Deterministic Sequencing of Exploration and Exploitation for Reinforcement Learning

Piyush Gupta and Vaibhav Srivastava

***Abstract*— We propose Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithm with interleaving exploration and exploitation epochs for model-based RL problems that aim to simultaneously learn the system model, i.e., a Markov decision process (MDP), and the associated optimal policy. During exploration, DSEE explores the environment and updates the estimates for expected reward and transition probabilities. During exploitation, the latest estimates of the expected reward and transition probabilities are used to obtain a robust policy with high probability. We design the lengths of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time.**

I. Introduction

Reinforcement Learning (RL) is used in solving complex sequential decision-making tasks in uncertain environments such as motion planning for robots [1], [2], personalized web services [3], [4], and the design of decision-support systems for human-supervisory control [5], [6]. Markov decision processes (MDPs) [7] provide a natural framework for optimal decision-making under uncertainty and are used to model and solve numerous model-based RL problems. The objective of these problems is to simultaneously learn the system model and the optimal policy. While MDP formulation accounts for environment uncertainty by using stochastic models, MDP policies are known to be sensitive to errors in these stochastic models [8], [9].

In many safety-critical systems, robust MDPs [10], [11] are used to mitigate performance degradation due to uncertainty in the learned MDP. However, to reduce the system uncertainty, the agent must explore the environment and visit parts of the state space associated with high estimation uncertainty. Most often, RL algorithms use simple randomized methods to explore the environment, e.g. applying ϵ-greedy policies [12] or adding random noise to continuous actions [13]. The objective of the robust MDPs is conflicting with the exploration objective, i.e., robust policy avoids the unexplored regions of the state space to optimize the worst-case performance while the objective of the exploration is to reduce the system uncertainty by visiting unexplored regions of the state space. Therefore, to balance the trade-off between learning the MDP and designing a robust policy, we design a Deterministic Sequencing of Exploration and Exploitation (DSEE) algorithm, in which exploration and exploitation epochs of increasing lengths are interleaved.

This work has been supported in part by NSF Award IIS-1734272, CMMI-1940950, and ECCS-2024649.

Piyush Gupta (guptapi1@msu.edu) and Vaibhav Srivastava (vaibhav@egr.msu.edu) are with the Department of Electrical and Computer Engineering, Michigan State University, East Lansing, Michigan, 48824, USA.

There exist efficient algorithms for solving RL problems with provable bounds on the sample complexity [14, Definition 1]. In [14], authors analyze the Model-based Interval Estimation (MBIE) algorithm that applies confidence bounds to compute an optimistic policy and show that the algorithm is PAC-optimal [14, Definition 2]. They provide an upper bound on the algorithm's sample complexity given by $O\left(\frac{|\mathcal{S}|^2|\mathcal{A}|}{(1-\gamma)^6\epsilon^3}\log(\delta^{-1})\right)$ which is the maximum number of time steps until when the MBIE policy is not $\epsilon-$optimal with at least probability $1-\delta$, where $|\mathcal{S}|$, $|\mathcal{A}|$, are the cardinality of the state space and action space, respectively, γ is the discount factor, and $\epsilon, \delta \in (0,1)$ are pre-defined constants. A similar bound on the sample complexity is obtained for the R-max algorithm [15] which distinguishes the "known" and "unknown" states based on how often they have been visited. It explores by acting to maximize rewards under the assumption that unknown states deliver the maximum reward. UCRL2 algorithm [16] relies on optimistic bounds on the reward functions and probability density functions and enjoys near-optimal regret bounds. A review of model-based RL algorithms with provable finite time guarantees can be found in [17, Chapter 38]. A major drawback of these algorithms is that they consider optimism in the face of uncertainty and hence, are not robust to the estimation uncertainties. Furthermore, these algorithms with random exploration might lead to a bad user experience in applications in which the RL agent seeks to learn human preferences for system optimization.

To address these shortcomings, we propose a DSEE algorithm for model-based RL in which we design a deterministic sequence of exploration and exploitation epochs. The DSEE approach has been used in multi-arm bandit problems [18]–[21] and multi-robot coordination [22]. It allows for differentiation between exploration and exploitation epochs. The announced exploration may lead to a better user experience for the agents (especially for human agents) than random exploration at any time. For example, many personalized web services calibrate their recommendations intermittently by announced exploration, i.e., through surveys and user selection. Another advantage of the DSEE algorithm is that it allows for efficient exploration of the environment in multi-agent systems. Specifically, in multi-agent systems, exploration can be well-planned to cover all regions of the state-space through agent coordination which can be easily arranged due to the deterministic structure of exploration and exploitation.

We design the DSEE algorithm with alternating sequences of exploration and exploitation. In exploration epochs, the

algorithm learns the MDP, while in exploitation epochs, it uses a robust policy based on the learned MDP and the associated uncertainty. We design the lengths of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time.

The major contributions of this work are twofold: (i) we propose a DSEE algorithm for model-based RL problems and (ii) we design the lengths of the exploration and exploitation epochs such that the cumulative regret for the DSEE algorithm grows as a sub-linear function of time.

This manuscript is structured as follows: in Section II, we provide background and formulate the problem. In Section III, we provide an overview of the DSEE algorithm. In Section IV, we analyze the DSEE algorithm and design the exploration and exploitation epochs such that the cumulative regret grows sub-linearly with time. We conclude in Section V.

