Model-Free Nonstationary Reinforcement Learning

: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control

Jihyeon Hyeong, Hyeyoon Kang, Byoungho Son*,

2025-12-09

(*presenter)

CONTENTS

1. **Introduction**
    1. Problem
    2. Example
    3. Definition
2. **Algorithm & Analysis**
    1. RestartQ-UCB (Hoeffding)
    2. RestartQ-UCB (Freedman)
    3. Double RestartQ-UCB
3. **Conclusion**
    1. Experiment
    2. Concluding Remark

Problem

Stationary MDP

reward: $r(s, a)$

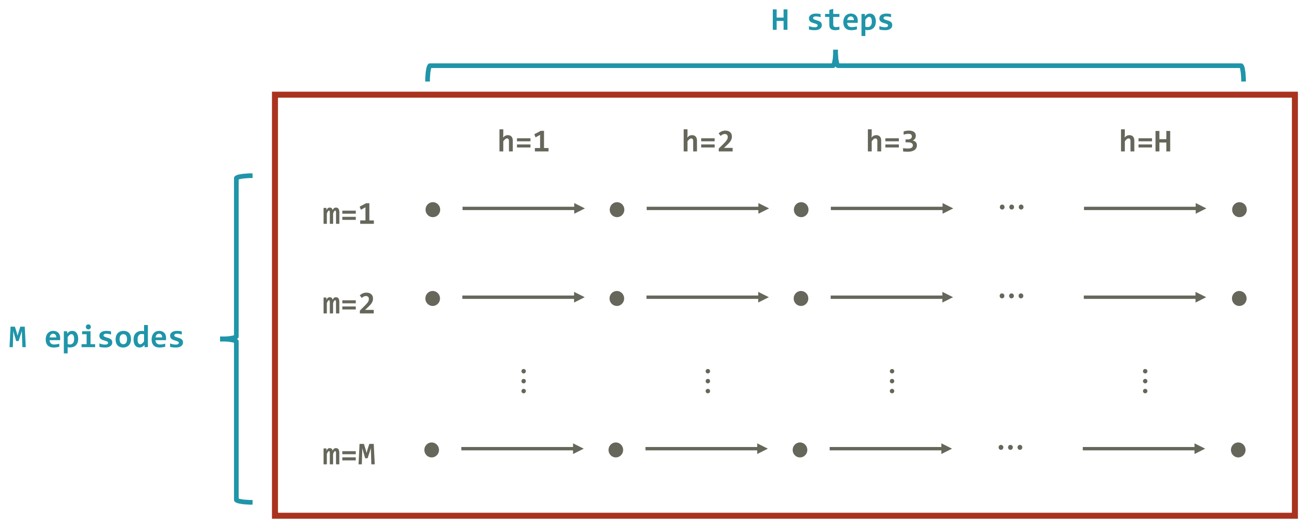
transition kernel: $P(s' \mid s, a)$

→

Nonstationary MDP

