Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

SAMPLE EFFICIENT EXPLORATION POLICY FOR ASYNCHRONOUS Q-LEARNING

Xinbo Shi[1,2], Jing Dong[3], and Yijie Peng[1,2]

[1]Wuhan Inst. for Artificial Intelligence & Guanghua School of Mgmt., Peking University, Beijing, CHINA
[2]Xiangjiang Laboratory, Changsha, CHINA
[3]Graduate School of Business, Columbia University, New York, NY 10027, USA

ABSTRACT

This paper investigates the sample efficient exploration policy for asynchronous Q-learning from the perspective of uncertainty quantification. Although algorithms like ε-greedy can balance exploration and exploitation, their performances heavily depend on hyperparameter selection, and a systematic approach to designing exploration policies remains an open question. Inspired by contextual Ranking and Selection problems, we focus on optimizing the probability of correctly selecting optimal actions (PCS) rather than merely estimating Q-values accurately. We establish a novel central limit theorem for asynchronous Q-iterations, enabling the development of two strategies: (1) an optimization-based policy that seeks an optimal computing budget allocation and (2) a parameter-based policy that selects from a parametrized family of policies. Specifically, we propose minimizing an asymptotic proxy of Q-value uncertainty with regularization. Experimental results on benchmark problems, including River Swim and Machine Replacement, demonstrate that the proposed policies can effectively identify sample-efficient exploration strategies.

1 INTRODUCTION

Reinforcement learning (RL) has advanced significantly by integrating deep learning techniques, achieving remarkable success in large-scale sequential decision-making problems. Similar to deep learning models, RL relies heavily on high-quality training data, making data collection a crucial factor, particularly for online algorithms such as deep Q-learning (DQN, Mnih et al. 2015), soft actor-critic (SAC, Haarnoja et al. 2018), and twin delayed deep deterministic policy gradient (TD3, Fujimoto et al. 2018). In real-world applications, data collection is often expensive and time-consuming, creating challenges for RL deployment. Despite its importance, online data collection strategies in RL remain underexplored.

For off-policy algorithms such as Q-learning, the exploration policy plays a critical role in data collection. Traditional approaches, such as the ε-greedy and Boltzmann exploration policies for discrete action spaces or Gaussian action noise for continuous spaces, help balance exploration and exploitation (see, Sutton 2018; Szepesvári 2022). However, these methods often require hyperparameter tuning, which is problem-specific and non-trivial. Their applications are largely confined in practice due to the unknown nature of the optimal hyperparameter tailored for specific problems (Auer et al. 2002). When the hyperparameters are not properly chosen, they may fail to ensure adequate exploration in complex environments. In a River Swim problem (Strehl and Littman 2008; Osband et al. 2013), we empirically demonstrate that greedy-guided policies, which predominantly sample the best-estimated action while rarely exploring suboptimal ones, can fail to visit the entire state space within a reasonable simulation budget, leading to poor Q-value estimates.

In this paper, we aim to improve data efficiency in RL by proposing a systematic approach to designing exploration policies. Our focus is on asynchronous Q-learning, where only the Q-value of the current state-action pair is updated at each step. This means that the exploration policy not only governs state-action transitions but also influences the frequency of Q-value updates, further underscoring the importance of exploration design.

We establish a novel central limit theorem (CLT) for asynchronous Q-learning. In the framework of temporal difference (TD) learning, Q-learning can be viewed as a Robbins-Monro type stochastic approximation method (Robbins and Monro 1951) for fixed-point finding problems. Although CLTs have previously been developed for similar algorithms, existing results typically require either global twice differentiability of the mean field (e.g., Fort 2015; Borkar et al. 2024) or rely on mathematically intractable moments of the temporal difference errors (e.g., Hu et al. 2024). Our contribution lies in addressing the non-smooth characteristics of the Q-learning mean field and deriving a closed-form expression for the asymptotic variance of the estimated Q-values. We also note that asynchronous reinforcement learning algorithms can naturally support parallelism to improve exploration through diverse starting points (Mnih 2016). At the same time, the design of exploration strategies for each parallel agent remains a nuanced issue that may benefit from further investigation.

We frame the problem of designing sample-efficient exploration policies through the theoretical lens of contextual Ranking and Selection (CR&S, Shi et al. 2023; Du et al. 2024; Li et al. 2024). The objective in CR&S is to maximize the probability of correctly selecting the best alternative rather than precisely estimating the mean performance of all alternatives. Analogously, we model online data collection as a best-action identification problem, recognizing that it requires significantly more samples to achieve a certain level of statistical accuracy in Q-values than to reach a good level of PCS. This insight motivates our approach of maximizing the probability of correctly identifying the optimal action $a^*(s)$ when it comes to RL, where $a^*(s)$ represents a maximizer of the action value given a state s and will shortly be introduced.

Inspired by the Ranking and Selection literature, which links the rate of decay of the probability of incorrect selection (PCS) to a signal-to-noise ratio (Chen 1995; Glynn and Juneja 2004), we propose an exploration policy similar to Zhu et al. (2023), that minimizes the relative asymptotic uncertainty of Q-value differences between the best and suboptimal actions. This asymptotic uncertainty is a proxy for PCS, guiding more effective action exploration. Unlike standard CR&S problems, where both states and actions can be freely sampled, exploration in Q-learning is constrained by state transitions dictated by the environment. As a result, the finite-sample behavior of Q-value estimates can differ significantly from their asymptotic behavior. This stands in sharp contrast to the behavior of independent and identically distributed (i.i.d.) sample averages, where such discrepancies are typically much smaller and better understood. To address this, we introduce a regularization term to the signal-to-noise index, penalizing cases where state-action pairs receive insufficient samples.

Leveraging this regularized index, we develop two exploration strategies: (1) an optimization-based policy that searches for an optimal computing budget allocation within a general policy space and (2) a parameter-based policy that searches among a finite set of parametrized policies, such as the family of ε-greedy policies. The proposed methods are fully adaptive and require no prior knowledge of the problem instances. Our experimental results demonstrate that the proposed adaptive methods perform at least as well as, if not better than, conventional strategies with post hoc optimal hyperparameters, highlighting their promise for practical applications.

The remainder of this paper is structured as follows. Section 2 reviews the asynchronous Q-learning algorithm and presents our novel central limit theorem. Section 3 introduces our proposed exploration policies. Section 4 evaluates these policies through experiments on the River Swim and Machine Replacement problems. Section 5 concludes with future research directions.

