A Deep Proactive Exploration Policy Based on Asymptotic Statistics for Asynchronous Q-Learning

Xinbo Shi, Jinyang Jiang, Ruihan Zhou, Yijie Peng
Guanghua School of Management
Peking University

Jing Dong
Graduate School of Business
Columbia University

Abstract

This paper presents a new methodology that adaptively optimizes exploration policy for the reinforcement learning problem. We shift the objective from value-function accuracy to direct policy identification, using the probability of correct selection as a metric. We establish a new central limit theorem for asynchronous Q-learning with adaptive step sizes and propose a regularized signal-to-noise ratio index for exploration policy designing. To address the computational cost of the high-dimensional optimization, we propose a novel pipeline with an offline, simulation-based proactive learning loop. This loop trains a deep neural network to serve as a fast, low-dimensional proxy for the complex optimization problem, allowing for efficient online policy updates in real-world applications. We validate our approach on the challenging RiverSwim environment, demonstrating superior performance compared to standard exploration heuristics.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in sequential decision-making by combining statistical learning with trial-and-error exploration. Yet, a central bottleneck remains: data collection in online RL is costly, and standard exploration heuristics such as ϵ-greedy or Boltzmann exploration depend heavily on hyperparameter tuning, often failing to ensure sufficient coverage in complex environments [Sutton and Barto, 2018, Szepesvari, 2010]. For Q-learning in particular, the exploration policy not only governs state transitions but also directly affects the update frequency of Q-values, making its design crucial for sample efficiency.

The primary goal of RL is to find an optimal policy, $a^*(s)$. Value-based methods like Q-learning traditionally approach this by estimating the optimal action-value function, $Q^*(s, a)$, typically through minimizing temporal-difference errors, which are often treated in practice as regression-like objectives with metrics such as mean squared error (MSE) or Bellman error. However, a critical limitation is that value function accuracy is a poor proxy for policy accuracy Fujimoto et al. [2022]. A model with low value error may still fail to correctly rank the optimal action and vice versa. Motivated by this misalignment, our work shifts the optimization goal from value estimation to direct policy identification. We introduce a more direct metric, the **probability of correct selection (PCS)**, which measures the likelihood of correctly identifying the optimal action $a^*(s)$ in a given state. This paradigm is inspired by the Ranking and Selection (R&S) problems in simulation optimization [Chen et al., 2020, Peng et al., 2018, Shi et al., 2023], which are designed to find the best alternative with minimal samples and high probability Hong [2018]. Unlike traditional RL exploration strategies that focus on reducing uncertainty across all actions Lu and Van Roy [2019], our approach integrates R&S experiences into an off-policy Q-learning framework, designing a behavior policy specifically to efficiently differentiate between optimal and suboptimal actions.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

We develop a new functional CLT for asynchronous Q-learning that explicitly accommodates adaptive step sizes, extending beyond prior analyses restricted to polynomial or constant step-size rules. This new CLT provides closed-form variance characterization through Lyapunov equations, enabling principled uncertainty quantification for Q-value estimates. Previous results in the literature often require global smoothness assumptions on the mean flow [Paulin, 2015, Borkar et al., 2024], do not admit adaptive step sizes [Szepesvári, 1997, Li et al., 2023], or fail to provide explicit variance formulas [Xie and Zhang, 2022, Zhang and Xie, 2024]. By contrast, our framework explicitly handles the non-smooth nature of the Q-learning mean flow and yields tractable variance expressions.

To optimize the exploration policy, we propose to optimize an asymptotic surrogate of the PCS [Glynn and Juneja, 2004, Zhu et al., 2023]. While variance-aware exploration can in principle be formulated as a high-dimensional optimization problem involving $D + D^2$ variables (with D the number of state-action pairs), directly solving it is computationally prohibitive. To overcome this, we design a pipeline that first generates problem instances and simulates exploration policies, then uses CLT-based variance analysis to quantify uncertainty, followed by learning and solving a reduced proxy optimization with only D variables. The resulting objective values are then used to train a neural network as a deep objective proxy. This surrogate allows fast and scalable approximation of the optimization objective, amortizing computation across tasks and enabling efficient online policy updates in real-world interactions. Numerical experiments on a pure-exploration RiverSwim problem demonstrates the feasibility and superiority of our method.

2 Methodology

Consider a discounted infinite-horizon Markov decision process (MDP) denoted as $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S} = [S]$ and $\mathcal{A} = [A]$ are the state and action spaces, respectively. The function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ specifies the expected immediate reward, P is the transition probability, and $\gamma \in (0, 1)$ is the discount factor. Here, $0 < S, A < \infty$ denote the cardinalities of the state and action spaces. A standard result in dynamic programming is that the maximum achievable expected cumulative reward can be represented by the optimal action-value function Q^*. This function satisfies the Bellman equation:

$$Q^*(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P_{s,a}(s') \max_{a' \in \mathcal{A}} Q^*(s', a').$$

The corresponding optimal policy is stationary and state-dependent, given by

$$a^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a), \quad s \in \mathcal{S}.$$