II. Background and Problem Formulation

We focus on the model-based RL problems which aim to simultaneously learn the system model, i.e., a Markov decision process (MDP), and the associated optimal policy. We seek to design policies that are robust to uncertainty in the learned MDP. However, learning the MDP requires visiting parts of the state space associated with high uncertainty in estimates and has exactly the opposite effect of a robust policy. Therefore, to balance the trade-off between learning the MDP and designing a robust policy, we design a DSEE algorithm, in which exploration and exploitation epochs of increasing lengths are interleaved. In exploration epochs, the algorithm learns the MDP, while in exploitation epochs, it uses a robust policy based on the learned MDP and the associated uncertainty.

Consider an MDP $(\mathcal{S}; \mathcal{A}, R, \mathbb{P}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, reward $R(s,a)$, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, is a random variable with support $[0, R_{\max}]$, $\mathbb{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{|\mathcal{S}|}$ is the transition distribution, and $\gamma \in (0,1)$ is the discount factor. Here, $\Delta_{|\mathcal{S}|}$ represents probability simplex in $\mathbb{R}^{|\mathcal{S}|}$, $|\cdot|$ represents the cardinality of a set. Let $\overline{R}(s,a)$ be the expected value of $R(s,a)$. We consider a finite MDP setting in which $|\mathcal{S}|$ and $|\mathcal{A}|$ are finite.

We assume that the rewards R and the state transition distribution $\mathbb{P}$ are unknown a priori. Hence, during exploration, we estimate $\overline{R}$ and $\mathbb{P}$ using online observations. Let (s,a) be any state-action pair where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. At any time t, let $n_t(s,a)$ be the number of times state-action pair (s,a) is observed until time t. For each (s,a), the empirical mean estimates $\hat{R}_t(s,a)$ and $\hat{\mathbb{P}}_t(s'|s,a)$, $s' \in \mathcal{S}$ are:

$$\hat{R}_t(s,a) = \frac{1}{n_t(s,a)} \sum_{i=1}^{n_t(s,a)} r_i(s,a), \text{and} \tag{1}$$

$$\hat{\mathbb{P}}_t(s'|s,a) = \frac{n_t(s,a,s')}{n_t(s,a)}, \tag{2}$$

respectively, where $r_i(s,a)$ is the immediate reward obtained in (s,a) during observation $i \in \{1, \ldots, n_t(s,a)\}$ until time t and $n_t(s,a,s')$ is the number of times the next state s' is observed from (s,a) out of $n_t(s,a)$ times.

Oftentimes, the uncertainty in probability transition matrices and mean reward function can be large, especially in the initial stages of learning due to limited observation data, which may lead to sub-optimal policies. Robust MDPs [10] mitigate the sub-optimal performance arising from this uncertainty by optimizing the worst-case performance over given uncertainty sets for reward function and probability transition matrices to obtain a robust policy. Given, at time t, uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ containing $\overline{R}$ and $\mathbb{P}$, respectively, the robust MDP solves the following robust Bellman equation:

$$V_t^R(s) = \max_{a \in \mathcal{A}} \min_{\tilde{R}_t \in \mathcal{R}_t^U,\ \tilde{\mathbb{P}}_t \in \mathcal{P}_t^U} \Big\{ \tilde{R}_t(s,a) + \gamma \sum_{s'} \tilde{\mathbb{P}}_t(s'|s,a) V_t^R(s') \Big\}, \tag{3}$$

to obtain a robust policy $\hat{\pi}_t^R = \operatorname{argmax}_{a \in \mathcal{A}} V_t^R$, which optimizes the worst-case performance through minimization with respect to the uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$, where V_t^R is the robust value function.

The choice of these uncertainty sets are critical for the performance of the robust algorithm. A poor modeling choice can increase the computational complexity and result in a highly conservative policy [9], [23]. To avoid these issues, during the exploitation epoch of the DSEE, we construct these uncertainty sets based on the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ from the previous exploration epochs and Hoeffding bounds [24] for $\hat{R}_t$ (Lemma 1) and $\hat{\mathbb{P}}_t$ (Lemma 2). Subsequently, we utilize robust MDP to learn a policy that is robust to the estimation uncertainties with high probability. The convergence of the robust MDP with uncertain transition matrices to the uncertainty-free MDP can be shown under the assumption that the uncertainty sets converge to singleton estimates almost surely [9], [25].

*Definition 1 (**Instantaneous and Cumulative Regret**):* For a discounted and ergodic RL [26], consider an algorithm $\mathbb{A}$ that, at the end of the $(t-1)$-th step, returns a policy π_t to be applied in the t-th step. For any state $s \in \mathcal{S}$, let $V^*(s)$ and $V^{\pi_t}(s)$ be the optimal value of the state and its value under the policy π_t, respectively. At any time t, the instantaneous regret $\mathfrak{R}(t)$ of the algorithm $\mathbb{A}$ is given by:

$$\mathfrak{R}(t) = \|V^*(s) - V^{\pi_t}(s)\|_\infty, \tag{4}$$

where $\|\cdot\|_\infty$ denotes the L^∞-norm of a vector, and the cumulative regret $\mathbf{R}_T$ until time horizon T is given by:

$$\mathbf{R}_T = \sum_{t=1}^{T} \mathfrak{R}(t) = \sum_{t=1}^{T} \|V^*(s) - V^{\pi_t}(s)\|_\infty. \tag{5}$$

We design the exploration and exploitation epochs of the DSEE algorithm such that its cumulative regret grows as a sub-linear function of time. In the next section, we provide an overview of the DSEE algorithm.