2 ASYNCHRONOUS Q-LEARNING

Consider a discounted infinite-horizon Markov decision process (MDP) defined by the tuple $(\mathscr{S}, \mathscr{A}, R, P, \gamma)$, where $\mathscr{S}$ and $\mathscr{A}$ are state space and action space, respectively, $R : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$ is the mean reward function, P is the transition probability, and $\gamma \in (0,1)$ is the discount factor. Let $S := |\mathscr{S}|$ and $A := |\mathscr{A}|$. The total number of state-action pairs is denoted as $D = |\mathscr{S} \times \mathscr{A}| = SA$. Define $r(s,a)$ as the random reward received after taking action a in state s, and thus the mean is $\mathbb{E}[r(s,a)] = R(s,a)$ and the variance is further denoted by $\sigma^2(s,a)$. For the transition probability $P : \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})$, each row $\mathbf{P}_{s,a} := (P_{s,a}(\cdot))$ represents

the probability vector over the next states given (s,a). Here $\Delta(\mathscr{S})$ denotes the family of probability measures over $\mathscr{S}$. For simplicity, we label states and actions in numerical order and organize matrix entries lexicographically. That is, the $((s-1)A+a-1)$-th row or column corresponds to the (s,a) pair.

We assume a stationary exploration policy π, where $\pi(a|s)$ denotes the probability of taking action a in state s. For such a stationary policy π, the projection matrix $\mathbf{\Pi}^\pi \in \mathbb{R}^{S\times D}$ is defined as:

$$\mathbf{\Pi}^\pi = \mathrm{diag}\{\pi(\cdot|1)^\top, \ldots, \pi(\cdot|S)^\top\} = \begin{bmatrix} \pi(\cdot|1)^\top & & \\ & \cdots & \\ & & \pi(\cdot|S)^\top \end{bmatrix}.$$

Then, the probability transition matrices for state-action pairs and state sequences are given by:

$$\mathbf{P}^\pi := \mathbf{P}\mathbf{\Pi}^\pi \in \mathbb{R}^{D\times D}, \quad \text{and} \quad \mathbf{P}_\pi := \mathbf{\Pi}^\pi \mathbf{P} \in \mathbb{R}^{S\times S}.$$

We will use matrices in bold symbols as probability measures and integral operators interchangeably when no confusion arises.

At each step $t \geq 1$, the agent observes the current environment state s_t, selects an action $a_t \sim \pi(\cdot|s_t)$, and receives a reward $r_t(s_t,a_t)$, which is an independent sample of $r(s_t,a_t)$. The next state is then sampled as $s_{t+1} \sim P_{s_t,a_t}(\cdot)$. We assume that $r_t(s_t,a_t)$ and s_{t+1} are independent of each other conditional on (s_t,a_t). The asynchronous Q-learning algorithm updates the Q-value using the rule:

$$Q_t(s,a) = Q_{t-1}(s,a) + \alpha_t \cdot \mathbf{1}\{(s,a) = (s_t,a_t)\}\left(r_t(s_t,a_t) + \max_{a'\in\mathscr{A}} Q_{t-1}(s_{t+1},a') - Q_{t-1}(s_t,a_t)\right),$$

where $\alpha_t \in [0,1]$ is the step size. With this updating formula, only the Q-value corresponding to the current state-action pair is updated at each step. For simplicity, we consider a widely applied polynomial step size $\alpha_t = k^\rho/(t+k)^\rho$ for some fixed $k > 0$ and $1/2 < \rho \leq 1$.

2.1 Asymptotic variance based on Borkar-Meyn theorem

We establish a central limit theorem (CLT) for Q_t by modeling Q-learning as a stochastic approximation recursion $Q_{t+1} = Q_t + \alpha_{t+1} f(Q_t, \Phi_{t+1})$, where Φ_t is a Markov chain with stationary distribution μ. This recursion aims to find the root of the problem $\mathbb{E}_{\Phi_t\sim\mu}[f(\cdot,\Phi_t)] = 0$. For Q-learning, the Markov chain is $\Phi_t = (s_t,a_t,r_t,s_{t+1})$, with stationary distribution $\mu(s_t,a_t,B,s_{t+1}) = \lambda(s_t,a_t)P(r_t \in B|s_t,a_t)P_{s_t,a_t}(s_{t+1})$ for any Borel set B, where $\lambda(s_t,a_t) = \mu(s_t,a_t,\mathbb{R},\mathscr{S})$ denotes the stationary distribution of (s_t,a_t), and $f:\mathbb{R}^D \to \mathbb{R}^D$ is given by $f_{s,a}(Q,\Phi_t) = \mathbf{1}\{(s,a) = (s_t,a_t)\}(r_t + \max_{a'\in\mathscr{A}} Q(s_{t+1},a') - Q(s_t,a_t))$.

Under the following assumption, we ensure the ergodicity of the Markov chain (s_t,a_t), guaranteeing the stationary distribution μ is well-defined.

Assumption 1 The Markov chain (s_t,a_t) induced by $\mathbf{P}^\pi$ under policy $\pi(a|s)$ is aperiodic and irreducible.

Assumption 1 implies that $\lambda(s,a)$ exists, is unique, and strictly positive for all (s,a). Given this, the Q-learning recursion converges almost surely to Q^* (Borkar and Meyn 2000), the solution to $\mathbb{E}_{\Phi_t\sim\mu}[f(\cdot,\Phi_t)] = 0$, which coincides with the optimal action-value function associated with the optimal action policy $a^*(s) := \arg\max Q^*(s,a)$.

We establish a novel CLT for the normalized Q-values $(Q_t - Q^*)/\sqrt{\alpha_t}$ using an ordinary differential equation technique following Borkar et al. (2024), which establishes a CLT for stochastic approximation recursions by assuming the global smoothness of f. Some notations are necessary before introducing the asymptotic variance. Let $\bar{f}:\mathbb{R}^D \to \mathbb{R}^D$ be given by $\bar{f}(Q) := \mathbb{E}_{\Phi_t\sim\mu}[f(Q,\Phi_t)]$. The vector-valued function $\bar{f}$ is also known as the mean flow of the recursion, representing the mean update of Q_t in the long run. Moreover, let $A := J\bar{f}(Q^*)$ denote the Jacobian matrix of $\bar{f}$, which reflects the marginal contribution of the estimation of Q-table to the mean flow. The following assumption is made to ensure the existence of A.

Assumption 2 For any state $s \in \mathscr{S}$, $\arg\max_{a\in\mathscr{A}} Q^*(s,a)$ is unique.

Assumption 2 is commonly used in the analysis of Q-learning. Then, we have the following lemma that ensures the local differentiability of $\bar{f}$.

Lemma 1 Under Assumption 2, A is well-defined and is equal to $\Lambda(\gamma \mathbf{P}^{\pi^*} - I)$, where Λ is a diagonal matrix with the (s,a)-th entry being $\lambda(s,a)$.