Suppose a stationary exploration policy $\pi(a|s)$ is implemented. At step $t \geq 1$, an agent observes state s_t, selects an action $a_t \sim \pi(\cdot|s_t)$, and receives a reward r_t. The next state is sampled as $s_{t+1} \sim P_{s_t,a_t}(\cdot)$. We assume r_t and s_{t+1} are conditionally independent given (s_t, a_t). The asynchronous Q-learning algorithm updates the Q-value for the state-action pair (s_t, a_t) using the following rule:

$$Q_{t+1}(s,a) = Q_t(s,a) + \beta_t(s,a)\mathbf{1}_{(s,a)}(s_t,a_t)\left(r_t(s_t,a_t) + \gamma \max_{a' \in \mathcal{A}} Q_t(s_{t+1}, a') - Q_t(s_t,a_t)\right)$$

where $\beta_t(s,a) \in (0,1)$ is a randomized step size. Specifically, we use $\beta_t(s,a) = p_t / w_t(s,a)$, where $p_t = k^\rho/(t+k)^\rho$ for $k > 0$ and $1/2 < \rho \leq 1$. And, $w_t(s,a)$ is a weight satisfying Assumption 3 in the online supplement. For example, $w_t(s,a) = (N_{t-1}(s,a) + 1)/(t+1)$ can be utilized for variance reduction and facilitating convergence, especially when the state space is large. Here, $N_{t-1}(s,a) = \sum_{\tau=1}^{t-1} \mathbf{1}_{(s,a)}(s_t, a_t)$ stands for the visiting times of (s,a).

We propose to use the average PCS as the performance metric. Formally, it is defined as

$$\text{PCS} := \frac{1}{S} \sum_{s \in \mathcal{S}} \mathbb{P}\left(Q_t(s, a^*(s)) > Q_t(s,a),\ \forall a \neq a^*(s)\right).$$

Each summand dictates the probability of correctly identifying $a^*(s)$ as the best action after exhausting t samples, given a state s. Although PCS is not analytically tractable, we have the following approximation based on the large deviations principle given by Glynn and Juneja [2004], i.e.,

$$-p_t \log(1 - \text{PCS}) \approx \min_{s \in \mathcal{S}} \min_{a \in \mathcal{A} \setminus \{a^*(s)\}} \frac{1}{2} \left(Q^*(s, a^*(s)) - Q^*(s,a)\right)^2 / \tilde{\sigma}^2(s,a),$$

where $\tilde{\sigma}^2(s,a)$ represents the asymptotic variance of $Q_t(s,a^*(s)) - Q_t(s,a),\ \forall s \in \mathcal{S},\ a \neq a^*(s)$, if it exists. Moreover, the above approximation takes the explicit form of a signal-to-noise ratio (SNR). A high SNR means that the expected difference is large relative to the variability, hence the difference is reliably detectable. These observations drive us to take SNR as a surrogate for PCS.

In implementation, we use plug-in estimate to calculate SNR, substituting unknown parameter Q^* with Q_t. Let $\lambda(s,a)$ denote the stable distribution of the Markov chain $\{(s_t,a_t) : t \geq 1\}$ given exploration policy $\pi(a|s)$. We notice Q_t is typically biased for Q^* when t is small, leading to a biased estimation for SNR. To be specific, Szepesvári [1997] provides a polynomial bound for the bias $\mathbb{E}|Q_t(s,a) - Q^*(s,a)| \leq C/\exp\{\ln t \cdot \min \lambda(s,a)/\max \lambda(s,a) \cdot (1-\gamma)\}$ for some constant C. To mitigate the bias issue, we introduce a regularized term to the surrogate objective:

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

where ξ is a hyperparameter.

It remains to characterize asymptotic behavior of $Q_t(s,a^*(s)) - Q_t(s,a)$. For this purpose, we develop a novel central limit theorem for Q_t leveraging the ordinary differential equation method in Borkar et al. [2024]. Suppose $D = SA$ and $Q_t = (Q_t(s,a)) \in \mathbb{R}^D$ is a vector arranged in lexicographical order. Denote W^*, Λ, Σ_R and Σ_T as diagonal matrices with the $((s-1)A + a)$-th element on the diagonal being $w^*(s,a) := \lim_{t\to\infty} w_t(s,a)$, $\lambda(s,a)$, $\sigma^2(s,a)$, and $P_{s,a}V^{*2} - (P_{s,a}V^*)^2$, respectively. Here, $\sigma^2(s,a)$ is the variance of the reward received for action a at state s, and $V^* = (V^*(s))$ is the state-value function with $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s,a)$. Moreover, let P^* be the transition kernel of (s_t,a_t) when the optimal action policy a^* is applied. Then, we have:

Theorem 1 (CLT). *Under Assumptions 1, 2, and 3, then the asynchronous Q-learning algorithm is consistent, i.e., $Q_t \to Q^*$ as $t \to \infty$. Moreover, if the initial Q-value Q_0 has eighth bounded absolute moment, and p_t satisfies either one of the following:*

1. $\frac{1}{2} < \rho < 1$, $k > 0$, *and* $\alpha = 0$*; or*
2. $\rho = 1$, $0 < 1/k < (1-\gamma) \cdot \min_{s,a} \lambda(s,a)/w^*(s,a)$, *and* $\alpha = 1/k$,

then, $\sqrt{p_{t-1}}^{-1}(Q_t - Q^*) \xrightarrow{w} N(0, \Sigma_Q)$*, where* Σ_Q *solves the Lyapunov equation*

$$A\Sigma_Q + \Sigma_Q A^\top + B = 0.$$

where $A = [\frac{1}{2}\alpha I + (W^*)^{-1}\Lambda(\gamma P^* - I)]$, $B = (W^*)^{-2}\Lambda(\Sigma_R + \gamma^2\Sigma_T)$.