III. DSEE Algorithm

We design the DSEE algorithm for model-based RL under the following assumptions:

(A1) State space $\mathcal{S}$ and action space $\mathcal{A}$ are finite sets.

Algorithm 1 Deterministic Sequencing of Exploration and Exploitation (DSEE)

Input: Set of states $\mathcal{S}$, Set of actions $\mathcal{A}$, Initial State s_0;
Set: $\eta > 1$, Sequences $\{\epsilon_j\}_{j\in\mathbb{N}}$, $\{\delta_j\}_{j\in\mathbb{N}}$, $s_0^{\text{end}} = s_0$, $s = s_0$;
Set: $t = 0$, $n(s,a) = 0, n(s,a,s') = 0, \mathrm{S}(s,a) = 0, \forall s, a, s'$;

for epoch $j = 1, 2, \ldots$ **do**
  % Exploration phase:
  $\rho_j \leftarrow \frac{\epsilon_j}{4+\frac{2R_{\max}\gamma}{(1-\gamma)^2}}$;
  $\mu \leftarrow 2\left[\log(2^{|\mathcal{S}|}-2) + \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_j}\right)\right]$;
  $U_j \leftarrow \max\left\{\frac{(R_{\max})^2\log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_j})}{2\rho_j^2}, \frac{\mu}{\rho_j^2}\right\}$
  while $n(s,a) < U_j,\ \forall(s,a)$
    $t \leftarrow t+1$;
    Pick $a \sim \text{UNIF}(\mathcal{A})$ in current state s **do**
    Observe reward R and the next state s'
    if s_{j-1}^{end} has been visited in epoch j **then**
      $n(s,a) \leftarrow n(s,a)+1$;
      $n(s,a,s') \leftarrow n(s,a,s')+1$;
      $\mathrm{S}(s,a) = \mathrm{S}(s,a) + \mathrm{R}$;
    end if
    $s \leftarrow s'$;
  end while
  $s_j^{\text{end}} \leftarrow s$;
  $\hat{R}_t(s,a) = \frac{\mathrm{S}(s,a)}{n(s,a)},\ \forall(s,a)$;
  $\hat{\mathbb{P}}_t(s'|s,a) = \frac{n(s,a,s')}{n(s,a)},\ \forall(s,a)$;
  % Exploitation phase:
  Construct uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ using (11);
  Compute $V_t^R(s)$ and $\hat{\pi}_t^R$ using (3);
  Implement $\hat{\pi}_t^R$ for $\lceil\eta^j\rceil$ time steps;
  $t \leftarrow t + \lceil\eta^j\rceil$;
end for

(A2) The MDP is ergodic under the uniform policy π, i.e., under a policy π that, in every state s, randomly selects the actions from $\mathcal{A}$ with equal probability, the MDP admits a unique stationary distribution $\phi_\pi(s) : \mathcal{S} \to \Delta_{|\mathcal{S}|}$, with $\phi_\pi(s) > 0$ for all s.

Ergodic MDP [26] (assumption (A2)) is a common assumption. It ensures that the stationary distribution is independent of the initial distribution and all states are recurrent, i.e., each state s is visited infinitely often and $\phi_\pi(s) > 0$. We use this assumption to estimate the number of times each state is visited in N time steps.

Algorithm 1 shows an overview of the DSEE algorithm. In the DSEE algorithm, we design a sequence of alternating exploration and exploitation epochs. Let α_i and β_i be the lengths of the i-th exploration and exploitation epoch, respectively, where $i \in \mathbb{N}$. During an exploration epoch, we uniformly sample the action in the current state and update the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$. For a given sequence of $\{\epsilon_i\}_{i\in\mathbb{N}}$ and $\{\delta_i\}_{i\in\mathbb{N}}$ that we design in Section IV, the length of the exploration epoch α_i is determined to reduce the estimation uncertainty such that $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \le \epsilon_i) \ge 1-\delta_i$ after the epoch, where $\mathbf{P}(\cdot)$ denotes the probability measure, and $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R$. In DSEE, we choose exponentially increasing lengths of the exploitation epochs β_i. During the exploitation epoch, we utilize the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ from previous exploration epochs and construct the uncertainty sets $\mathcal{R}^U$ and $\mathcal{P}^U$ at time t. We use these uncertainty sets with a robust Bellman equation to learn a policy that is robust to the estimation uncertainties with high probability. In next section, we analyze the DSEE algorithm, and design the sequence of $\{\epsilon_i\}_{i\in\mathbb{N}}$ and $\{\delta_i\}_{i\in\mathbb{N}}$, such that the cumulative regret (5) grows as a sub-linear function of time.

IV. Analysis of DSEE algorithm

We now characterize the regret of the DSEE algorithm under the assumptions (A1-A2) and design the exploration and exploitation epochs. The optimal value $V^*(s_t)$ of the state s_t is given by:

$$V^*(s_t) = \overline{R}(s_t, \pi^*(s_t)) + \gamma\mathbb{E}[V^*(s_{t+1})|s_t, \pi^*(s_t)], \quad (6)$$

where π^* is an optimal policy that satisfies:

$$\pi^*(s_t) = \underset{a_t}{\text{argmax}}\left\{\overline{R}(s_t,a_t) + \gamma\mathbb{E}[V^*(s_{t+1})|s_t,a_t]\right\}. \quad (7)$$