Proof. Fix a positive constant $\varepsilon < Q^*(s,a^*(s)) - \max_{a \in \mathscr{A} \setminus \{a^*(s)\}} Q^*(s)$, $\forall s \in \mathscr{S}$. The constant ε exists due to Assumption 1. For $(s,a) \in \mathscr{S} \times \mathscr{A}$ given, if $\|Q - Q^*\|_\infty < \varepsilon$, then it follows immediately that $\max_{a \in \mathscr{A}} Q(s,a) = Q(s,a^*(s)) =: V(s)$. For such Q, it follows by the definition of $f_{s,a}$ that,

$$\begin{aligned}
\bar{f}_{s,a}(Q) &= \mathbb{E}_{\Phi_t \sim \mu}\left[\mathbf{1}\{(s,a) = (s_t,a_t)\}\left(r_t(s_t,a_t) + \gamma Q(s_{t+1},a^*(s_{t+1})) - Q(s_t,a_t)\right)\right] \\
&= \mathbb{E}_{(s_t,a_t)\sim\lambda}\left[\mathbf{1}\{(s,a) = (s_t,a_t)\}\left(R(s,a) + \gamma \mathbf{P}_{s,a} V - Q(s,a)\right)\right] \\
&= \lambda(s,a)\left(R(s,a) + \gamma \mathbf{P}_{s,a} V - Q(s,a)\right),
\end{aligned}$$

where the second equality follows from the iterated law of expectations. Then, for any $(s',a') \in \mathscr{S} \times \mathscr{A}$, it follows from the definition of V that $\frac{\partial}{\partial Q(s',a')} \bar{f}_{s,a}(Q) = \lambda(s,a)(\gamma \mathbf{P}_{s,a}(s') \mathbf{1}\{a' = a^*(s')\} - \mathbf{1}\{(s,a) = (s',a')\}) = \lambda(s,a)(\gamma \mathbf{P}^{\pi^*}_{s,a}(s',a') - \mathbf{1}\{(s,a) = (s',a')\})$. □

Now, define $\tilde{f}(Q,\Phi) := f(Q,\Phi) - \bar{f}(Q)$ as a noise by which the update of Q-table deviates from the mean flow, and define $\zeta_{t+1}(Q) := \tilde{f}(Q,\Phi_{t+1}) - \mathbb{E}\left[\tilde{f}(Q,\Phi_{t+1}) \middle| \Phi_t\right]$. And denote

$$\Sigma_\zeta := \mathbb{E}_{\Phi_t \sim \mu}\left[\zeta_{t+1}(Q^*)\zeta_{t+1}(Q^*)^\top\right]$$

as the stationary variance of ζ_{t+1}. The Borkar-Meyn theorem (Borkar et al. 2024) characterizes the asymptotic variance of $(Q_t - Q^*)/\sqrt{\alpha_t}$ as the solution to the Lyapunov equation

$$\left[\frac{1}{2}\alpha I + A\right]\Sigma_Q + \Sigma_Q\left[\frac{1}{2}\alpha I + A\right]^\top + \Sigma_\zeta = 0, \tag{1}$$

for some constant α. However, their analysis for general stochastic approximation involves decomposing $\tilde{f}$ as the sum of a martingale difference sequence and a residual term, utilizing a *Poisson* equation, and consequently, Σ_ζ cannot be characterized in closed form. Moreover, they require $\bar{f}$ to be globally twice continuously differentiable. For the special case of Q-learning, there are still two gaps to fill. We first take advantage of the martingale property of the Q-learning mean flow to circumvent the necessity of solving the Poisson equation and thus provide a closed-form characterization of Σ_ζ. We then extend the analysis to Q-learning by working with the local smoothness of $\bar{f}$.

Theorem 1 (CLT) Under Assumptions 1 and 2, and suppose one of the following two cases is true:

(i) $\frac{1}{2} < \rho < 1$, $k > 0$ arbitrary, and $\alpha = 0$;
(ii) $\rho = 1$, $0 < 1/k < 2(1-\gamma)\min_{s,a}\lambda(s,a)$, and $\alpha = 1/k$.

Then, we have

$$\frac{Q_t - Q^*}{\sqrt{\alpha_t}} \overset{w}{\Longrightarrow} N(0,\Sigma_Q),$$

where Σ_Q is given in (1), Σ_ζ is a diagonal matrix with the (s,a)-th entry being $\lambda(s,a)\sigma^2(s,a) + \gamma^2\lambda(s,a)\left(\mathbf{P}_{s,a}V^{*2} - (\mathbf{P}_{s,a}V^*)^2\right)$, where $V^* \in \mathbb{R}^S$ is given by $V^*(s) = Q^*(s,a^*(s))$.

In these two cases, the step sizes are large enough, either because the decaying speed of α_t is sublinear or the scaling constant k is large, so that the TD noise dominates the asymptotic behavior of Q_t. Theorem 1 characterizes Σ_ζ in the Lyapunov equation (1) as the sum of the variance of the random reward and that of the next-step value function. It allows us to analyze the impact of exploration policies on the uncertainty in Q_t through the stationary distribution λ.

3 EXPLORATION POLICIES

We formulate the problem of data collection for online Q-learning as a best-action identification problem. Given a fixed sampling budget T, we aim to maximize the average probability of correctly selecting the best action $a^*(s)$ over each state s after exhausting T samples, which we denote as PCS. Formally, it is defined as:

$$\text{PCS} = \frac{1}{S}\sum_{s\in\mathscr{S}} \mathbb{P}_\pi\left(Q_T(s,a^*(s)) \geq Q_T(s,a),\ \forall a \in \mathscr{A}\backslash\{a^*(s)\}\right).$$

Leveraging asymptotic normality, we approximate $Q_T(s,a^*) - Q_T(s,a)$ using a surrogate normal distribution. Following Glynn and Juneja (2004), we have

$$\begin{aligned}
-\frac{1}{\alpha_T}\log(1-\text{PCS}) &\approx \min_{s\in\mathscr{S}} -\frac{1}{\alpha_T}\sum_{a\in\mathscr{A}\backslash\{a^*(s)\}} \log\mathbb{P}_\pi\left(\sqrt{\alpha_t}(Q_T(s,a^*(s)) - Q_T(s,a)) \leq 0\right) \\
&\approx \min_{s\in\mathscr{S}} -\frac{1}{\alpha_T}\sum_{a\in\mathscr{A}\backslash\{a^*(s)\}} \log\mathbb{P}_\pi\left(Z(s,a) \leq \sqrt{\alpha_T}(Q^*(s,a) - Q^*(s,a^*(s))\right) \\
&= \min_{s\in\mathscr{S}}\min_{a\in\mathscr{A}\backslash\{a^*(s)\}} \frac{1}{2}h(s,a),
\end{aligned}$$

where $Z(s,a)$ is a centered normal variable with variance

$$\tilde{\sigma}^2(s,a) = \Sigma_Q((s,a^*(s)),(s,a^*(s))) - 2\Sigma_Q((s,a^*(s)),(s,a)) + \Sigma_Q((s,a),(s,a)),$$

and $h(s,a) = \left(Q^*(s,a^*(s)) - Q^*(s,a)\right)^2/\tilde{\sigma}^2(s,a)$. Intuitively, $h(s,a)$ takes the form of a signal-to-noise ratio and quantifies the relative uncertainty when comparing two Q-values. In the literature of CR&S, the signal-to-noise ratio has been widely applied to measure the efficiency of exploration policies (Shi et al. 2023; Li et al. 2024).