The Assumptions 1, 2, and 3 could be found in the online supplement. The proof of Theorem 1 is deferred to the future version of this project.

The optimal allocation is then calculated by solving an optimization problem that maximizes the proposed index. Similar to Zhu et al. [2023], we consider optimizing the following problem:

$$\min_{\Lambda \geq 0, \Sigma_Q} \text{ISNR-REG} \tag{1}$$

$$s.t. \begin{cases} \left[\dfrac{1}{2}\alpha I + (W^*)^{-1}\Lambda(\gamma P^* - I)\right]\Sigma_Q + \Sigma_Q\left[\dfrac{1}{2}\alpha I + (\gamma P^* - I)^\top \Lambda (W^*)^{-1}\right] \\ \qquad\qquad = -(W^*)^{-2}\Lambda(\Sigma_R + \gamma^2\Sigma_T), \\ \displaystyle\sum_{a' \in \mathcal{A}} \lambda(s,a') = \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \lambda(s',a') P_{s',a'}(s), \quad \forall s \in \mathcal{S}, \\ \displaystyle\sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \lambda(s',a') = 1, \quad \lambda(s,a) > \varepsilon, \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \end{cases}$$

where $\varepsilon > 0$ is introduced to force enough exploration, especially when t is small. The first equation links the essential decision variable Λ to the asymptotic variance matrix Σ_Q. The second equations ensure $\lambda(s,a)$ is compatible with the transition probability P. The third equation normalizes $\lambda(s,a)$. Then, the exploration policy follows $\pi(a|s) = \lambda(s,a)/\sum_{a\in\mathcal{A}} \lambda(s,a)$. In implementation, we begin with an initial exploration policy π_0, carry out the Q-learning algorithm, and periodically re-solve the optimization problem (1). The details are described in Algorithm 1, which we refer to as Q-learning optimal computing budget allocation (Q-OCBA) algorithm.

3 Deep Q-OCBA Exploration Policy

The Q-OCBA algorithm, as a sequential exploration policy, necessitates the repeated re-solving of the optimization problem (1). This optimization module is computationally prohibitive due to several factors. First, the problem's formulation is non-convex and involves quadratic constraints with an indefinite coefficient matrix. Second, the high dimensionality of the decision variables contributes significantly to the computational burden, as the variance matrix is D^2-dimensional and the Λ matrix contains D non-degenerate elements. Consequently, the high computational time makes the approach impractical for real-world applications.

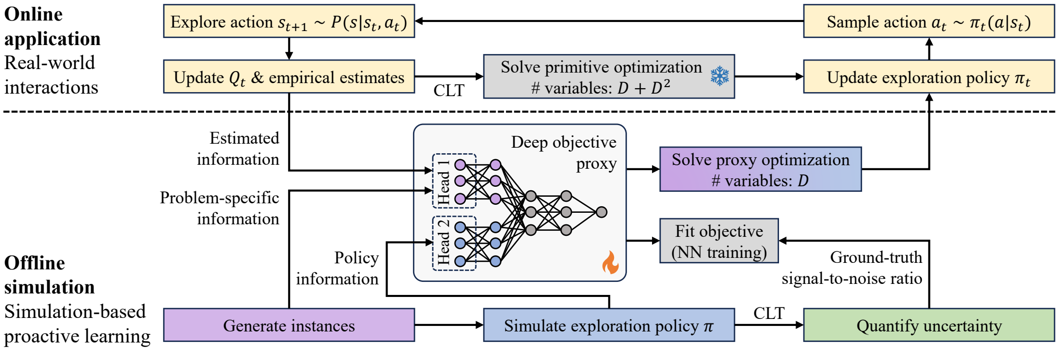
Online application
Real-world interactions
Explore action $s_{t+1} \sim P(s|s_t, a_t)$
Sample action $a_t \sim \pi_t(a|s_t)$
Update Q_t & empirical estimates
CLT
Solve primitive optimization
# variables: $D + D^2$
Update exploration policy π_t
Estimated information
Problem-specific information
Policy information
Deep objective proxy
Head 1
Head 2
Solve proxy optimization
# variables: D
Fit objective
(NN training)
Ground-truth signal-to-noise ratio
Offline simulation
Simulation-based proactive learning
Generate instances
Simulate exploration policy π
CLT
Quantify uncertainty

Figure 1: A learning pipeline with two workflows: an online primitive workflow above the dashed line for real-world interactions, and a new offline workflow below it that uses pretrained neural networks to replace the slow optimization module.