We define an approximate optimal value function $\hat{V}_t$ that utilizes the estimates $\hat{R}_t$ and $\hat{\mathbb{P}}_t$ at time t. Therefore, $\hat{V}_t(s_t)$ is given by:

$$\hat{V}_t(s_t) = \hat{R}_t(s_t, \hat{\pi}_t(s_t)) + \gamma\hat{\mathbb{E}}\left[\hat{V}_t(s_{t+1})|s_t, \hat{\pi}_t(s_t)\right], \quad (8)$$

where $\hat{\mathbb{E}}\left[\hat{V}_t(s_{t+1})|s_t,\hat{\pi}_t(s_t)\right]$ is used to denote $\sum_{s_{t+1}}\hat{\mathbb{P}}_t(s_{t+1}|s_t,\hat{\pi}_t(s_t))\hat{V}_t(s_{t+1})$ and $\hat{\pi}_t$ is an optimal policy for the approximate optimal value function given by:

$$\hat{\pi}_t(s_t) = \underset{a_t}{\text{argmax}}\left\{\hat{R}_t(s_t,a_t) + \gamma\hat{\mathbb{E}}\left[\hat{V}_t(s_{t+1})|s_t,a_t\right]\right\}. \quad (9)$$

*Theorem 1 (**Concentration of robust value function**):* Let $\|\cdot\|_1$ denote the L^1-norm of a vector. For any given $\epsilon_t, \delta_t \in (0,1)$, there exists an $n \in O\left(\frac{|\mathcal{S}|}{\epsilon_t^2} + \frac{1}{\epsilon_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that if each state-action pair (s,a) is observed $n_t(s,a) \ge n$ times until time t, then for each state s, the following inequality holds:

$$\mathbf{P}\left(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \le \epsilon_t\right) \ge 1-\delta_t, \quad (10)$$

where $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R = \text{argmax}_{a\in\mathcal{A}} V_t^R$. The robust value function V_t^R is defined in (3) with $\rho_t = \frac{\epsilon_t}{2}\left(2+\frac{R_{\max}\gamma}{(1-\gamma)^2}\right)^{-1}$ and

$$\mathcal{R}_t^U = \left\{R^U(s,a) : |R^U(s,a) - \hat{R}_t(s,a)| \le \rho_t,\ \forall(s,a)\right\},$$
$$\mathcal{P}_t^U = \left\{\mathbb{P}^U(s,a) : \left\|\mathbb{P}^U(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \le \rho_t,\ \forall(s,a)\right\}. \quad (11)$$

We prove Theorem 1 using the following Lemmas 1-4.

*Lemma 1 (**Concentration of rewards**):* Suppose until time step t, the state-action pair (s,a) is observed $n_t(s,a)$ times and bounded immediate rewards $r_i(s,a)$,

$i \in \{1, \ldots, n_t(s,a)\}$, are obtained at these instances. Then the following inequality holds:

$$\mathbf{P}\left(\left|\overline{R}(s,a) - \hat{R}_t(s,a)\right| \leq \epsilon_t^R\right) \geq 1 - \delta_t^R, \qquad (12)$$

where $\hat{R}_t(s,a)$ is the empirical mean reward defined in (1) and $\epsilon_t^R = \sqrt{\frac{(R_{\max})^2 \log(2/\delta_t^R)}{2n_t(s,a)}}$.

Proof: For brevity of notation, let ϵ_R and δ_R denote ϵ_t^R and δ_t^R, respectively. For bounded random variables $r_i(s,a)$, using the Hoeffding bounds [24], we have

$$\mathbf{P}\left(\left|\overline{R}(s,a) - \hat{R}_t(s,a)\right| \leq \epsilon_R\right) \geq 1 - 2e^{-\frac{2n_t(s,a)\epsilon_R^2}{(R_{\max})^2}}. \quad (13)$$

Choosing $\delta_R = 2e^{-\frac{2n_t(s,a)\epsilon_R^2}{(R_{\max})^2}}$, we get the desired result. ■

*Lemma 2 (**Concentration of transition probabilities**):* Suppose until time step t, the state-action pair (s,a) is observed $n_t(s,a)$ times and let $\mathbb{P}(s,a) \in \Delta_{|\mathcal{S}|}$ be the true transition probability distribution for (s,a). Then for any (s,a), the following inequality holds:

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \epsilon_t^P\right) \geq 1 - \delta_t^P, \qquad (14)$$

where $\|\cdot\|_1$ is the L^1 norm of a vector and $\hat{\mathbb{P}}_t(s,a)$ is the empirical transition probability vector with components $\hat{\mathbb{P}}_t(s'|s,a)$ defined in (2), and $\epsilon_t^P = \sqrt{\frac{2[\log(2^{|\mathcal{S}|}-2) - \log(\delta_t^P)]}{n_t(s,a)}}$.

Proof: For brevity of notation, let ϵ_P and δ_P denote ϵ_t^P and δ_t^P, respectively. Using [27, Theorem 2.1], we have:

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \epsilon_P\right) \geq 1 - (2^{|S|} - 2)e^{-\frac{n_t(s,a)\epsilon_P^2}{2}}. \qquad (15)$$

Setting $\delta_P = (2^{|S|} - 2)e^{-\frac{n_t(s,a)\epsilon_P^2}{2}}$, yields the desired result. ■

Lemmas 1 and 2 provide concentration bounds on the reward and transition probability based on how often a state-action pair is visited.

*Lemma 3: **(Concentration of reward and transition probability functions)*** Let $\delta_t^R = \delta_t^P = \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$. Then, for any $\rho_t > 0$, there exists an $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that when each state-action pair (s,a) is observed $n_t(s,a) \geq n$ times, then the following statements hold for any (s,a):

(i) $\mathbf{P}\left(|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$,

(ii) $\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$.