However, Q-learning often exhibits overestimation bias, which can affect estimating the signal-to-noise ratio. Szepesvári (1997) provides a polynomial bound for the bias, to be specific, $|Q_T(s,a) - Q^*(s,a)| \leq B/T^{\min\lambda(s,a)/\max\lambda(s,a)\cdot(1-\gamma)}$ for some constant B and any pair (s,a). In other words, the bias of Q-values decays slowly when the exploration policy is imbalanced. Since $h(s,a)$ involves the unknown true Q-values, the estimation of the signal-to-noise ratio may be unreliable if certain state-action pairs are insufficiently explored. To mitigate this issue, we introduce a regularization term:

$$\text{ISNR-REG} := \max_{s\in\mathscr{S}}\max_{a\in\mathscr{A}\backslash\{a^*(s)\}} \log\frac{\tilde{\sigma}^2(s,a)}{(Q^*(s,a) - Q^*(s,a^*(s)))^2} + \xi\log\frac{\max\lambda(s,a)}{\min\lambda(s,a)},$$

where $\xi > 0$ is a hyperparameter. The first term penalizes high variance, while the second term discourages extreme imbalance in exploration. Empirically speaking, the choice $\xi = 1$ works fairly well for small scale problems. While for larger scale problems, letting ξ be inversely proportional to the problem size allows for more sampling effort allocated to certain important (s,a) pairs and thus delivers good empirical performance.

This index is a regularized quantification of uncertainty in the differences of Q-values for a finite budget. The following sections introduce two sequential exploration policies based on this index.

3.1 An optimal computing budget allocation for asynchronous Q-learning

The optimal allocation can be calculated by solving an optimization problem that maximizes the proposed index. Similar to Zhu et al. (2023), we consider an optimization problem with objective

$$\min_{\Lambda\geq 0,\Sigma_Q} \text{ISNR-REG} \tag{2}$$

subject to constraints

$$\begin{cases} \left[\frac{1}{2}\alpha I + \Lambda(\gamma \mathbf{P}^{\pi^*} - I)\right]\Sigma_Q + \Sigma_Q\left[\frac{1}{2}\alpha I + (\gamma \mathbf{P}^{\pi^*} - I)^\top \Lambda\right] + \Sigma_\zeta = 0, & (3) \\ \sum_{a' \in \mathscr{A}} \lambda(s,a') = \sum_{\substack{s' \in \mathscr{S} \\ a' \in \mathscr{A}}} \lambda(s',a')\mathbf{P}_{s',a'}(s), \quad \forall s \in \mathscr{S}, & (4) \\ \sum_{\substack{s' \in \mathscr{S} \\ a' \in \mathscr{A}}} \lambda(s',a') = 1. & (5) \end{cases}$$

The second and third constraints ensure that λ remains a valid stationary distribution, while the first constraint links λ to the asymptotic variance of Q-values. Given Λ, the Lyapunov equation uniquely determines Σ_Q, ensuring the feasibility of the optimization.

Proposition 1 Given $\Lambda \geq 0$ fixed, assuming either one of the two cases in Theorem 1 is true, then (3) admits a unique solution. Moreover, if Σ_ζ is symmetric and (semi-)positive definite, then the unique solution Σ_Q is symmetric and (semi-)positive definite as well.

Proposition 1 guarantees that the solution Σ_Q exists and is a valid variance matrix without posing additional constraints requiring that Σ_Q is symmetric and positive definite, which facilitates the optimization. In fact, Proposition 1 holds true when $\mathbf{P}$, π^* and Σ_ζ are replaced by valid estimates. Hence, we propose an optimal computing budget allocation type exploration policy for asynchronous Q-learning based on the above optimization problem.

Algorithm 1 describes the implementation details of this exploration policy. For every T_0 steps where T_0 is a positive integer, we update the exploration policy by solving an empirical version of (2) to have an updated approximation for Λ. Then, the exploration policy follows $\pi(a|s) = \lambda(s,a)/\sum_{a \in \mathscr{A}} \lambda(s,a)$.

Algorithm 1 Optimization-based Exploration Policy

Input: Budget T, interval T_0. Estimation parameter ρ, k, α satisfying assumptions in Theorem 1. Hyperparameter $n_0 = 1$, $R_0 = 0$, $\sigma_0^2 > 0$, $\xi > 0$.
Initialize: timestep $t = 0$; the state $s \in \mathscr{S}$; random Q-table Q; $\forall (s,a) \in \mathscr{S} \times \mathscr{A}$, counts of samples $N(s,a) = 0$, reward mean estimate $\hat{R}(s,a) = R_0$, and reward variance estimate $\hat{\sigma}^2(s,a) = \sigma_0^2/n_0$; $\forall s, s' \in \mathscr{S}, a \in \mathscr{A}$, transition kernel estimate $\hat{P}_{s,a}(s') = 1/S$.
while $t < T$ **do**
  $\forall s \in \mathscr{S}$, $\hat{V}(s) \leftarrow \max_{a \in \mathscr{A}} \hat{Q}(s,a)$, $\hat{a} \leftarrow \arg\max_{a \in \mathscr{A}} \hat{Q}(s,a)$.
  $\hat{\Sigma}_\sigma = \mathrm{diag}(\ldots, \hat{\sigma}^2(s,a), \ldots)$, $\hat{\Sigma}_T = \mathrm{diag}(\hat{\mathbf{P}}\hat{V}^2 - (\hat{\mathbf{P}}\hat{V})^2)$.
  if t is divisible by T_0 **then**
    Solve optimization problem (2) for Λ, with $\mathbf{P}$ replaced by $\hat{\mathbf{P}}$, π^* by $\hat{a}$, Σ_ζ by $\Lambda(\hat{\Sigma}_\sigma + \gamma^2 \hat{\Sigma}_T)$.
    $\forall (s,a) \in \mathscr{S} \times \mathscr{A}$, $\pi(a|s) \leftarrow \lambda(s,a)/\sum_{a \in \mathscr{A}} \lambda(s,a)$.
  end if
  Sample $a \sim \pi(\cdot|s)$ and take action a, observe reward r and next state s'.
  Update Q-table and update count by $N(s,a) \leftarrow N(s,a) + 1$.
  Update the mean and variance of reward by $\tilde{R} \leftarrow [\hat{R}(s,a) * (N(s,a) - 1 + n_0) + r]/(N(s,a) + n_0)$, $\hat{\sigma}^2(s,a) \leftarrow [(\hat{\sigma}^2(s,a) + \hat{R}^2(s,a)) * (N(s,a) - 1 + n_0) + r^2]/(N(s,a) + n_0) - \tilde{R}^2$, and $\hat{R}(s,a) \leftarrow \tilde{R}$.
  $\forall \tilde{s} \in \mathscr{S}$, update transition kernel $\hat{P}_{s,a}(s') \leftarrow [\hat{P}_{s,a}(s') * (N(s,a) - 1 + n_0) + \mathbf{1}(s' = \tilde{s})]/(N(s,a) + n_0)$
  Update $s \leftarrow s'$, $t \leftarrow t + 1$.
end while