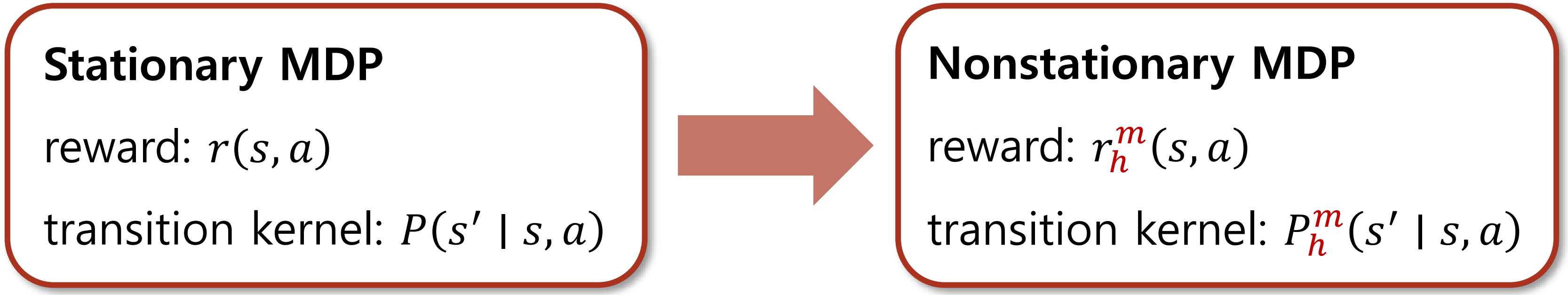
reward: $r_h^m(s, a)$

transition kernel: $P_h^m(s' \mid s, a)$



Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Problem



Stationary MDP

reward: $r(s, a)$

transition kernel: $P(s' \mid s, a)$

Nonstationary MDP

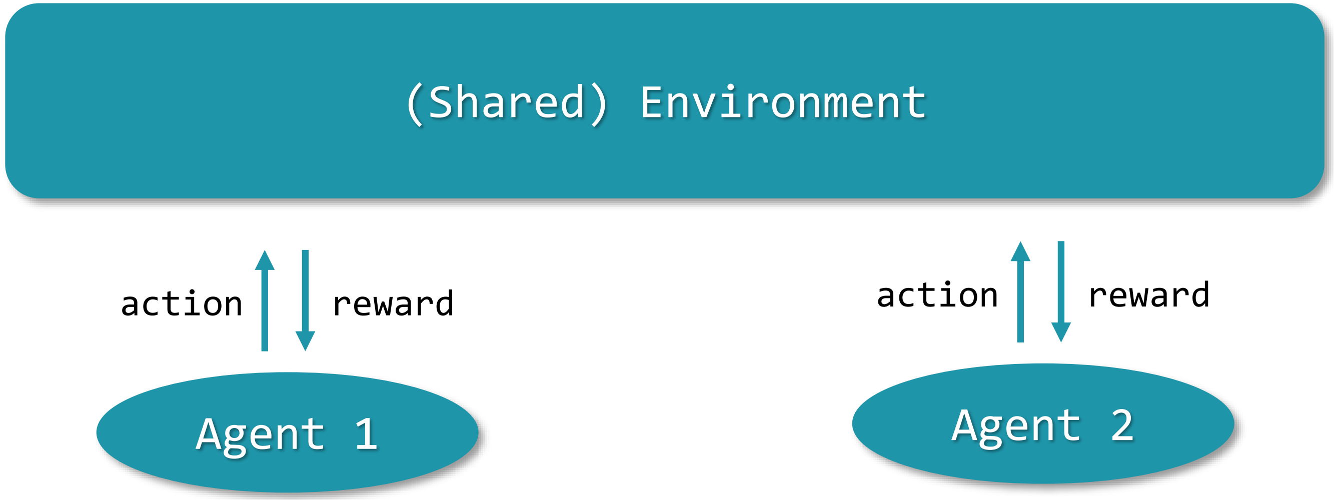
reward: $r_h^m(s, a)$

transition kernel: $P_h^m(s' \mid s, a)$

Problem: Can we design a **near-optimal** **model-free** learning algorithm over **Nonstationary** MDPs?

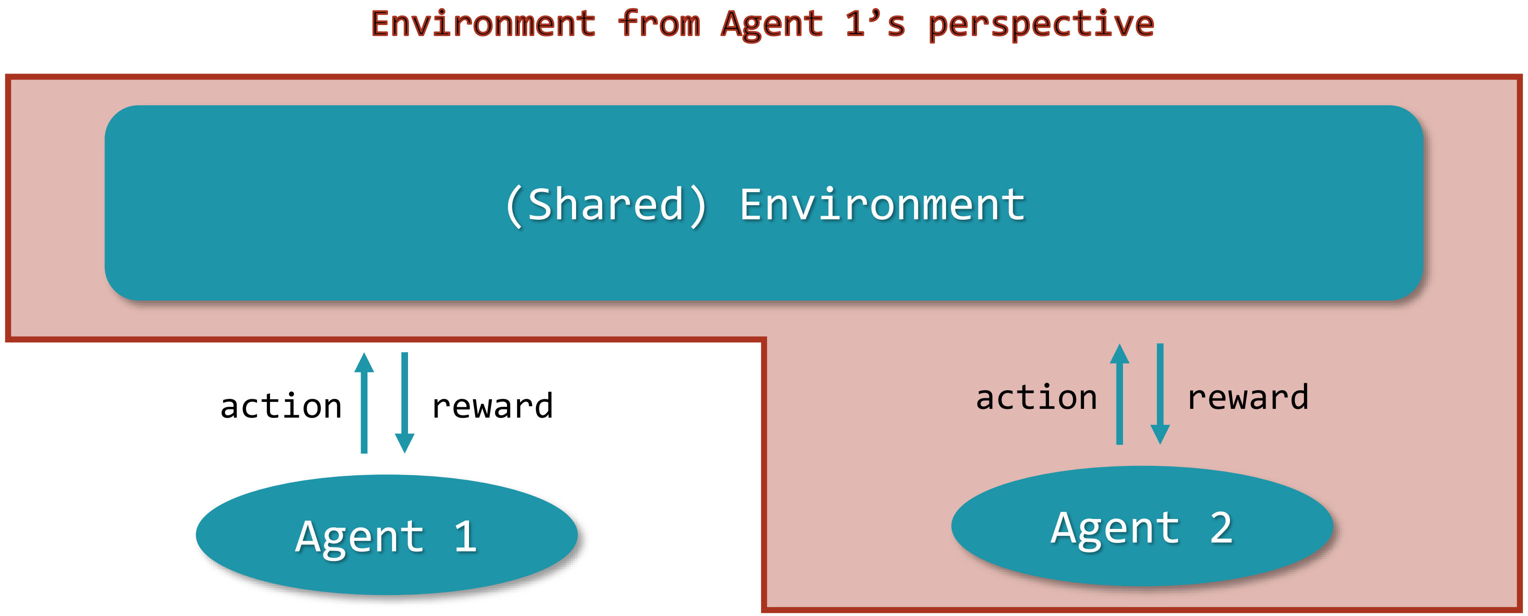
Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Example - 2-player Games



Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

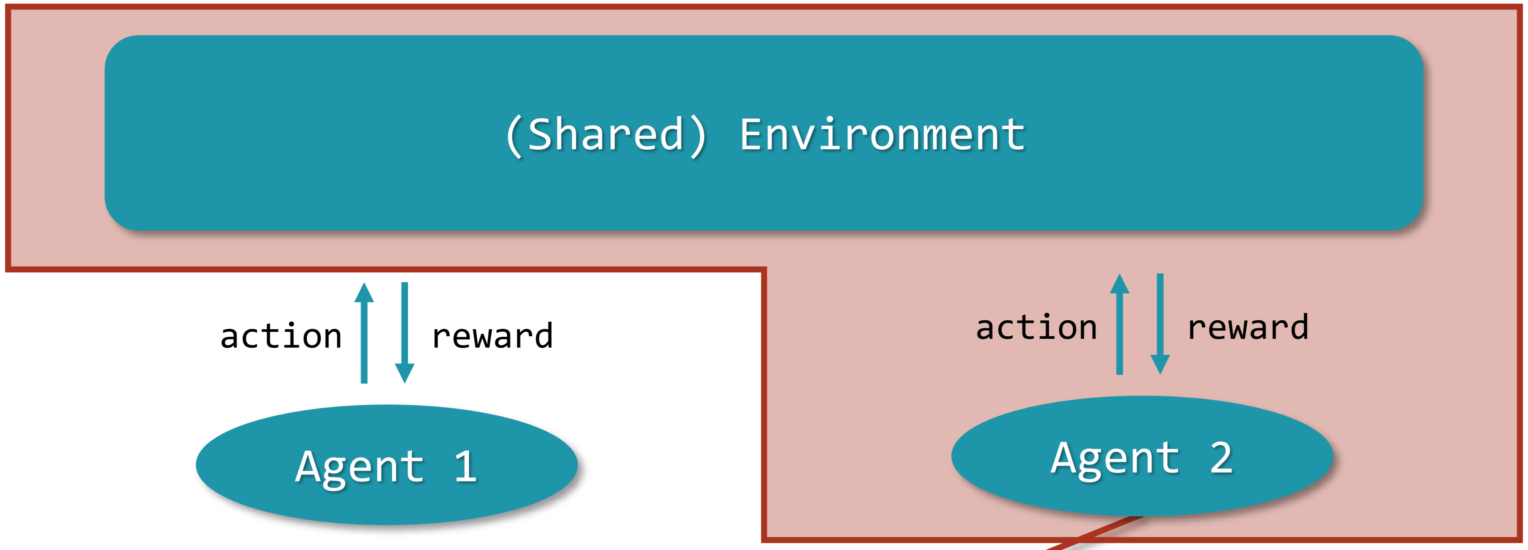
Example - 2-player Games



Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Example - 2-player Games

Environment from Agent 1’s perspective ⇒ NONSTATIONARY!