Proof: Using Lemmas 1 and 2, we know that $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$ holds for any (s,a) with at least probability $1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$ for $\rho_t \geq \sqrt{\frac{(R_{\max})^2 \log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_t})}{2n_t(s,a)}}$ and $\rho_t \geq \sqrt{\frac{2\left[\log(2^{|\mathcal{S}|}-2) - \log\left(\frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}\right)\right]}{n_t(s,a)}}$, respectively. Hence, we have that

$$n_t(s,a) \geq \max\left\{\frac{(R_{\max})^2 \log(\frac{4|\mathcal{S}||\mathcal{A}|}{\delta_t})}{2\rho_t^2}, \frac{\mu}{\rho_t^2}\right\}, \qquad (16)$$

where $\mu = 2\left[\log(2^{|\mathcal{S}|} - 2) + \log\left(\frac{2|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right]$, is sufficient to guarantee that $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$ holds for any (s,a) with at least probability $1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$. Hence, we can choose $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ such that (16) holds, and hence, the lemma follows. ■

*Remark 1 (**Concentration inequalities**):* The concentration inequalities in Lemmas 1, 2, and 3 use a deterministic value of $n_t(s,a)$. However, these bound also apply if $n_t(s,a)$ is a realization of a random process that is independent of $\hat{R}_t(s,a)$ and $\hat{\mathbb{P}}_t(s,a)$, which would be the case in this paper.

*Remark 2 (**Uncertainty set**):* Using union bounds over all (s,a) and Lemma 3, it follows that uncertainty sets $\mathcal{R}_t^U$ and $\mathcal{P}_t^U$ stated in Theorem 1 are ρ_t-level uncertainty sets for $R(s,a)$ and $\mathbb{P}(s,a)$, respectively, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$ with at least probability $1 - \delta_t$. Thus, the policy obtained at time t in an exploitation epoch is robust to estimation uncertainties with at least probability $1 - \delta_t$.

*Lemma 4 (**Loss in robust value function**):* Suppose $|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t$ and $\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t$, for any (s,a), at time t. Then $\mathcal{L}_t(s) = V^*(s) - V^{\hat{\pi}_t^R}(s)$ satisfies:

$$\mathcal{L}_t(s) \leq 2\rho_t\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right), \qquad (17)$$

where $V^{\hat{\pi}_t^R}(s)$ is the value of state s under the robust policy $\hat{\pi}_t^R$ at time t.

Proof [Sketch]: The upper bound on the loss is obtained by following a similar analysis as in [28] that provides an upper bound on the loss by considering the uncertainty in $\mathbb{P}$ only. The analysis in [28] can be extended by considering the uncertainty in $\overline{R}$ as well. The details of the proof can be found in [29].

Lemma 4 provides the bounds on the loss in robust value function w.r.t. the optimal value function using the concentration bounds on the rewards and transition probabilities.

Proof of Theorem 1: Using Lemma 3, we know that when each state-action pair (s,a) is sampled $n \in O\left(\frac{|\mathcal{S}|}{\rho_t^2} + \frac{1}{\rho_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$ times, then the following inequalities holds for any (s,a):

$$\mathbf{P}\left(|\overline{R}(s,a) - \hat{R}_t(s,a)| \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}, \qquad (18)$$

$$\mathbf{P}\left(\left\|\mathbb{P}(s,a) - \hat{\mathbb{P}}_t(s,a)\right\|_1 \leq \rho_t\right) \geq 1 - \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}. \qquad (19)$$

Hence, using Lemma 4 and applying union bounds, we obtain that the following holds with at least probability $1 - (\delta_t^R + \delta_t^P)|\mathcal{S}||\mathcal{A}|$

$$V^*(s) - V^{\hat{\pi}_t^R}(s) \leq 2\rho_t\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right). \qquad (20)$$

Setting $\rho_t = \frac{\epsilon_t}{2}\left(2 + \frac{R_{\max}\gamma}{(1-\gamma)^2}\right)^{-1}$ and $\delta_t^R = \delta_t^P = \frac{\delta_t}{2|\mathcal{S}||\mathcal{A}|}$,

$$\mathbf{P}\left(V^*(s) - V^{\hat{\pi}_t^R}(s) \leq \epsilon_t\right) \geq 1 - \delta_t, \ \forall s \in \mathcal{S}$$

$$\implies \mathbf{P}\left(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_t\right) \geq 1 - \delta_t. \qquad (21)$$

Additionally, the order of n in terms of ϵ_t becomes $n \in O\left(\frac{|\mathcal{S}|}{\epsilon_t^2} + \frac{1}{\epsilon_t^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_t}\right)\right)$. ■

In Theorem 1, we obtain the number of times n each state-action pair needs to be visited to reduce the estimation uncertainty in rewards and transition probabilities to obtain an ϵ_t-optimal policy with probability at least $1-\delta_t$. Now we estimate the total number of exploration steps that are needed to ensure that each state-action pair is visited at least n times.