3.2 Parameter-based exploration policy selection

Solving the optimization problem can be computationally expensive because it involves a high-dimensional quadratic optimization problem. Optimizing λ requires solving not only the Lyapunov function but also the derivative of the solution with respect to λ. Meanwhile, we note that solving the Lyapunov equation alone for the asymptotic variance can be very efficient. Therefore, we propose a parameter-based policy that selects among a predefined set of exploration policies by evaluating their effectiveness using the ISNR-REG index.

We first introduce an efficient approach in solving the Lyapunov equation. We will re-arrange the matrices defined above indexed by (s,a) pairs for simplicity by letting rows (or columns) indexed by (s,a) with $a \neq a^*(s)$ rank first and be followed by the remaining rows (or columns) indexed by $(s,a^*(s))$. In other words, we will multiply a permutation matrix $R = \left[\cdots \boldsymbol{e}_{(s,a)}^\top \cdots | \boldsymbol{e}_{(1,a^*(1))}^\top \cdots \boldsymbol{e}_{(S,a^*(S))}^\top\right]^\top = [R_1^\top | R_2^\top]^\top$ to each matrix from the left and multiply $R^\top = R^{-1}$ on the right. Then A, Σ_Q and Σ_ζ can be represented by block matrices

$$RAR^\top = \begin{bmatrix} -\Lambda_1 & \gamma\Lambda_1\mathbf{P}_1 \\ 0 & \gamma\Lambda_2\mathbf{P}_2 - \Lambda_2 \end{bmatrix}, \quad R\Sigma_Q R^\top = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad R\Sigma_\zeta R^\top = \begin{bmatrix} \Sigma_{11}^\zeta & 0 \\ 0 & \Sigma_{22}^\zeta \end{bmatrix},$$

where $\Lambda_1 = \text{diag}\{\ldots, \lambda(s,a), \ldots\}_{a\neq a^*(s)}$ is a diagonal matrix in $\mathbb{R}^{(D-S)\times(D-S)}$, $\Lambda_2 = \text{diag}\{\lambda(1,a^*(1)), \ldots, \lambda(S,a^*(S))\} \in \mathbb{R}^{S\times S}$, and $\mathbf{P}_1 = R_1\mathbf{P}$, $\mathbf{P}_2 = R_2\mathbf{P}$. Then, the Lyapunov equation is reduced to

$$(\frac{\alpha}{2}I - \Lambda_1)\Sigma_{11} + \Sigma_{11}(\frac{\alpha}{2}I - \Lambda_1) = \Sigma_{11}^\zeta - \gamma\Lambda_1\mathbf{P}_1\Sigma_{12}^\top - \gamma\Sigma_{12}\mathbf{P}_1^\top\Lambda_1, \tag{6}$$

$$(\frac{\alpha}{2}I - \Lambda_1)\Sigma_{12} + \Sigma_{12}(\frac{\alpha}{2}I + \gamma\mathbf{P}_2^\top\Lambda_2 - \Lambda_2) = -\gamma\Lambda_1\mathbf{P}_1\Sigma_{22}, \tag{7}$$

$$(\frac{\alpha}{2}I + \gamma\Lambda_2\mathbf{P}_2 - \Lambda_2)\Sigma_{22} + \Sigma_{22}(\frac{\alpha}{2}I + \gamma\mathbf{P}_2^\top\Lambda_2 - \Lambda_2) = \Sigma_{22}^\zeta. \tag{8}$$

The first equation can be solved efficiently by

$$\Sigma_{11} = \left(\frac{1}{\alpha - \Lambda_1(i,i) - \Lambda_1(j,j)}\right)_{ij} \circ (\Sigma_{11}^\zeta - \gamma\Lambda_1\mathbf{P}_1\Sigma_{12}^\top - \gamma\Sigma_{12}\mathbf{P}_1^\top\Lambda_1), \tag{9}$$

where $\circ$ denote element-wise product, i.e., $(a_{ij}) \circ (b_{ij}) := (a_{ij}b_{ij})_{ij}$. In practice, we only have to solve the second and the third equations numerically, since Σ_{11} has an explicit characterization. This reduces the number of variables to be solved numerically from D^2 to $2DS - S^2$. Assume $\frac{\alpha}{2}I + \gamma\Lambda_2\mathbf{P}_2 - \Lambda_2 = VDV^{-1}$ where D is a diagonal matrix and V is of full-rank. It follows from equation (8) that

$$DV^{-1}\Sigma_{22}(V^{-1})^\top + V^{-1}\Sigma_{22}(V^{-1})^\top D = V^{-1}\Sigma_{22}^\zeta(V^{-1})^\top, \tag{10}$$

and it turns out that

$$\Sigma_{22} = V\left[\left(\frac{1}{D_i + D_j}\right)_{ij} \circ V^{-1}\Sigma_{22}^\zeta(V^{-1})^\top\right]V^\top.$$

Equation (7) can be solved similarly by

$$\Sigma_{12} = \left[\left(\frac{1}{\alpha/2 - \Lambda_1(i,i) + D_j}\right)_{ij} \circ (-\gamma\Lambda_1\mathbf{P}_1\Sigma_{22}(V^{-1})^\top)\right]V^\top. \tag{11}$$

Let $\{\pi_m : 1 \leq m \leq M\}$ be a set of base policies where $M \geq 1$. We estimate their performance and select the best one sequentially. Specifically, we replace line 7 in Algorithm 1 by Algorithm 2. This method allows efficient adaptation to different problem settings while avoiding direct optimization of λ.