We then propose a deep-learning driven methodology to mitigate this issue. The new approach is motivated by the observation that the cause of the curse of dimensionality arises from the intermediate variable Σ_Q. Although Σ_Q can be solved from a linear system when Λ is given, the need to minimize the objective as a function of Λ necessitates differentiating Σ_Q w.r.t. Λ, applying the chain rule. In order to facilitate optimization, we train a deep neural network to express $\mathrm{ISNR} - \mathrm{REG}$ as a function in merely Λ. To train the network, we simulate problem instance parameters proactively, including α, γ, ξ, P^*, and W^*. For each instance, we then simulate feasible stable distribution Λ's and evaluate the objective values exactly. The simulated data are gathered to feed the neural network. With the pretrained network, we obtain a D-dimensional problem with a neural network surrogate objective, subject to the second and third constraints as in (1), which can be solved easily using non-convex optimizers. The gradients of the surrogate are easily accessible, facilitating the optimization. The pipeline is illustrated in Figure 1.

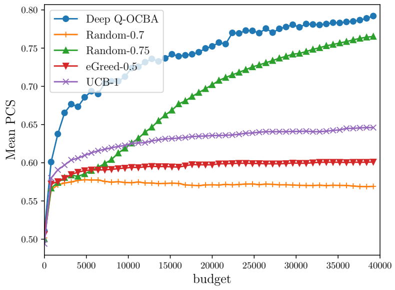

Figure 2: PCS curve as a function of simulation budget, estimated based on 100 macro-replications.

We demonstrate the performance of the proposed methodology in a pure exploration problem known as RiverSwim. It is commonly used to validate exploration policies in reinforcement learning [Osband et al., 2013, Hans et al., 2016]. The problem details can be found in the online supplement. In an instance with $S = 30$ states and 2 actions, we compare the proposed algorithm with a problem-dependent heuristic, random exploration. It relies on a single parameter $p \in (0, 1)$ and its performance is highly parameter-sensitive. As demonstrated in Figure 2, the random exploration policy $p = 0.75$ performs best among the compared methods. Unfortunately, the optimal parameter is unknown ex ante. Another random exploration policy with $p = 0.7$ delivers poor PCS. The proposed deep Q-OCBA algorithm, is superior to random exploration with the oracle optimal parameter. Its edge over other methods when the budget is small further emphasizes its advantage in practice. Further discussions can be found in the online supplement (https://github.com/anduiiin/NIPSWS_RSQLearning-online-supplement).

References

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Csaba Szepesvari. *Algorithms for Reinforcement Learning*. Morgan and Claypool Publishers, 2010. ISBN 1608454924.

Scott Fujimoto, Shixiang Gu, and Sergey Levine. A minimalist approach to offline reinforcement learning. *Proceedings of the 39th International Conference on Machine Learning (PMLR)*, 162, 2022. URL `https://proceedings.mlr.press/v162/fujimoto22a/fujimoto22a.pdf`.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

Yijie Peng, Edwin KP Chong, Chun-Hung Chen, and Michael C Fu. Ranking and selection as stochastic control. *IEEE Transactions on Automatic Control*, 63(8):2359–2373, 2018.

Xinbo Shi, Yijie Peng, and Gongbo Zhang. Top-two thompson sampling for contextual top-mc selection problems, 2023.

L. J. Hong. Ranking and selection procedures for large-scale simulation optimization. *Journal of the Operational Research Society*, 69(1):1–13, 2018. URL `https://people.orie.cornell.edu/shane/pubs/RandSRuntime.pdf`.

Xiuyuan Lu and Benjamin Van Roy. Information-theoretic confidence bounds for reinforcement learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/411ae1bf081d1674ca6091f8c59a266f-Paper.pdf`.

Daniel Paulin. Concentration inequalities for markov chains by marton couplings and spectral methods. 2015.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ode method for asymptotic statistics in stochastic approximation and reinforcement learning, 2024.

Csaba Szepesvári. The asymptotic convergence-rate of q-learning. In M. Jordan, M. Kearns, and S. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997. URL `https://proceedings.neurips.cc/paper_files/paper/1997/file/cd0dce8fca267bf1fb86cf43e18d5598-Paper.pdf`.

Xiang Li, Wenhao Yang, Jiadong Liang, Zhihua Zhang, and Michael I. Jordan. A statistical analysis of polyak-ruppert averaged q-learning. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2207–2261. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/li23b.html`.

Chuhan Xie and Zhihua Zhang. A statistical online inference approach in averaged stochastic approximation. *Advances in Neural Information Processing Systems*, 35:8998–9009, 2022.

Yixuan Zhang and Qiaomin Xie. Constant stepsize q-learning: Distributional convergence, bias and extrapolation, 2024.

Peter Glynn and Sandeep Juneja. A large deviations perspective on ordinal optimization. In *In Proceedings of the 2004 Winter Simulation Conference (WSC)*, volume 1. IEEE, 2004.

Yi Zhu, Jing Dong, and Henry Lam. Uncertainty quantification and exploration for reinforcement learning. *Operations Research*, 0(0):null, 2023. doi: 10.1287/opre.2023.2436.

Ian Osband, Daniel Russo, and Benjamin Van Roy. Efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, volume 26, pages 3054–3062, 2013.

Alexander Hans, Steffen Udluft, and Daniel Schneegass. The value of a good explanation: on the importance of the environment model in rl. In *Proceedings of the 23rd International Conference on Artificial Intelligence*, pages 15–22, 2016.