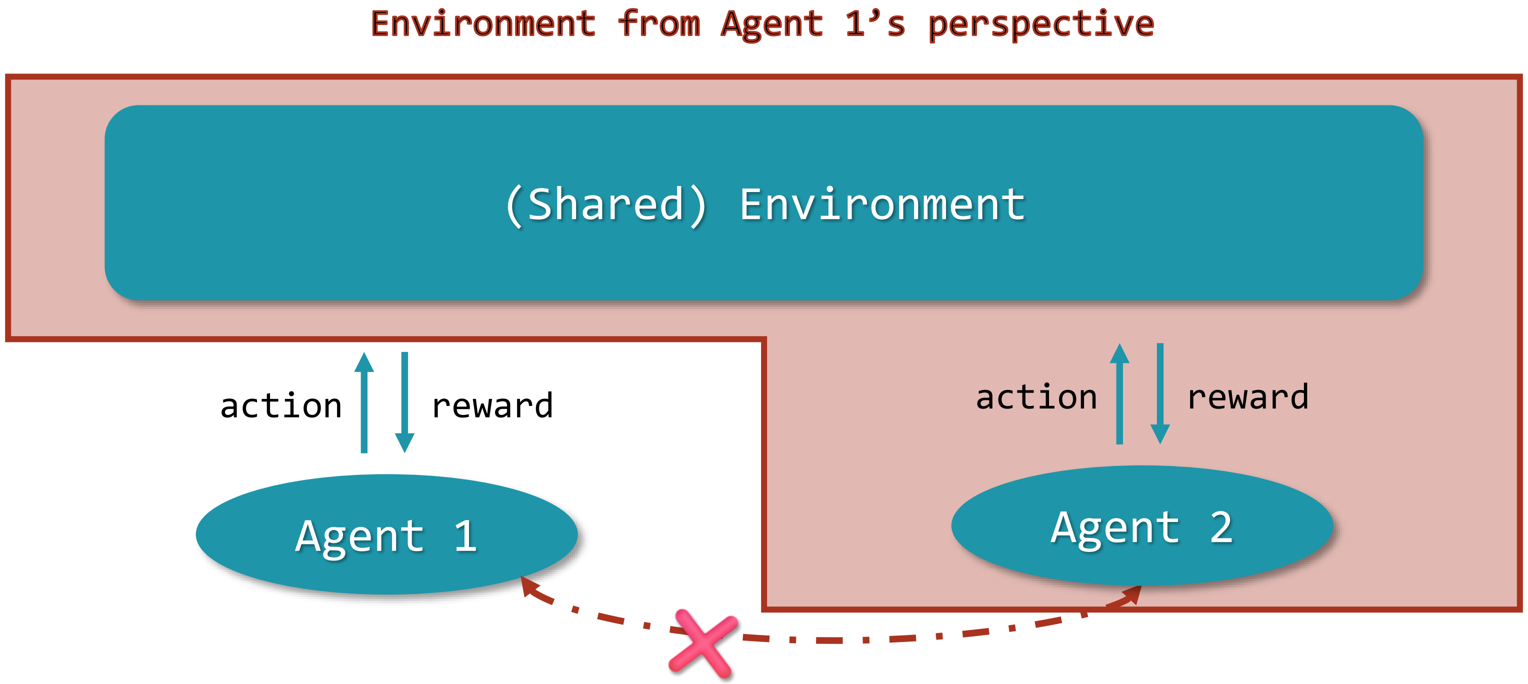


Consider:

- Shared environment is stationary
- Agent 2 can take arbitrary actions (uncontrollable by Agent 1)
- As the game proceeds, **Agent 2 learns and updates its policy** across episodes

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Example - 2-player Games



Decentralized: Agent 1 cannot observe the actions taken by Agent 2

⇒ Agent 1 has no access to the complete model of the environment

⇒ Agent 1 should learn the policy by simulation (**Model-free Learning**)

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Definition – Dynamic Regret, Variation Budget

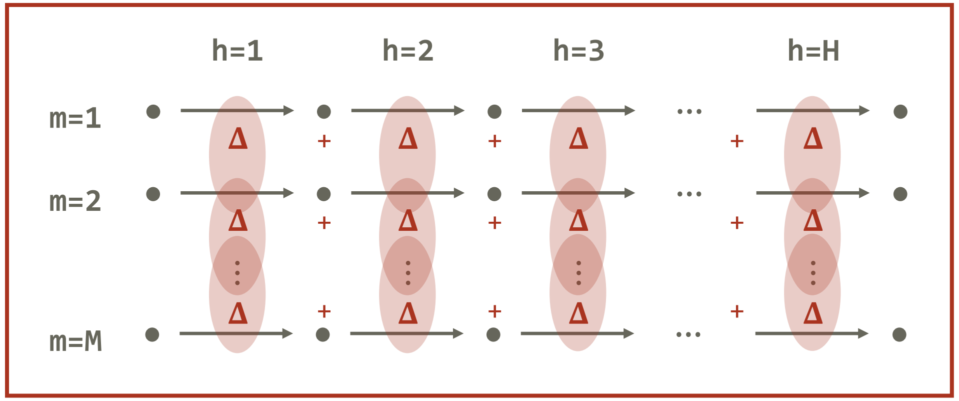
- **Dynamic Regret**: The measure of the algorithm’s performance
  - Static Regret: Compare to best **single** policy for all episodes
  - **Dynamic Regret**: Compare to best policy **for each episode**
    - $\mathcal{R}(\pi, M) \overset{\text{def}}{=} \sum_{m=1}^{M} (V_1^{m,\star}(s_1^m) - V_1^{m,\pi}(s_1^m))$
    - Measures **the optimality of policy** — appropriate for nonstationary environments

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Definition – Dynamic Regret, Variation Budget

- **Dynamic Regret**: The measure of the algorithm's performance
  - Static Regret: Compare to best **single** policy for all episodes
  - **Dynamic Regret**: Compare to best policy **for each episode**
    - $\mathcal{R}(\pi, M) \overset{\text{def}}{=} \sum_{m=1}^{M} (V_1^{m,\star}(s_1^m) - V_1^{m,\pi}(s_1^m))$
    - Measures **the optimality of policy** — appropriate for nonstationary environments