Algorithm 2 Selection of Parameterized Exploration Policy

1: **for** $1 \leq m \leq M$ **do**
2: Estimate stationary distribution $\Lambda_m := \text{diag}(x_m)$ using policy π_m by solving $x_m = x_m \hat{\mathbf{P}}\mathbf{\Pi}^{\pi_m}$.
3: Solve $\hat{\Sigma}_Q$ by (10), (11), and (9) in order, using plug-in estimations.
4: Calculate ISNR-REG$_m$ using plug-in estimations, with Λ replaced by Λ_m and Σ_Q replaced by $\hat{\Sigma}_Q$.
5: **end for**
6: **return** optimal parameterized policy π_{m^*} where $m^* := \arg\min \text{ISNR-REG}_m$.

4 SYNTHETIC EXPERIMENTS

In this section, we evaluate the proposed exploration policies through experiments on two benchmark problems: the River Swim problem and the Machine Replacement problem. Our objective is to assess the effectiveness of our policies in improving sample efficiency and enhancing the PCS. We compare our methods against standard exploration strategies including ε-greedy, Boltzmann exploration, and the Upper Confidence Bound algorithm, which are denoted by 'Egreedy', 'Boltzmann', and 'UCB', respectively. We also carry out an ablation study to illustrate the effect of the proposed exploration policy.

4.1 River Swim Problem

The river swim problem is a widely used benchmark in RL. The environment consists of a finite, discrete set of states $[S] = \{1, 2, \ldots, S\}$, arranged in a linear topology. There are two possible actions: to swim upstream (rightward) or downstream (leftward). The challenge in this problem lies in the fact that the optimal policy requires persistent exploration to reach and assess the state at the rightmost end.

At each step t, when the agent selects the upstream action in state $s_t \in [S]$, it transitions to $s_{t+1} = (s_t + 1) \vee S$ with a small probability p_r, or remains in the same state with probability $1 - p_r - \delta$, or transitions to the downstream state with a small probability δ. Conversely, choosing the downstream action, the agent transitions to $s_{t+1} = (s_t - 1) \wedge 1$ with probability 1. We consider scenarios with $S = 15$ or $S = 30$ states. The upstream transition probability p_r is set to 0.4, while the small probability δ is set to 0.1. Rewards are distributed sparsely across the state space: a small reward of $r(1) = 1$ is given at the leftmost state 1, and a significantly larger reward of $r(S) = 10$ is assigned to the rightmost state S. All intermediate states $\{2, \ldots, S-1\}$ yield zero rewards.

Intuitively, the optimal policy is to choose the upstream action when the current state is close to the upmost state and to choose the downstream action when the state is small. Therefore, this problem underlines sufficient exploration for an efficient policy. In fact, greedy-like algorithms that spend most samples on the optimal action tend to stick to one side of the river and may fail to learn the best action in the upstream states that are never visited, which necessitates adopting a farsighted exploration policy.

We test our optimization-based algorithm 1 denoted as 'Sequential' against standard policies, and compare the parameter-based algorithm 2 with the underlying base policies. We call Matlab built-in optimizer, which is based on the interior point method, to solve approximated (2) in Algorithm 1. The first time the optimizer is called, we randomize the starting point of the algorithm to search for a global optimal solution. After that, we use the optimal solution output from the most recent run as a starting point to enhance the computational efficiency. We also include an 'Oracle' agent which knows the true parameters and thus an optimal solution to the optimization problem (2) is available. Throughout the River Swim problem, we use constant $\xi = 1$ for the regularization term.

For the benchmark policies, in addition to 'Egreedy', 'Boltzmann', and 'UCB', we also test a heuristic algorithm, which explores the upstream action randomly with a fixed probability across all states, which we denote as 'Random'. Hyperparameters of these benchmarks are indicated in the legend and are picked manually after the experiments to maximize the ultimate PCS. In other words, these hyperparameters are ex-post optimal but are unknown beforehand. In this experiment, we initialize the Q-table randomly. Specially, each Q-value is independently and identically distributed following $U([0, 40])$.

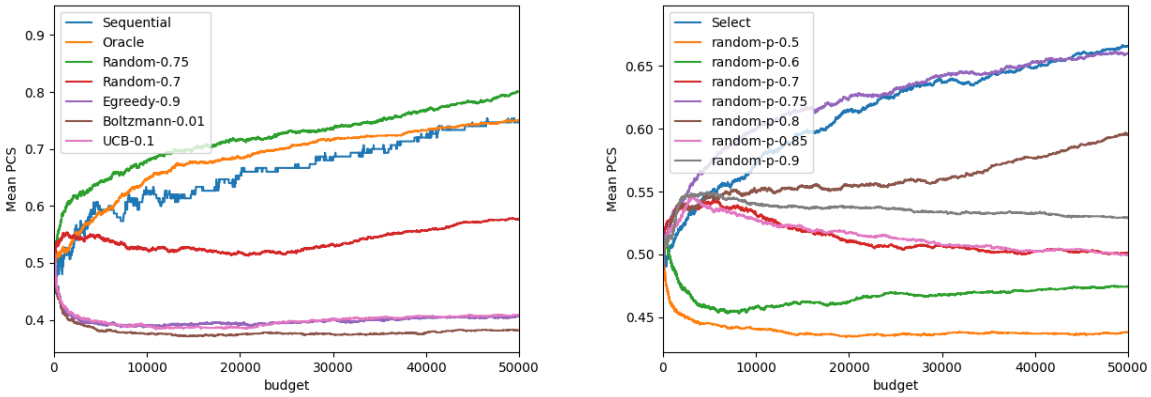

Figure 1: The average probability of correct selection based on 100 macro-replications. Left: a small instance with 15 states. Right: a larger instance with 30 states.

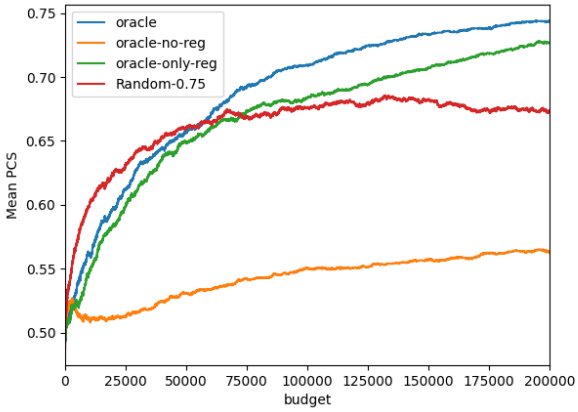

Figure 2: The average probability of correct selection in states based on 100 macro-replications. This instance includes 30 states.