William Feller. *An introduction to probability theory and its applications, Volume 2*, volume 2. John Wiley & Sons, 1991.

A Experiment Details

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 $S = 30$ 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 $r(1) = 4$ 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.

Our model encodes each instance–input pair through parallel pathways and a lightweight attention mechanism. The instance encoder has two branches: a convolutional branch processes the instance-transition matrix (reshaped to 60×60 with a single channel) through three convolutional kernels with channel sizes 64, 128, and 256, followed by adaptive average pooling to obtain a 256-dimensional feature, which is further projected to 256 dimensions with a linear layer; and a tabular branch maps 122 auxiliary features through two fully connected layers (122-256-128), producing a 128-dimensional representation. The two outputs are concatenated into a 384-dimensional vector and fused through a two-layer MLP (384-256-128) to form the final 128-dimensional instance representation. Each candidate input (60-dimensional) is independently mapped into the same latent space using a two-layer MLP (60-256-128). A multi-head paired attention module (4 heads, 32 dimensions per head) then conditions each input on its corresponding instance embedding using gated dot-product scores, projecting the output back to 128 dimensions, which is combined with the input embedding via residual addition and normalization. The resulting embedding is refined by a residual feed-forward block (128-256-128) and finally passed through a two-layer prediction head (128-128-1) to produce the scalar output.

B Algorithms

Algorithm 1 Q-Learning Optimal Computing Budget Allocation (Q-OCBA) Policy

```
Input: Review periods $B$, cumulative number of data collected after stage $b$, $t_b$, $b = 1, \ldots, B$, with $t_B = N$, i.e., $N$ is the total sampling budget (and we define $t_0 = 0$), estimation parameters $\rho, k, \alpha$ satisfying assumptions in Theorem 1, auxiliary parameter $\xi > 0$ (a penalty constant), $0 < \varepsilon \ll 1$, and initial exploration policy $\pi_1$.
Initialize: iteration numbering $b = 1$; randomized initial Q-table $Q_0$.
while $b \leq B$ do
    Simulate data from exploration policy $\pi_b$ for $t_b - t_{b-1}$ steps, update $Q_{t_{b-1}+1}, \ldots, Q_{t_b}$ in an online manner, and maintain estimations $\hat{P}_{t_b}$ and $\hat{\Sigma}_{R,t_b}$;
    Use plug-in estimation based on $Q_{t_b}$ and $\hat{P}_{t_b}$ to obtain $\hat{\pi}^*_{t_b}$, $\hat{V}^*_{t_b}$ and $\hat{\Sigma}_{T,t_b}$.
    Solve optimization problem (1) for $\Lambda_b$, with $W^*$ replaced by $W_{t_b}$, $P$ replaced by $\hat{P}_{t_b}$, $\pi^*$ replaced by $\hat{\pi}^*_{t_b}$, $\Sigma_R$ replaced by $\hat{\Sigma}_{R,t_b}$, and $\Sigma_T$ replaced by $\hat{\Sigma}_{T,t_b}$.
    Set $\pi_b(s|a) = \lambda_b(s,a) / \sum_{a' \in \mathcal{A}} \lambda_b(s,a')$ and $b = b + 1$.
end while
```

C Assumptions

For any $M > 0$, we define $T_k := \sum_{t=0}^{k-1} p_t$, $\tau_0 = 0$ and $\tau_{n+1} := \inf\{t > \tau_n : T_t \geq T_{\tau_n} + M\}$, inductively. We assume the following assumptions.

Assumption 1. *For any state $s \in \mathcal{S}$, $\arg\max_{a\in\mathcal{A}} Q^*(s,a)$ is unique.*

Assumption 2. *The transition probability matrix P^π is aperiodic and irreducible.*

Assumption 3. *Assume there exists a constant C_3, which is subject to change of M, such that $\forall(s,a)$,*

1. *$W_t > 0$ is a predictable process, i.e., $w_t \in \mathcal{F}_{t-1}$, converging to W^*, a.s.; and*
2. *for $t \leq l$ that satisfies $T_l - T_t \leq M$, $|w_l^{-1}(s,a) - w_t^{-1}(s,a)| \leq C_3 |w_t^{-1}(s,a)|$; and*
3. *$\lim_{n\to\infty} \sup_{\tau_n \leq t \leq \tau_{n+1}} \|w_t^{-1}(s,a) - w^{*-1}(s,a)\|_{r_n} = 0$ as $n \to \infty$, if $r_n = O(n^2)$; moreover, $\sup_{\tau_n \leq t \leq \tau_{n+1}} p_t^{-1/2} \|w_t^{-1}(s,a) - w^{*-1}(s,a)\|_{r_n} < C_3$, if $r_n = O(1)$.*

D Lemmas, Propositions, and Proofs

Lemma 1 (Hurwitz). *In either of the following cases,*

1. *$1/2 < \rho < 1$, $k > 0$ arbitrary, and $\alpha = 0$; or*
2. *$\rho = 1$, $0 < 1/k < 2(1-\gamma)\cdot(1 \wedge \min_{s,a} \lambda(s,a^*(s))/w^*(s,a^*(s)))$, and $\alpha = 1/k$,*

the drift coefficient A in Theorem 1 is Hurwitz. Therefore, the Lyapunov equation (1) is well-posed. Furthermore, if Σ_ζ is symmetric and (semi-)positive definite, the unique solution to the Lyapunov equation (1) is symmetric and (semi-)positive definite as well.

