. 2019, pp. 509–518.
[26] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Learning to optimize under non-stationarity," in *Proc. AISTATS*, Okinawa, Japan, Jan. 2018, pp. 1074–1082.
[27] W. C. Cheung, D. Simchi-Levi, and R. Zhu, "Reinforcement learning for non-stationary Markov decision processes: The blessing of (more) optimism," in *Proc. ICML*, Vienna, Austria, Jan. 2020, pp. 1841–1850.
[28] C.-Y. Wei and H. Luo, "Non-stationary reinforcement learning without prior knowledge: An optimal black-box approach," in *Proc. COLT*, Boulder, CO, USA, Feb. 2021, pp. 1498–1531.
[29] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theor. Comput. Sci.*, vol. 410, no. 19, pp. 1876–1902, Apr. 2009.
[30] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Chichester, U.K.: Wiley, Apr. 1994.
[31] P. J. Schweitzer, "On undiscounted Markovian decision processes with compact action spaces," *RAIRO-Oper. Res.*, vol. 19, no. 1, pp. 71–86, 1985.
[32] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*. London, U.K.: Oxford Univ. Press, May 2013.

RUOQI WEN received the B.Sc. degree in mathematics and applied mathematics from Hefei University of Technology, Hefei, China, in June 2021. She is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. Her research interests include multi-agent reinforcement learning, regret bound, endogenous and exogenous uncertainty analysis, wireless networks, and autonomous vehicle control.



RONGPENG LI (Senior Member, IEEE) was a Research Engineer with the Wireless Communication Laboratory, Huawei Technologies Company Ltd., Shanghai, China, from August 2015 to September 2016. He was a Visiting Scholar with the Department of Computer Science and Technology, University of Cambridge, Cambridge, U.K., from February 2020 to August 2020. He is currently an Associate Professor with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His research interests include networked intelligence for communications evolving (NICE). He received the Wu Wenjun Artificial Intelligence Excellent Youth Award in 2021. He serves as an Editor for *China Communications*.