*Lemma 5 (**Adapted from [30, Theorem 3]**):* For an ergodic Markov chain with state space $\mathcal{S}$ and stationary distribution ϕ_{ss}, let $\tau = \tau(\sigma)$ be the σ-mixing time[1] with $\sigma \leq \frac{1}{8}$. Let ϕ_0 be the initial distribution on $\mathcal{S}$ and let $\|\phi_0\|_{\phi_{\text{ss}}} = \sqrt{\sum_{s\in\mathcal{S}} \frac{\phi_0(s)^2}{\phi_{\text{ss}}(s)}}$. Let $n_{\text{vis}}(s_i, N)$ be the number of times state $s_i \in \mathcal{S}$ is visited until time N. Then, for any $0 \leq \kappa \leq 1$, there exists a constant $c > 0$ (independent of σ and κ) such that:

$$\mathbf{P}\left(n_{\text{vis}}(s_i, N) \geq (1-\kappa)N\phi_{\text{ss}}(s_i)\right) \geq 1 - c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N \phi_{\text{ss}}(s_i)}{72\tau}}. \quad (22)$$

Proof: See [30, Theorem 3] for the proof. ■

In the exploration epoch, at any state $s_i \in \mathcal{S}$, we choose actions uniformly randomly. Consider the Markov chain on $\mathcal{S}$ that is associated with the uniform action selection policy. Let $\{\phi_0(s)\}_{s\in\mathcal{S}}$ and $\{\phi_{\text{ss}}(s)\}_{s\in\mathcal{S}}$, respectively, be the associated initial and stationary distribution. We can also consider an equivalent lifted Markov chain on $\mathcal{S}\times\mathcal{A}$ with states (s_i, a_j) such that $s_i \in \mathcal{S}$ and $a_j \in \mathcal{A}$. The lifted Markov chain has the initial and stationary distribution, $\phi_0(s,a) = \frac{\phi_0(s_i)}{|\mathcal{A}|}$ and $\phi_{\text{ss}}(s,a) = \frac{\phi_{\text{ss}}(s_i)}{|\mathcal{A}|}$, respectively. Hence, we can apply Lemma 5 to obtain the probability of visiting a state-action pair (s,a) at least $(1-\kappa)N\phi_{\text{ss}}(s,a)$ times after N time steps under the uniform action selection policy.

We now design a sequence of exploration and exploitation epochs. Let α_i and β_i be the lengths of the i-th exploration and exploitation epoch, respectively. Let $\phi_{\text{ss}}^{\min} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \phi_{\text{ss}}(s,a)$ and $N_i = \frac{\bar{N}_i}{(1-\kappa)\phi_{\text{ss}}^{\min}}$, where $\bar{N}_i$ is the upper bound in (16) associated with (ϵ_i, δ_i). Let $\delta^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}^{\min}}{72\tau}}$. Note that the desired values of $(1-\kappa)N\phi_{\text{ss}}(s_i,a_j)$ and δ^{α_i} can be obtained by tuning N and κ in (22).

*Theorem 2 (**Regret bound for DSEE algorithm**):* Let the length of exploitation epochs in DSEE be exponentially increasing, i.e. $\beta_i = \eta^i,\ \eta > 1$. Let $\epsilon_i = \eta^{-\frac{i}{3}}$ and $\delta_i = \eta^{-\frac{i}{3}}$ such that $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_i) \geq 1-\delta_i$ after exploration epoch i. For any $\delta \in (0,1)$, set $\delta^{\alpha_i} = \frac{6\delta}{|S||A|\pi^2 i^2}$. Then, the cumulative regret for the DSEE algorithm $\mathbf{R}_T \in O((T)^{\frac{2}{3}}\log(T))$ grows sub-linearly with time T with probability at least $1-\delta$.

Proof: We note that the system state at the start of the i-th exploration epoch might be different from the final state at the end of the $(i-1)$-th exploration epoch. Therefore, we remember the final state of the previous exploration epoch and wait for the same state to restart the new exploration epoch. For the ergodic MDP under the uniform action selection policy (assumption A2), we know that the expected hitting time is finite [32]. Let $U \in \mathbb{R}_{>0}$ be a constant upper bound on the expected hitting time to reach the final state in the previous exploration epoch from an arbitrary initial state in the current exploration epoch. Hence, the cumulative regret during the i-th exploration epoch of length α_i is upper-bounded by $(U+\alpha_i)\mathfrak{R}_{\max}$, where $\mathfrak{R}_{\max} = \frac{R_{\max}}{1-\gamma}$ is the maximum instantaneous regret.

Since at start of the exploitation epoch i of length β_i, $\mathbf{P}(\|V^*(s) - V^{\hat{\pi}_t^R}(s)\|_\infty \leq \epsilon_i) \geq 1-\delta_i$, the expected cumulative regret during the exploitation epoch is $(1-\delta_i)\beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}$. Therefore, the total cumulative regret after k sequences of exploration and exploitation each is upper bounded by:

$$\begin{aligned}\mathbf{R}_{T_k} &\leq \sum_{i=1}^{k}\left((\alpha_i+U)\mathfrak{R}_{\max} + (1-\delta_i)\beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}\right)\\ &\leq \sum_{i=1}^{k}\left((\alpha_i+U)\mathfrak{R}_{\max} + \beta_i\epsilon_i + \delta_i\beta_i\mathfrak{R}_{\max}\right). \quad (23)\end{aligned}$$

Let T_i be the time at the end of the i-th exploitation epoch. Then, $\sum_{j=1}^{k}\beta_j < T_k \leq \sum_{j=1}^{k}(\alpha_j+U) + \sum_{j=1}^{k}\beta_j$. We design the length of the exploitation epochs to be exponentially increasing, i.e., $\beta_i = \eta^i$, for $\eta > 1$. Thus, $T_k \in O(\sum_{j=1}^{k}\eta^j) = O(\eta^k)$. Let $\epsilon_i = \eta^{-di}$ and $\delta_i = \eta^{-gi}$, where $d \in (0,1)$ and $g \in (0,1)$ are constants that we design later. Thus, (23) can be written as:

$$\mathbf{R}_{T_k} \leq \mathfrak{R}_{\max}\left(\sum_{i=1}^{k}\alpha_i + kU\right) + \sum_{i=1}^{k}\eta^{i(1-d)} + \mathfrak{R}_{\max}\sum_{i=1}^{k}\eta^{i(1-g)}. \quad (24)$$