Proof. Recall all entries of $\bar{A} = (W^*)^{-1}\Lambda(\gamma P^* - I)$ are non-positive since P^* is a probability transition matrix. It follows from the Gershgorin circle theorem that for each eigenvalue z of $\bar{A}$, either $|z - 1| \leq \gamma$, or for some $s \in \mathcal{S}$,

$$\left|z - \lambda(s,a^*(s))(\gamma P_{s,a^*(s)}(s) - 1)/w^*(s,a^*(s))\right| \leq \lambda(s,a^*(s))\gamma(1 - P_{s,a^*(s)}(s))/w^*(s,a^*(s)).$$

In the former case, $\mathrm{Re}(z) \leq -(1-\gamma) < 0$, and in the latter case,

$$\mathrm{Re}(z) \leq -(1-\gamma)\lambda(s,a^*(s))/w^*(s,a^*(s)) < 0.$$

If condition 1. holds, $A = \bar{A}$ is Hurwitz. If condition 2. holds, each eigenvalue z' of $\frac{1}{2}\alpha I$ coincides with $\frac{1}{2k} + z$ for some eigenvalue z of $\bar{A}$. Therefore, in either case in the last paragraph, $\mathrm{Re}(z') = \frac{1}{2k} + \mathrm{Re}(z) < 0$. Hence A is Hurwitz.

Now, suppose Σ_ζ is symmetric. Then, since A is Hurwitz, the unique solution admits an analytical form $\Sigma_Q = \int_0^\infty \exp\{At\}\Sigma_\zeta \exp\{A^\top t\}dt$. Hence Σ_Q is symmetric. If Σ_ζ is semi-positive definite, then $x^\top \Sigma_Q x = \int_0^\infty x^\top \exp\{At\}\Sigma_\zeta \exp\{A^\top t\}x dt \geq 0$ for any $x \in \mathbb{R}^D$. If Σ_ζ is further positive definite, we claim Σ_Q is positive definite as well. Suppose otherwise that there exists $x \neq 0$ such that $x^\top \Sigma_Q x = 0$, we must have $x^\top \exp\{At\}\Sigma_\zeta \exp\{A^\top t\}x = 0$ for $t \geq 0$ almost everywhere. Since Σ_ζ is positive definite, $\exp\{A^\top t\}x = 0$, which implies that $\exp\{A^\top t\}$ does not have full rank, a contradiction to the fact that A is Hurwitz. □

Lemma 2. *The following statements regarding τ_n hold:*

1. *for $\frac{1}{2} < \rho < 1$, we have $\tau_n + k = \Theta(n^{1/(1-\rho)})$ as $n \to \infty$ and $(\tau_{n+1} - \tau_n)/\tau_n = O(n^{-1})$;*
2. *for $\rho = 1$, we have $k\exp\{nMk^{-\rho} - 1\} \leq \tau_n + k \leq k\exp\{n(M+1)k^{-\rho}\}$, and $(\tau_{n+1} - \tau_n)/\tau_n < (1+k)(\exp\{M+1\} - 1)$.*

Proof. By definition of τ_n, we have $nM \leq T_{\tau_n} \leq n(M+1)$. Moreover, it follows from the integral inequality that

$$\int_k^{k+t} x^{-\rho}dx \leq k^{-\rho}T_t = \sum_{j=0}^{t-1}(j+k)^{-\rho} \leq 1 + \int_k^{k+t} x^{-\rho}dx.$$

Moreover,

$$k^{-\rho}(M+1) \geq \sum_{j=\tau_n}^{\tau_{n+1}-1} (j+k)^{-\rho} \geq \int_{\tau_n+k}^{\tau_{n+1}+k} x^{-\rho} dx.$$

For $\rho < 1$, it follows from the first inequality that $(k^{1-\rho}+n(1-\rho)Mk^{-\rho}-(1-\rho))^{1/(1-\rho)} \leq \tau_n+k \leq (k^{1-\rho} + n(1-\rho)(M+1)k^{-\rho})^{1/(1-\rho)}$. And, it follows from the second inequality and the concavity of the mapping $x \mapsto x^{1-\rho}$ that $\int_{\tau_n+k}^{\tau_{n+1}+k} x^{-\rho}dx = (1-\rho)^{-1}((\tau_{n+1}+k)^{1-\rho} - (\tau_n+k)^{1-\rho}) \geq (\tau_n+k)^{-\rho}(\tau_{n+1}-\tau_n)$. Since $(\tau_n+k)^{1-\rho} \geq k^{1-\rho} + n(1-\rho)Mk^{-\rho} - (1-\rho)$, we see that $(\tau_{n+1}-\tau_n)/\tau_n = O(n^{-1})$ as $n \to \infty$.