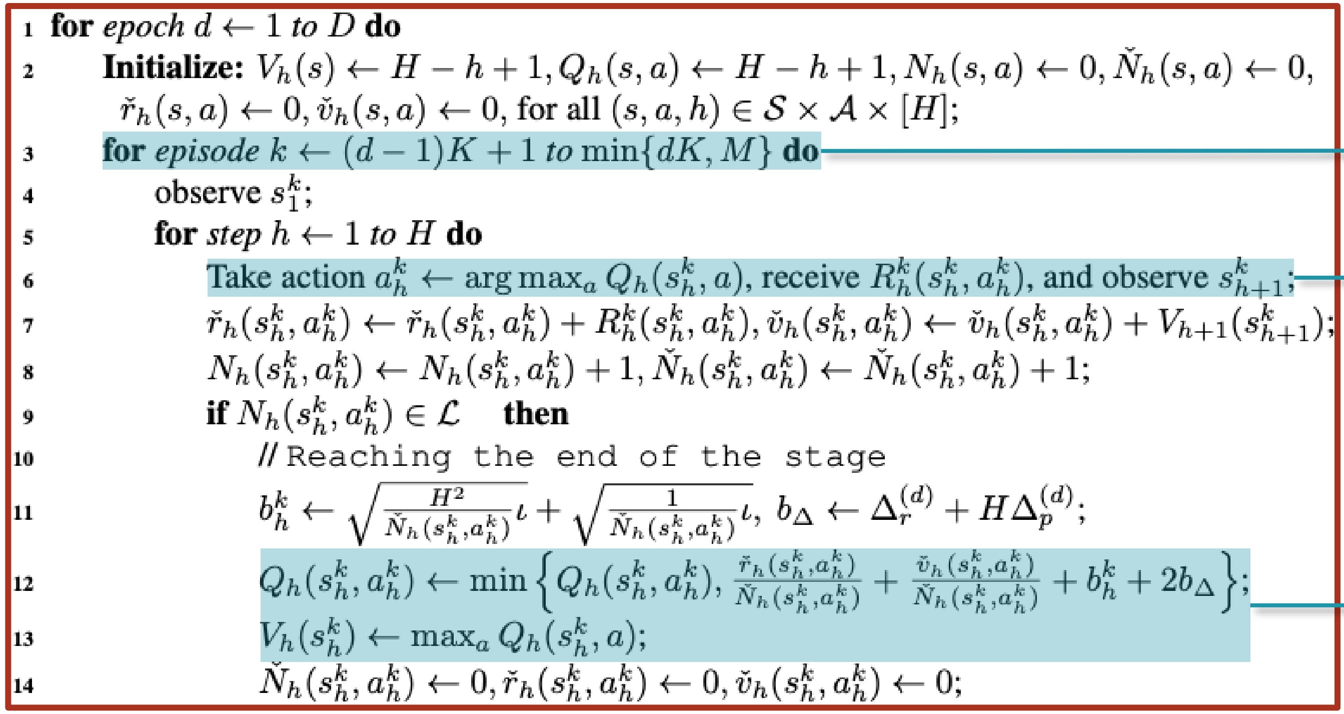
- **Variation Budget**: The measure of the model's non-stationarity
  - $\Delta = \Delta_r + \Delta_p$
    - $\Delta_r \overset{\text{def}}{=} \sum_{m=1}^{M-1} \sum_{h=1}^{H} \sup_{s,a} |r_h^m(s,a) - r_h^{m+1}(s,a)|$
    - $\Delta_p \overset{\text{def}}{=} \sum_{m=1}^{M-1} \sum_{h=1}^{H} \sup_{s,a} \|P_h^m(\cdot|s,a) - P_h^{m+1}(\cdot|s,a)\|_1$



Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- **RestartQ-UCB**
  - A familiar **Q-Learning** algorithm…



```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0,$
     $\check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1^k$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h^k \leftarrow \arg\max_a Q_h(s_h^k, a)$, receive $R_h^k(s_h^k, a_h^k)$, and observe $s_{h+1}^k$;
            $\check{r}_h(s_h^k, a_h^k) \leftarrow \check{r}_h(s_h^k, a_h^k) + R_h^k(s_h^k, a_h^k), \check{v}_h(s_h^k, a_h^k) \leftarrow \check{v}_h(s_h^k, a_h^k) + V_{h+1}(s_{h+1}^k)$;
            $N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1, \check{N}_h(s_h^k, a_h^k) \leftarrow \check{N}_h(s_h^k, a_h^k) + 1$;
            if $N_h(s_h^k, a_h^k) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h^k \leftarrow \sqrt{\frac{H^2}{\check{N}_h(s_h^k, a_h^k)}\iota} + \sqrt{\frac{1}{\check{N}_h(s_h^k, a_h^k)}\iota}, b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k), \frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)} + \frac{\check{v}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)} + b_h^k + 2b_\Delta\right\}$;
                $V_h(s_h^k) \leftarrow \max_a Q_h(s_h^k, a)$;
                $\check{N}_h(s_h^k, a_h^k) \leftarrow 0, \check{r}_h(s_h^k, a_h^k) \leftarrow 0, \check{v}_h(s_h^k, a_h^k) \leftarrow 0$;
```