The left panel in Figure 1 presents the PCS for the small instance with $S = 15$ states. Our results indicate that traditional greedy-like algorithms tend to get trapped in low-reward states and fail to learn the optimal actions eventually in most states with insufficient exploration. The random policy with hyperparameter 0.75 achieves higher PCS than the proposed methods. However, it requires fine-tuned parameters to perform well, which is unknown in practice. A slightly smaller hyperparameter such as 0.7 does not deliver such a large PCS. In contrast, our optimization-based exploration policy sequentially adjusts exploration, leading to improved PCS. Surprisingly, the sequential algorithm based on plug-in estimations of the Q-table and the transition kernel performs almost as well as the Oracle one, implying its robustness.

The right panel in Figure 1 tests algorithm 2 using 'random' as a base policy. We consider a larger instance with $S = 30$ states. The optimal base policy is with hyperparameter 0.75. We observe that a slightly perturbed hyperparameter, e.g., 0.7, can lead to poor performances and even a decreasing PCS curve. Nevertheless, the proposed parameter-based approach selects effective policies without requiring extensive hyperparameter tuning, demonstrating comparable performance to the best-tuned baselines.

Ablation study. To better understand the impact of different components in our proposed policies, we conduct an ablation study. We compare three variants: (1) a version without regularization in index ISNR-REG, (2) a version that only includes the regularization term, and (3) the full version of our approach. Equivalent, we set $\xi = 0, \infty, 1$ for these variants, and they are denoted as 'oracle-no-reg', 'oracle-only-reg', 'oracle', respectively, in Figure 2. The results reveal that removing the regularization term leads to suboptimal exploration, as certain state-action pairs remain under-sampled. Conversely, using only the regularization term without the signal-to-noise ratio objective improves exploration balance but fails to efficiently prioritize valuable state-action pairs. The full version of our method achieves the highest PCS, confirming the effectiveness of our regularization strategy in combination with the signal-to-noise ratio.

4.2 Machine Replacement Problem

The Machine Replacement problem models a maintenance decision process where an agent must decide whether to continue using a machine or replace it to minimize long-term costs (Lake and Muhlemann 1979). It balances immediate replacement costs against the cost of holding an old machine at the risk of system failures. Suppose k machines are running in an infinite-period discrete-time system. In period $t = 1, 2, \ldots,$ each machine $i \leq k$ has a status $s_{t,i} \in [n]$ indicating its durability, where n is the number of possible statuses. The state of the entire system and the state space are denoted by $s_t = (s_{t,1}, s_{t,2}, \ldots, s_{t,k})$ and $\mathscr{S} = [n]^k$, respectively. In each period t, the machine operator takes action $a_{t,i} \in \{0, 1\}$ for each $i \leq k$ such that it is equal to 1 when machine i is replaced. Therefore, the action is denoted by $a_t = (a_{t,1}, a_{t,2}, \ldots, a_{t,k})$ and the action space is $\mathscr{A} = \{0, 1\}^k$. If a machine is replaced, its status will be reset to n in the next period. Otherwise, the status decrease by the level of its wear and tear.

The machine replacement problem can be viewed as a variant of the multi-item inventory problem with a fixed order-up-to level and without backlogs. We consider a general parallel machine replacement problem with multiple machines where the holding cost and the replacement cost of each machine are contingent on the status of other machines (Childress and Durango-Cohen 2005). The holding cost of a set of worn machines with state s_t is

$$H(s_t) = \mathbf{1}\{\min_{1 \leq i \leq k} s_{t,i} = 0\} - 0.05 \cdot \min_{1 \leq i \leq k} s_{t,i},$$

and the replacement cost is

$$O(a_t) = \begin{cases} 0.2 + \sum_{i=1}^{k} a_{t,i}, & \exists\ a_{t,i} = 1, \\ 0, & \text{o.w.} \end{cases}$$

The holding cost is the out-of-stock cost, which equals to 1 when any machine can not longer work, minus the salvage value, which is assumed to be linear in the smallest durability of machines. The replacement cost consists of a fixed cost and a variable cost. The reward function is $R(s, a) = -H(s) - O(a)$.

We model the transition of states as compound Poisson distributions, similar to multi-item inventory problems in Federgruen et al. (1984), to capture the correlation of statuses of machines. Consider k tasks and each task $\tau \leq k$ features a set M_τ, which contains all machines that participate in task τ. In each period t, task τ causes wear and tear to participating machines at level amounting to $D_{\tau,t} \overset{i.i.d.}{\sim} \mathrm{P}(\lambda_\tau)$, where $\mathrm{P}(\lambda_\tau)$ denotes a Poisson distribution with expectation λ_τ. The status of machine i follows the dynamic $s_{t+1,i} = (s_{t,i} - \sum_{\tau \leq k} \mathbf{1}\{i \in M_\tau\} D_{\tau,t}) \vee 0$. That is, the new status either equals to the old status minus the sum of wear-and-tear levels or to 0 if the former quantity is less than 0. In this experiment, for $\tau = 1$, we let $M_1 = [k]$ and $\lambda_1 = 1/k$. For $2 \leq \tau \leq k$, we let $M_\tau = \{1, \tau\}$ and $\lambda_\tau = 1/k$. Unlike the threshold policy for single-machine replacement problems or the (s, S) inventory policy, the optimal policy can be highly non-trivial.

Figure 3 illustrates the performance of algorithm 2 on two instances of the problem with different base policies. The first instance uses $k = 2, n = 3$, and thus the size the state space is $n^k = 9$. Since there are no known structured exploration policies tailored for this problem, we consider ε-greedy algorithms as

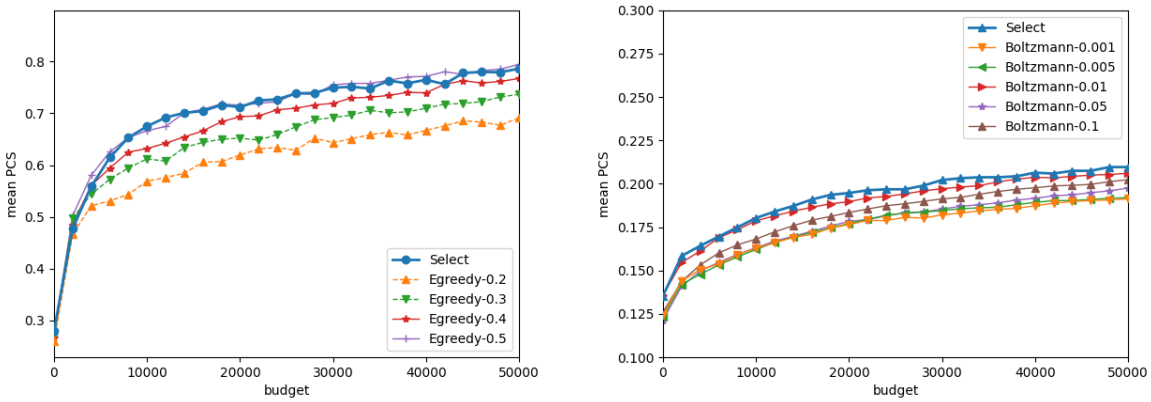

Figure 3: The average probability of correct selection in states based on 100 macro-replications. Left: 2 machines and 3 statuses for each machine. The base policies are ε-greedy policies. Right: 3 machines and 5 statuses for each machine. The base policies are Boltzmann exploration policies.