For $\rho = 1$, we have $\int_k^{k+t} x^{-1}dx = \ln(1+t/k)$. The first inequality leads to $k\exp\{nMk^{-\rho}-1\} \leq \tau_n + k \leq k\exp\{n(M+1)k^{-\rho}\}$. From the second inequality above, we have $\exp\{M+1\} - 1 \geq (\tau_{n+1}-\tau_n)/(\tau_n+k) \geq (\tau_{n+1}-\tau_n)/((1+k)\tau_n)$. □

Proposition 1. *Under Assumption 2, either (i) the constant steps sizes, i.e.,* $w_t(s,a) = w^*(s,a) = 1$, *or (ii) the inverse-visit-frequency step sizes, i.e.,* $w_t(s,a) = (N_{t-1}(s,a)+1)/(t+1)$, *satisfies the conditions in Assumption 3.*

Proof. For case (i), W_t is a constant and thus Assumption 3 holds trivially.

For case (ii), given Assumption 2, $w_t(s,a) = (N_{t-1}(s,a)+1)/(t+1)$ is $\mathcal{F}_{t-1}$-measurable and converges to $\lambda(s,a)$. Therefore, Assumption 3.1 holds. To see the validity of Assumption 3.2, a straightforward calculation yields $w_l^{-1}(s,a) - w_t^{-1}(s,a) = \frac{l+1}{N_{l-1}(s,a)+1} - \frac{t+1}{N_{t-1}(s,a)+1} \leq \frac{l-t}{N_{t-1}(s,a)+1} = \frac{l-t}{t+1} w_t^{-1}(s,a)$ and $w_l^{-1}(s,a) - w_t^{-1}(s,a) \geq -\frac{t+1}{N_{t-1}(s,a)+1}$. Choosing $C_3 = \max\{1, \sup_{t\leq l,\ T_l - T_t \leq M} \frac{l-t}{t+1}\}$ suffices, if the second term is finite. In fact, using a similar argument in the proof of Lemma 2, we have $k^{-\rho}M \geq k^{-\rho}(T_l - T_t) = \sum_{j=t+1}^{l} (j+k)^{-\rho} \geq \int_{t+k+1}^{l+k+1} x^{-\rho}dx \geq \frac{l-t}{t+1}\frac{t+1}{(t+k+1)^\rho}$. It turns out that $\limsup_{t\to\infty} \sup_{T_l-T_t\leq M} \frac{l-t}{t+1} \leq k^{-\rho}M$, which justifies Assumption 3.2.

Given Assumption 2, Paulin [2015] implies the Hoeffding's inequality

$$\mathbb{P}\left(\left|\frac{N_{t-1}(s,a)}{t} - \lambda(s,a)\right| \geq x\right) \leq \exp\{-C_1 t x^2\}, \tag{2}$$

where C_1 is a constant depending on the mixing time of the underlying Markov chain. Fix any s, a, and write $\delta_t := w_t^{-1}(s,a) - {w^*}^{-1}(s,a)$. And we define the following analogy to δ_t:

$$\tilde{\delta}_t := w_t^{-1}(s,a) - \frac{t+1}{tw^*(s,a)+1} = \tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s,a)+1}\} + \tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}\}.$$

If $|\tilde{\delta}_t| > \frac{t+1}{tw^*(s,a)+1}$, then $w_t^{-1}(s,a) \geq \frac{2(t+1)}{tw^*(s,a)+1} > 0$ and $\tilde{\delta}_t \leq \frac{t}{N_{t-1}(s,a)} - {w^*}^{-1}(s,a)$. Therefore, it follows from (2), for $x \geq \frac{t+1}{tw^*(s,a)+1} \geq 1$,

$$\mathbb{P}\left(w_t^{-1}(s,a) - \frac{t+1}{tw^*(s,a)+1} \geq x\right) \leq \mathbb{P}\left(\frac{t}{N_{t-1}(s,a)} - {w^*}^{-1}(s,a) \geq x\right) \leq \exp\{-C_1' t\},$$

where $C_1' = C_1 \frac{{w^*}^2(s,a)}{({w^*}^{-1}(s,a)+1)^2}$. It follows by definition that $w_t(s,a) \geq \frac{1}{1+t}$. Hence,

$$\begin{aligned}\left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n}^{r_n} &= \int_{\frac{t+1}{tw^*(s,a)+1}}^{t+1} r_n x^{r_n-1} \mathbb{P}\left(w_t^{-1}(s,a) - \frac{t+1}{tw^*(s,a) \geq 1} + x\right) dx \\ &\leq \exp\{-C_1' t\} \times \int_0^{t+1} r_n x^{r_n-1} dx = \exp\{-C_1' t\} \cdot (t+1)^{r_n}.\end{aligned}$$

Therefore, for n sufficiently large, if $1/2 < \rho < 1$,

$$\sup_{\tau_n \leq t \leq \tau_{n+1}} \left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n} \leq \exp\{-C_1' \frac{\tau_n}{r_n}\}(\tau_{n+1}+1) = O\left(\exp\{-C_1' \frac{\tau_n}{r_n}\}\tau_n\right).$$

The second equality above holds because $\sup_{n\geq 1} \frac{\tau_{n+1}}{\tau_n} \leq \infty$ according to Lemma 2.

