

A Experiment Details

The river swim problem is a widely used benchmark in RL. The environment consists of a finite, discrete set of states $[S] = \{1, 2, \dots, S\}$, 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 $\{2, \dots, S - 1\}$ 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

- 1: **Input:** Review periods B , cumulative number of data collected after stage b , t_b , $b = 1, \dots, B$, with $t_B = N$, i.e., N is the total sampling budget (and we define $t_0 = 0$), estimation parameters ρ, k, α satisfying assumptions in Theorem 1, auxiliary parameter $\xi > 0$ (a penalty constant), $0 < \varepsilon \ll 1$, and initial exploration policy π_1 .
 - 2: **Initialize:** iteration numbering $b = 1$; randomized initial Q-table Q_0 .
 - 3: **while** $b \leq B$ **do**
 - 4: Simulate data from exploration policy π_b for $t_b - t_{b-1}$ steps, update $Q_{t_{b-1}+1}, \dots, Q_{t_b}$ in an online manner, and maintain estimations \hat{P}_{t_b} and $\hat{\Sigma}_{R,t_b}$;
 - 5: 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 $\hat{\Sigma}_{T,t_b}$.
 - 6: Solve optimization problem (1) for Λ_b , with W^* replaced by W_{t_b} , P replaced by \hat{P}_{t_b} , π^* replaced by $\hat{\pi}_{t_b}^*$, Σ_R replaced by $\hat{\Sigma}_{R,t_b}$, and Σ_T replaced by $\hat{\Sigma}_{T,t_b}$.
 - 7: Set $\pi_b(s|a) = \lambda_b(s, a) / \sum_{a' \in \mathcal{A}} \lambda_b(s, a')$ and $b = b + 1$.
 - 8: **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. $\lim_{n \rightarrow \infty} \sup_{\tau_n \leq t \leq \tau_{n+1}} \|w_t^{-1}(s, a) - w^{*-1}(s, a)\|_{r_n} = 0$ as $n \rightarrow \infty$, 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 $s \in \mathcal{S}$,

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

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

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

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, $\text{Re}(z') = \frac{1}{2k} + \text{Re}(z) < 0$. 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. \square

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 $n \rightarrow \infty$ 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

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

275 Moreover,

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

276 For $\rho < 1$, it follows from the first inequality that $(k^{1-\rho} + n(1-\rho)Mk^{-\rho} - (1-\rho))^{1/(1-\rho)} \leq \tau_n + k \leq$
 277 $(k^{1-\rho} + n(1-\rho)(M+1)k^{-\rho})^{1/(1-\rho)}$. And, it follows from the second inequality and the concavity
 278 of the mapping $x \mapsto x^{1-\rho}$ that $\int_{\tau_n+k}^{\tau_{n+1}+k} x^{-\rho} dx = (1-\rho)^{-1}((\tau_{n+1}+k)^{1-\rho} - (\tau_n+k)^{1-\rho}) \geq$
 279 $(\tau_n+k)^{-\rho}(\tau_{n+1}-\tau_n)$. Since $(\tau_n+k)^{1-\rho} \geq k^{1-\rho} + n(1-\rho)Mk^{-\rho} - (1-\rho)$, we see that
 280 $(\tau_{n+1}-\tau_n)/\tau_n = O(n^{-1})$ as $n \rightarrow \infty$.

281 For $\rho = 1$, we have $\int_k^{k+t} x^{-1} dx = \ln(1+t/k)$. The first inequality leads to $k \exp\{nMk^{-\rho} - 1\} \leq$
 282 $\tau_n + k \leq k \exp\{n(M+1)k^{-\rho}\}$. From the second inequality above, we have $\exp\{M+1\} - 1 \geq$
 283 $(\tau_{n+1}-\tau_n)/(\tau_n+k) \geq (\tau_{n+1}-\tau_n)/((1+k)\tau_n)$. \square