base policies. The candidate hyperparameters range from 0.2 to 0.5, where hyperparameter 0.5 performs the best. As shown in the left panel, the parameter-based policy works almost as well as the optimal base policy. In the second experiment, we set $k = 3, n = 5$. The size of the state space is $n^k = 125$. We use Boltzmann exploration as base policies which balances the exploration of suboptimal actions in a more sophisticated way than ε-greedy. The proposed policy is superior to base policies, implying that it can not only select a satisfying hyperparameter for the base policy, but also surpass base policies due to the flexibility of switching between policies adaptively.

5 CONCLUSION

We investigated sample-efficient exploration policies for asynchronous Q-learning, drawing insights from CR&S problems. By establishing a novel central limit theorem for asynchronous Q-iterations, we developed and analyzed two exploration strategies: an optimization-based policy and a parameter-based policy. Prioritizing PCS over precise Q-value estimation, our approach enhances data efficiency in reinforcement learning. This work lays the groundwork for further exploration. Future research may extend our methods to more general Q-learning algorithms, such as double Q-learning and adaptive step-size approaches. Additionally, applying these strategies to deep Q-learning presents an intriguing direction.

ACKNOWLEDGMENTS

The work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants 72325007, 72250065, and 72022001. This work was also supported in part by Xiangjiang Laboratory Key Project under Grant 23XJ02004 and the Science and Technology Innovation Program of Hunan Province under Grant 2024RC7003.

REFERENCES

Auer, P., N. Cesa-Bianchi, and P. Fischer. 2002. "Finite-time Analysis of the Multiarmed Bandit Problem". *Machine Learning* 47:235–256.

Borkar, V., S. Chen, A. Devraj, I. Kontoyiannis, and S. Meyn. 2024. "The ODE Method for Asymptotic Statistics in Stochastic Approximation and Reinforcement Learning". *arXiv preprint arXiv: 2110.14427*.

Borkar, V. S., and S. P. Meyn. 2000. "The ODE Method for Convergence of Stochastic Approximation and Reinforcement Learning". *SIAM Journal on Control and Optimization* 38(2):447–469.

Chen, C.-H. 1995. "An effective approach to smartly allocate computing budget for discrete event simulation". In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, Volume 3, 2598–2603.

Childress, S., and P. Durango-Cohen. 2005. "On parallel machine replacement problems with general replacement cost functions and stochastic deterioration". *Naval Research Logistics (NRL)* 52(5):409–419.

Du, J., S. Gao, and C.-H. Chen. 2024. "A contextual ranking and selection method for personalized medicine". *Manufacturing & Service Operations Management* 26(1):167–181.

Federgruen, A., H. Groenevelt, and H. C. Tijms. 1984. "Coordinated Replenishments in a Multi-Item Inventory System with Compound Poisson Demands". *Management Science* 30(3):344–357.

Fort, G. 2015. "Central limit theorems for stochastic approximation with controlled Markov chain dynamics". *ESAIM: Probability and Statistics* 19:60–80.

Fujimoto, S., H. Hoof, and D. Meger. 2018. "Addressing function approximation error in actor-critic methods". In *International conference on machine learning*, 1587–1596. PMLR.

Glynn, P., and S. Juneja. 2004. "A large deviations perspective on ordinal optimization". In *In Proceedings of the 2004 Winter Simulation Conference (WSC)*, Volume 1. IEEE.

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In *International conference on machine learning*, 1861–1870. PMLR.

Hu, J., V. Doshi *et al*. 2024. "Central Limit Theorem for Two-Timescale Stochastic Approximation with Markovian Noise: Theory and Applications". In *International Conference on Artificial Intelligence and Statistics*, 1477–1485. PMLR.

Lake, D. H., and A. P. Muhlemann. 1979. "An Equipment Replacement Problem". *Journal of the Operational Research Society* 30(5):405–411 https://doi.org/10.1057/jors.1979.100.

Li, H., H. Lam, and Y. Peng. 2024. "Efficient learning for clustering and optimizing context-dependent designs". *Operations Research* 72(2):617–638.

Mnih, V. 2016. "Asynchronous Methods for Deep Reinforcement Learning". *arXiv preprint arXiv:1602.01783*.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, *et al*. 2015. "Human-level control through deep reinforcement learning". *nature* 518(7540):529–533.

Osband, I., D. Russo, and B. Van Roy. 2013. "(More) efficient reinforcement learning via posterior sampling". *Advances in Neural Information Processing Systems* 26.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics* 22(3):400 – 407 https://doi.org/10.1214/aoms/1177729586.

Shi, X., Y. Peng, and G. Zhang. 2023. "Top-Two Thompson Sampling for Contextual Top-mc Selection Problems". *arXiv preprint arXiv: 2306.17704*.

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

Sutton, R. S. 2018. "Reinforcement learning: An introduction". *A Bradford Book*.

Szepesvári, C. 1997. "The Asymptotic Convergence-Rate of Q-learning". In *Advances in Neural Information Processing Systems*, edited by M. Jordan, M. Kearns, and S. Solla, Volume 10: MIT Press.

Szepesvári, C. 2022. *Algorithms for reinforcement learning*. Springer nature.

Zhu, Y., J. Dong, and H. Lam. 2023. "Uncertainty Quantification and Exploration for Reinforcement Learning". *Operations Research* https://doi.org/10.1287/opre.2023.2436.

AUTHOR BIOGRAPHIES

XINBO SHI is a Ph.D. candidate in Guanghua School of Management at Peking University, Beijing, China. His research interest includes simulation optimization and machine learning. His email address is xshi@stu.pku.edu.cn.

JING DONG is a DeRosa Family Associate Professor in the Decision, Risk, and Operations division at the Graduate School of Business, Columbia University. Her primary research interests are in applied probability and stochastic simulation, with an emphasis on applications in service operations management. Her current research focuses on developing data-driven stochastic modeling to improve patient flow in hospitals. Her e-mail is jd2736@columbia.edu.

YIJIE PENG is an Associate Professor in Guanghua School of Management at Peking University, Beijing, China. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare. His email address is pengyijie@pku.edu.cn.