for each episode…

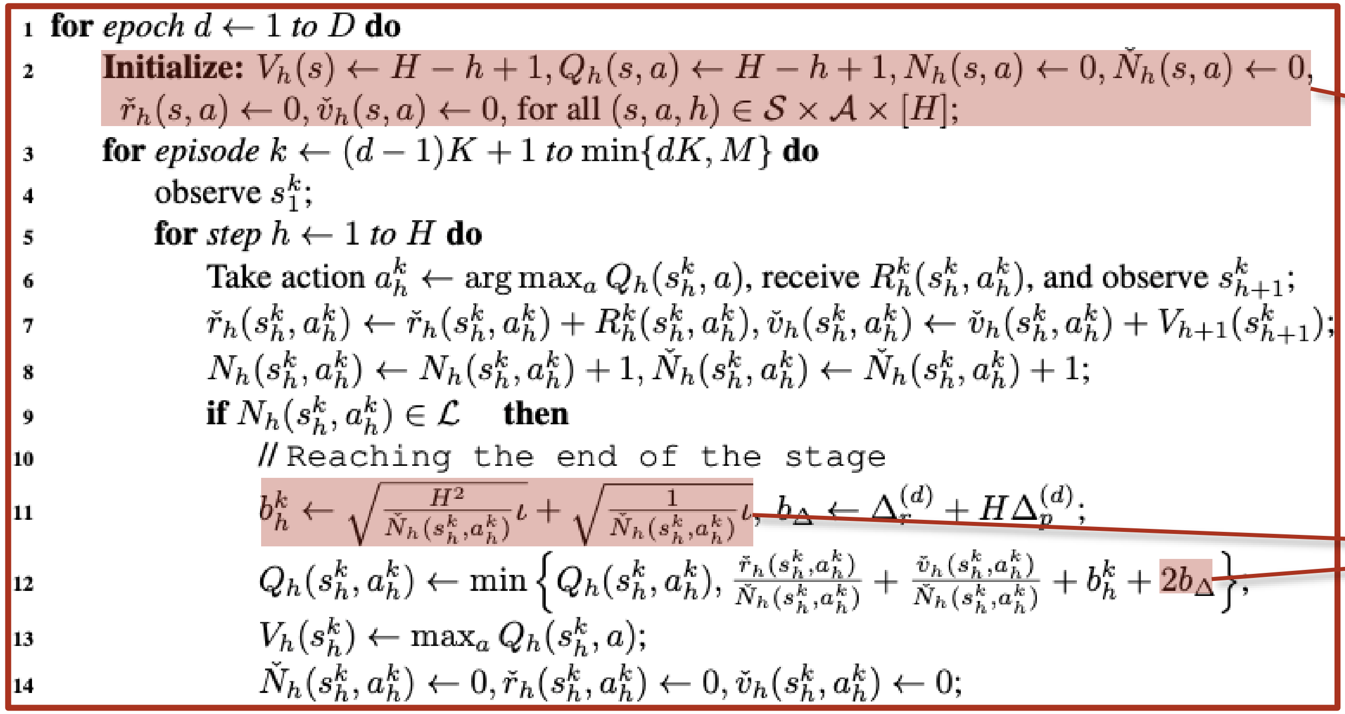
sample (s,a,r,s’)

update Q & V

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- **RestartQ-UCB**
  - A familiar **Q-Learning** algorithm… but over a **nonstationary** environment!



```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H - h + 1, Q_h(s,a) \leftarrow H - h + 1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0,$
    $\check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K + 1$ to $\min\{dK, M\}$ do
        observe $s_1^k$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h^k \leftarrow \arg\max_a Q_h(s_h^k, a)$, receive $R_h^k(s_h^k, a_h^k)$, and observe $s_{h+1}^k$;
            $\check{r}_h(s_h^k, a_h^k) \leftarrow \check{r}_h(s_h^k, a_h^k) + R_h^k(s_h^k, a_h^k), \check{v}_h(s_h^k, a_h^k) \leftarrow \check{v}_h(s_h^k, a_h^k) + V_{h+1}(s_{h+1}^k)$;
            $N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1, \check{N}_h(s_h^k, a_h^k) \leftarrow \check{N}_h(s_h^k, a_h^k) + 1$;
            if $N_h(s_h^k, a_h^k) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h^k \leftarrow \sqrt{\frac{H^2}{\check{N}_h(s_h^k, a_h^k)}\iota} + \sqrt{\frac{1}{\check{N}_h(s_h^k, a_h^k)}\iota}, b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k), \frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)} + \frac{\check{v}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)} + b_h^k + 2b_\Delta\right\}$;
                $V_h(s_h^k) \leftarrow \max_a Q_h(s_h^k, a)$;
                $\check{N}_h(s_h^k, a_h^k) \leftarrow 0, \check{r}_h(s_h^k, a_h^k) \leftarrow 0, \check{v}_h(s_h^k, a_h^k) \leftarrow 0$;
```

① Periodically forget everything (RestartQ-UCB)

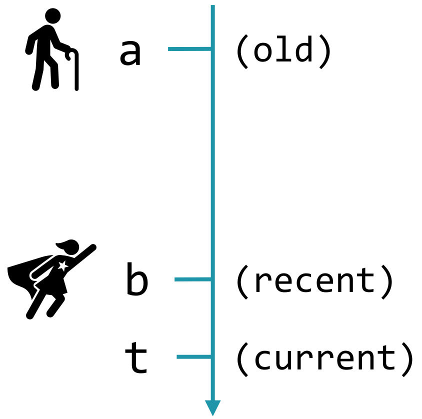
② Compute UCB bonus terms (RestartQ-UCB)