For a state-action pair (s,a), where $s \in \mathcal{S}$ and $a \in \mathcal{A}$, let $\delta_{(s,a)}^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}(s,a)}{72\tau}}$. Therefore, using Lemma 5, at the end of the i-th epoch,

$$\mathbf{P}\left(n_{\text{vis}}(s,a,N_i) \geq (1-\kappa)N_i\phi_{\text{ss}}(s,a)\right) \geq 1 - \delta_{(s,a)}^{\alpha_i}. \quad (25)$$

Recall $\delta^{\alpha_i} := c\|\phi_0\|_{\phi_{\text{ss}}} e^{-\frac{\kappa^2 N_i \phi_{\text{ss}}^{\min}}{72\tau}}$, where $\phi_{\text{ss}}^{\min} := \min_{(s,a)}\phi_{\text{ss}}(s,a)$. Substituting $N_i = \frac{\bar{N}_i}{(1-\kappa)\phi_{\text{ss}}^{\min}}$ in (25),

$$\mathbf{P}\left(n_{\text{vis}}(s,a,N_i) \geq \bar{N}_i\right) \geq 1 - \sum_{m=1}^{|\mathcal{S}||\mathcal{A}|}\delta^{\alpha_i}, \quad (26)$$

for each state-action pair (s,a). Therefore, in N_i time steps of the lifted Markov chain, each (s,a) is visited at least $\bar{N}_i$ times with high probability. Thus, N_i is an upper bound on $\sum_{j=1}^{i}\alpha_j$ with probability in (26). Therefore, using union bounds, with high probability $1 - \sum_{j=1}^{k}\sum_{m=1}^{|\mathcal{S}||\mathcal{A}|}\delta^{\alpha_j}$, $\sum_{j=1}^{k}\alpha_j \leq N_k = \frac{\bar{N}_k}{(1-\kappa)\phi_{\text{ss}}^{\min}}$, and hence,

$$\mathbf{R}_{T_k} \leq \frac{\mathfrak{R}_{\max}\bar{N}_k}{(1-\kappa)\phi_{\text{ss}}^{\min}} + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k}\eta^{i(1-d)} +$$

[1] σ-mixing time for an ergodic Markov chain in the minimal time until the distribution of Markov chain is σ-close in total variation distance to its steady state distribution [31].

$$\mathfrak{R}_{\max} \sum_{i=1}^{k} \eta^{i(1-g)}. \quad (27)$$

Using Theorem 1, $\bar{N}_k \in O\left(\frac{|\mathcal{S}|}{\epsilon_k^2} + \frac{1}{\epsilon_k^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_k}\right)\right)$. Therefore,

$$\begin{aligned}
\mathbf{R}_{T_k} &\leq \frac{\mathfrak{R}_{\max}\lambda}{(1-\kappa)\phi_{\text{ss}}^{\min}}\left(\frac{|\mathcal{S}|}{\epsilon_k^2} + \frac{1}{\epsilon_k^2}\log\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta_k}\right)\right) + \\
&\quad + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k}\eta^{i(1-d)} + \mathfrak{R}_{\max}\sum_{i=1}^{k}\eta^{i(1-g)} \\
&\leq \frac{\mathfrak{R}_{\max}\lambda}{(1-\kappa)\phi_{\text{ss}}^{\min}}\left(\eta^{2dk}|\mathcal{S}| + \eta^{2dk}\log\left(\eta^{gk}|\mathcal{S}||\mathcal{A}|\right)\right) + \\
&\quad + kU\mathfrak{R}_{\max} + \sum_{i=1}^{k}\eta^{i(1-d)} + \mathfrak{R}_{\max}\sum_{i=1}^{k}\eta^{i(1-g)},
\end{aligned} \quad (28)$$

for some constant λ. Recall that $T_k \in O(\eta^k)$, which implies $k \in O(\log(T_k))$. Let Z be the right-hand side of (28). Then, we have:

$$\begin{aligned}
Z &\in O\left((T_k)^{2d} + (T_k)^{2d}\log(T_k) + (T_k)^{(1-d)} + (T_k)^{(1-g)}\right). \\
&\in O((T_k)^{\frac{2}{3}}\log(T_k)), \quad (29)
\end{aligned}$$

by choosing $d = g = \frac{1}{3}$. Hence, the cumulative regret $\mathbf{R}_{T_k} \in O((T_k)^{\frac{2}{3}}\log(T_k))$ grows sub-linearly with time T_k with probability at least $1 - \sum_{i=1}^{k}\sum_{m=1}^{|\mathcal{S}||\mathcal{A}|}\delta^{\alpha_i}$.

Setting $\delta^{\alpha_i} = \frac{6\delta}{|S||A|\pi^2 i^2}$, we have $\sum_{i=1}^{k}\sum_{m=1}^{|\mathcal{S}||\mathcal{A}|}\delta^{\alpha_i} \leq \delta$. ■

V. Conclusions

We proposed a DSEE algorithm with interleaving exploration and exploitation epochs for model-based RL problems that aims to simultaneously learn the system model, i.e., an MDP, and the associated optimal policy. During exploration, we uniformly sample the action in each state and update the estimates of the mean rewards and transition probabilities. These estimates are used in the exploitation epoch to obtain a robust policy with high probability. We designed the length of the exploration and exploitation epochs such that the cumulative regret grows as a sub-linear function of time.