If $|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}$, then $0 \leq w_t^{-1}(s,a) \leq 2{w^*}^{-1}(s,a)$. Consequently,

$$|\tilde{\delta}_t| = \left|\frac{t+1}{N_{t-1}(s,a)+1} - \frac{t+1}{tw^*(s,a)+1}\right| \leq \left|\frac{2t{w^*}^{-1}(s,a)}{tw^*(s,a)+1}\right| \cdot \left|\frac{N_{t-1}(s,a)}{t} - w^*(s,a)\right|,$$

and by (2) again, it turns out that for $x > 0$,

$$\begin{aligned}&\mathbb{P}\left(\left|\frac{t+1}{N_{t-1}(s,a)+1} - \frac{t+1}{tw^*(s,a)+1}\right| \cdot \mathbf{1}\{|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}\} \geq x\right)\\ \leq&\mathbb{P}\left(\left|\frac{N_{t-1}(s,a)}{t} - w^*(s,a)\right| \geq \left|\frac{tw^*(s,a)+1}{2t{w^*}^{-1}(s,a)}\right| x\right) \leq \exp\{-C_1'' t x^2\},\end{aligned}$$

where $C_1'' = C_1 \frac{{w^*}^4(s,a)}{4}$. Hence,

$$\begin{aligned}\left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n}^{r_n} \leq& \int_{-{w^*}^{-1}(s,a)}^{{w^*}^{-1}(s,a)} r_n |x|^{r_n-1} \exp\{-C_1'' t x^2\} dx\\ \leq& r_n (2C_1'' t)^{-r_n/2} \sqrt{2\pi} \mathbb{E}|Z|^{r_n-1},\end{aligned}$$

where $Z \sim N(0,1)$. According to Feller [1991], we have $\mathbb{E}|Z|^{r_n-1} = 2^{\frac{r_n-1}{2}} \Gamma(\frac{r_n}{2})/\sqrt{\pi}$, where Γ is the gamma function. Then $\left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n} \leq O(r_n^{1/r_n} (\frac{t}{2})^{-\frac{1}{2}} \Gamma(\frac{r_n}{2})^{\frac{1}{r_n}})$. It follows from the Stirling's formula that this moment is smaller than or equal to $O(\sqrt{r_n/t}) \leq O(\sqrt{r_n/\tau_n})$.

Finally, we see that $0 \leq \delta_t - \tilde{\delta}_t \leq \frac{1-w^*(s,a)}{{w^*}^2(s,a)} \cdot \frac{1}{t} = O(\frac{1}{t})$. By the triangle inequality,

$$\begin{aligned}\sup_{\tau_n \leq t \leq \tau_{n+1}} \|\delta_t\|_{r_n} \leq& \sup_{\tau_n \leq t \leq \tau_{n+1}} \|\delta_t - \tilde{\delta}_t\|_{r_n} + \left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n}\\ &+ \left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| \leq \frac{t+1}{tw^*(s,a)+1}\}\right\|_{r_n}\\ =&O\left(\frac{1}{\tau_n}\right) + O\left(\exp\{-C_1' \frac{\tau_n}{r_n}\}\tau_n\right) + O\left(\sqrt{\frac{r_n}{\tau_n}}\right).\end{aligned}$$

For $r_n = O(n^2)$, Lemma 2 implies that $1/\tau_n = o(1)$. If $1/2 < \rho < 1$, $\exp\{-C_1' \frac{\tau_n}{r_n}\}\tau_n = \exp\{-\Omega(n^{\frac{1}{1-\rho}-2})\} O(n^{\frac{1}{1-\rho}}) = o(1)$, and $\sqrt{\frac{r_n}{\tau_n}} = O(n^{1-\frac{1}{2(1-\rho)}}) = o(1)$. If $\rho = 1$, it follows from the numerical inequality $\exp\{-x\} = o(1/x^p)$, $\forall\ p > 0$, $x > 0$, that $\exp\{-C_1' \frac{\tau_n}{r_n}\}\tau_n = \exp\{-\Omega(\exp\{nM/k\}/n^2)\} \cdot O(\exp\{n(M+1)/k\}) = o(1)$. To partially conclude, the first statement in Assumption 3.3 holds.

In addition, for $t \leq \tau_{n+1}$, $p_t^{-1/2} \leq p_{\tau_{n+1}}^{-1/2} = O(p_{\tau_n}^{-1/2}) = O(\tau_n^{\rho/2})$, where the last equality follows from Lemma 2. We have

$$\sup_{\tau_n \leq t \leq \tau_{n+1}} p_{t-1}^{-1/2} \|\delta_t\|_{r_n} \leq O\left(\tau_n^{\frac{\rho}{2}-1}\right) + O\left(\tau_n^{\frac{\rho}{2}+1} \exp\{-C_1' \frac{\tau_n}{r_n}\}\right) + O\left(\tau_n^{\frac{\rho-1}{2}} \sqrt{r_n}\right).$$

Now suppose $r_n = O(1)$. It is apparent that $O(\tau_n^{\rho/2-1}) = O(1)$. Since $p_{\tau_n} = \tau_n^{-\rho}$, the second term is still dominated by the exponential term and equal to $o(1)$. For the last term, $\tau_n^{\frac{\rho-1}{2}} \sqrt{r_n} = O(\tau_n^{\frac{\rho-1}{2}})$, which is $O(n^{-\frac{1}{2}})$ if $\rho < 1$ and $O(1)$ if $\rho = 1$. This completes the proof. □