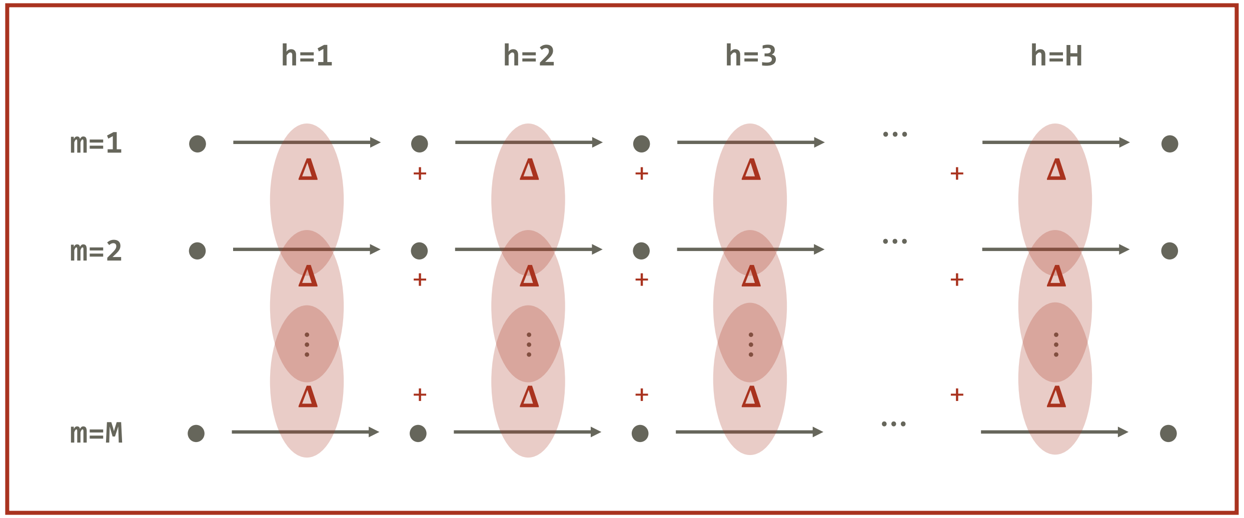
Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- ① **Why periodically forget everything to handle nonstationarity?**