References

[1] J. Xin, H. Zhao, D. Liu, and M. Li, "Application of deep reinforcement learning in mobile robot path planning," in *2017 Chinese Automation Congress (CAC)*. IEEE, 2017, pp. 7112–7116.

[2] P. Gupta, D. Coleman, and J. E. Siegel, "Towards safer self-driving through great pain (physically adversarial intelligent networks)," *arXiv preprint arXiv:2003.10662*, 2020.

[3] A. G. Barto, P. S. Thomas, and R. S. Sutton, "Some recent applications of reinforcement learning," in *Proceedings of the Eighteenth Yale Workshop on Adaptive and Learning Systems*, 2017.

[4] S. Ferretti, S. Mirri, C. Prandi, and P. Salomoni, "Automatic web content personalization through reinforcement learning," *Journal of Systems and Software*, vol. 121, pp. 157–169, 2016.

[5] P. Gupta and V. Srivastava, "Structural properties of optimal fidelity selection policies for human-in-the-loop queues," *arXiv preprint arXiv:2201.09990*, 2022.

[6] ——, "Optimal fidelity selection for human-in-the-loop queues using semi-Markov decision processes," in *American Control Conference*, 2019, pp. 5266–5271.

[7] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[8] L. F. Bertuccelli, "Robust Decision-Making with Model Uncertainty in Aerospace Systems," Ph.D. dissertation, Massachusetts Institute of Technology, 2008.

[9] P. Gupta and V. Srivastava, "On robust and adaptive fidelity selection for human-in-the-loop queues," in *2021 European Control Conference (ECC)*, 2021, pp. 872–877.

[10] W. Wiesemann, D. Kuhn, and B. Rustem, "Robust Markov decision processes," *Mathematics of Operations Research*, vol. 38, no. 1, pp. 153–183, 2013.

[11] A. Nilim and L. El Ghaoui, "Robust Markov Decision Processes with Uncertain Transition Matrices," Ph.D. dissertation, University of California, Berkeley, 2004.

[12] C. J. C. H. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, Cambridge University, 1989.

[13] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[14] A. L. Strehl and M. L. Littman, "An analysis of model-based interval estimation for Markov decision processes," *Journal of Computer and System Sciences*, vol. 74, no. 8, pp. 1309–1331, 2008.

[15] S. M. Kakade, "On the Sample Complexity of Reinforcement Learning," Ph.D. dissertation, University College London, 2003.

[16] T. Jaksch, R. Ortner, and P. Auer, "Near-optimal regret bounds for reinforcement learning," *Journal of Machine Learning Research*, vol. 11, no. 51, pp. 1563–1600, 2010.

[17] T. Lattimore and C. Szepesvári, *Bandit Algorithms*. Cambridge University Press, 2020.

[18] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 5, pp. 759–767, 2013.

[19] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multiarmed bandit with unknown dynamics," *IEEE Transactions on Information Theory*, vol. 59, no. 3, pp. 1902–1916, 2013.

[20] L. Wei and V. Srivastava, "On abruptly-changing and slowly-varying multiarmed bandit problems," in *American Control Conference*, Milwaukee, WI, Jun. 2018, pp. 6291–6296.

[21] N. Nayyar, D. Kalathil, and R. Jain, "On regret-optimal learning in decentralized multiplayer multiarmed bandits," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 1, pp. 597–606, 2016.

[22] L. Wei, A. McDonald, and V. Srivastava, "Multi-robot gaussian process estimation and coverage: Deterministic sequencing algorithm and regret analysis," in *IEEE International Conference on Robotics and Automation*, 2021, pp. 9080–9085.

[23] L. F. Bertuccelli, A. Wu, and J. P. How, "Robust adaptive Markov decision processes: Planning with model uncertainty," *IEEE Control Systems Magazine*, vol. 32, no. 5, pp. 96–109, 2012.

[24] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

[25] G. N. Iyengar, "Robust dynamic programming," *Mathematics of Operations Research*, vol. 30, no. 2, pp. 257–280, 2005.

[26] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[27] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the L_1 deviation of the empirical distribution," *Hewlett-Packard Labs, Tech. Rep*, 2003.

[28] A. Mastin and P. Jaillet, "Loss bounds for uncertain transition probabilities in Markov decision processes," in *IEEE Conference on Decision and Control*, 2012, pp. 6708–6715.

[29] P. Gupta and V. Srivastava, "Deterministic sequencing of exploration and exploitation for reinforcement learning," *arXiv preprint arXiv: 2209.05408*, Sep. 2022.

[30] K.-M. Chung, H. Lam, Z. Liu, and M. Mitzenmacher, "Chernoff-Hoeffding bounds for Markov chains: Generalized and simplified," *arXiv preprint arXiv:1201.0559*, 2012.

[31] D. Aldous, L. Lovász, and P. Winkler, "Mixing times for uniformly ergodic markov chains," *Stochastic Processes and their Applications*, vol. 71, no. 2, pp. 165–185, 1997.

[32] H. Chen and F. Zhang, "The expected hitting times for finite Markov chains," *Linear Algebra and its Applications*, vol. 428, no. 11-12, pp. 2730–2749, 2008.