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

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

288 For case (ii), given Assumption 2, $w_t(s, a) = (N_{t-1}(s, a) + 1)/(t+1)$ is \mathcal{F}_{t-1} -measurable
 289 and converges to $\lambda(s, a)$. Therefore, Assumption 3.1 holds. To see the validity of Assumption
 290 3.2, a straightforward calculation yields $w_t^{-1}(s, a) - w_t^{-1}(s, a) = \frac{l+1}{N_{l-1}(s, a)+1} - \frac{t+1}{N_{t-1}(s, a)+1} \leq$
 291 $\frac{l-t}{N_{t-1}(s, a)+1} = \frac{l-t}{t+1} w_t^{-1}(s, a)$ and $w_l^{-1}(s, a) - w_t^{-1}(s, a) \geq -\frac{t+1}{N_{t-1}(s, a)+1}$. Choosing $C_3 =$
 292 $\max\{1, \sup_{t \leq l, T_l - T_t \leq M} \frac{l-t}{t+1}\}$ suffices, if the second term is finite. In fact, using a similar argument
 293 in the proof of Lemma 2, we have $k^{-\rho}M \geq k^{-\rho}(T_l - T_t) = \sum_{j=t+1}^l (j+k)^{-\rho} \geq \int_{t+k+1}^{l+k+1} x^{-\rho} dx \geq$
 294 $\frac{l-t}{t+1} \frac{t+1}{(t+k+1)^\rho}$. It turns out that $\limsup_{t \rightarrow \infty} \sup_{T_l - T_t \leq M} \frac{l-t}{t+1} \leq k^{-\rho}M$, which justifies Assumption
 295 3.2.

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

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

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 $x \geq \frac{t+1}{tw^*(s, a)+1} \geq 1$,

$$\mathbb{P}\left(w_t^{-1}(s, a) - \frac{t+1}{tw^*(s, a)+1} \geq x\right) \leq \mathbb{P}\left(\frac{t}{N_{t-1}(s, a)} - w^{*-1}(s, a) \geq x\right) \leq \exp\{-C'_1 t\},$$

297 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,

$$\begin{aligned} \left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s, a)+1}\}\right\|_{r_n}^{r_n} &= \int_{\frac{t+1}{tw^*(s, a)+1}}^{t+1} r_n x^{r_n-1} \mathbb{P}\left(w_t^{-1}(s, a) - \frac{t+1}{tw^*(s, a)+1} \geq x\right) dx \\ &\leq \exp\{-C'_1 t\} \times \int_0^{t+1} r_n x^{r_n-1} dx = \exp\{-C'_1 t\} \cdot (t+1)^{r_n}. \end{aligned}$$

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

$$\sup_{\tau_n \leq t \leq \tau_{n+1}} \left\|\tilde{\delta}_t \mathbf{1}\{|\tilde{\delta}_t| > \frac{t+1}{tw^*(s, a)+1}\}\right\|_{r_n} \leq \exp\{-C'_1 \frac{\tau_n}{r_n}\} (\tau_{n+1} + 1) = O\left(\exp\{-C'_1 \frac{\tau_n}{r_n}\} \tau_n\right).$$

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

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

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

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

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

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

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

302 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
303 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
304 from the Stirling's formula that this moment is smaller than or equal to $O(\sqrt{r_n/t}) \leq O(\sqrt{r_n/\tau_n})$.

305 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,

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

306 For $r_n = O(n^2)$, Lemma 2 implies that $1/\tau_n = o(1)$. If $1/2 < \rho < 1$, $\exp\{-C_1' \frac{\tau_n}{r_n}\} \tau_n =$
307 $\exp\{-\Omega(n^{\frac{1}{1-\rho}-2})\} O(n^{\frac{1}{1-\rho}}) = o(1)$, and $\sqrt{\frac{r_n}{\tau_n}} = O(n^{1-\frac{1}{2(1-\rho)}}) = o(1)$. If $\rho = 1$, it follows
308 from the numerical inequality $\exp\{-x\} = o(1/x^p)$, $\forall p > 0, x > 0$, that $\exp\{-C_1' \frac{\tau_n}{r_n}\} \tau_n =$
309 $\exp\{-\Omega(\exp\{nM/k\}/n^2)\} \cdot O(\exp\{n(M+1)/k\}) = o(1)$. To partially conclude, the first state-
310 ment in Assumption 3.3 holds.

311 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
312 from Lemma 2. We have

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

313 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
314 is still dominated by the exponential term and equal to $o(1)$. For the last term, $\tau_n^{\frac{\rho-1}{2}} \sqrt{r_n} = O(\tau_n^{\frac{\rho-1}{2}})$,
315 which is $O(n^{-\frac{1}{2}})$ if $\rho < 1$ and $O(1)$ if $\rho = 1$. This completes the proof. \square