Q-Learning Timeline



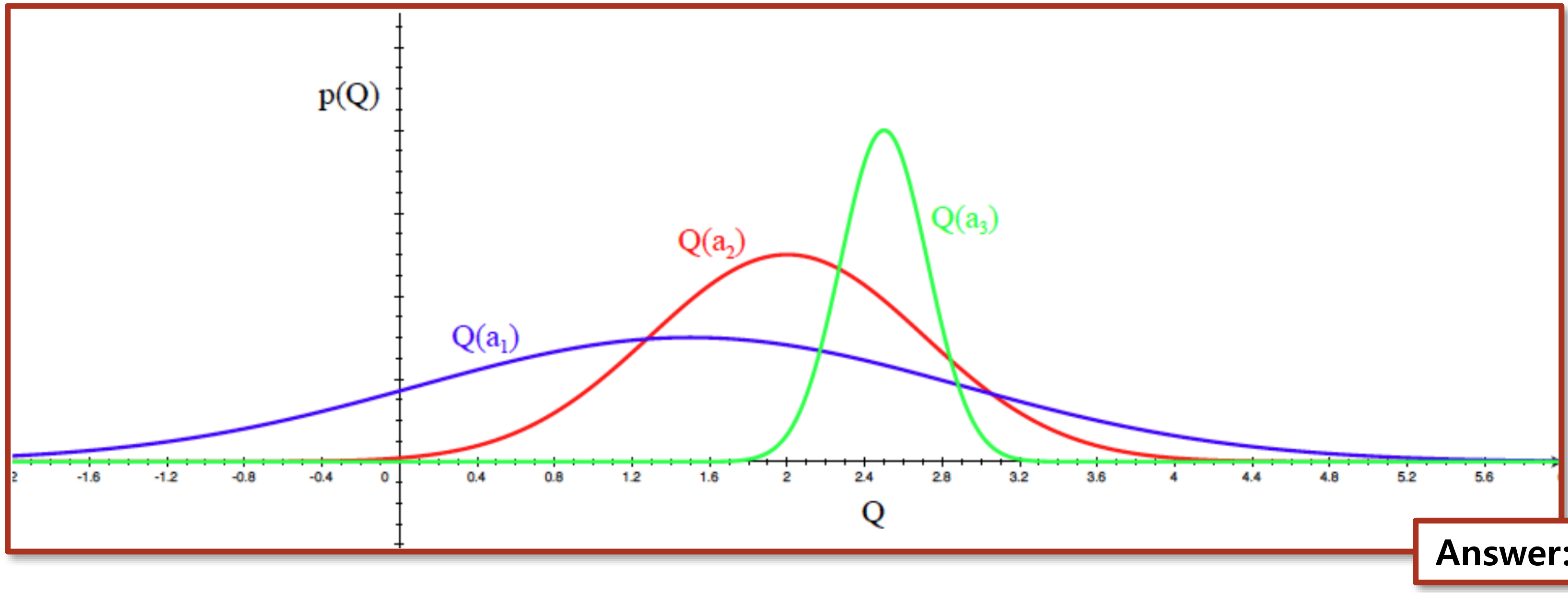


- Note that, our env is as much nonstationary as variation budget Δ
- Env at t is similar to env at b, but distant to env a
- This means, at timepoint t, Q_b is useful but Q_a is outdated
- So we better forget Q_a when learning Q_t!

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- ② **What are UCB(Upper Confidence Bound) terms for?**
  - **Question**: Suppose we want to estimate Q-values for actions a_1, a_2, a_3 as below. which action should we choose first for exploration? (width of distrib. = uncertainty)



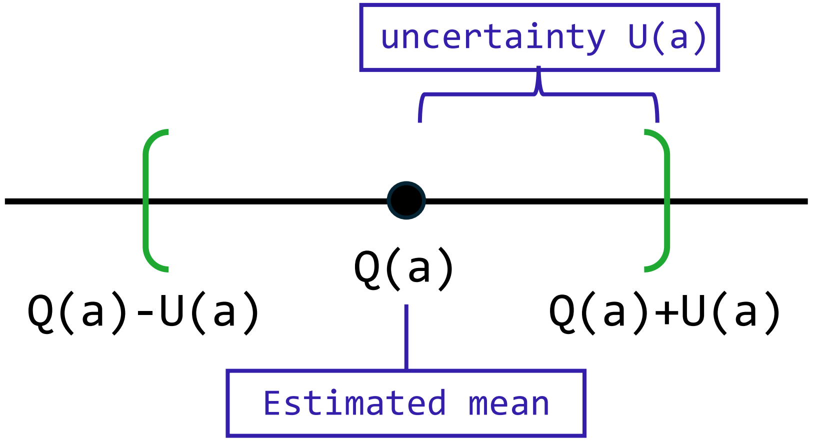
Answer: a1

  - **Principle1**: “The more uncertain we are about $Q(a)$, the more important it is to explore the action”

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- ② **What are UCB(Upper Confidence Bound) terms for?**
  - **UCB Algorithm**: “Pick the action according to the **upper bound** of the confidence interval



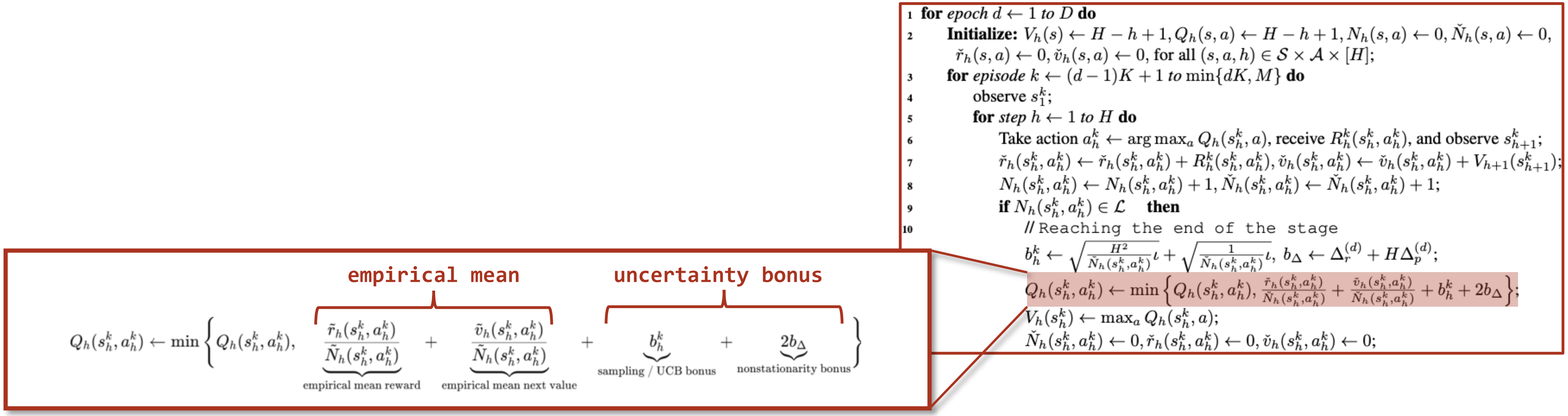
$$\boxed{a = \arg\max_{a} \; Q(a) + U(a)}$$

  - **Principle2**: “Optimism in the face of uncertainty”

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- ② **What are UCB(Upper Confidence Bound) terms for?**



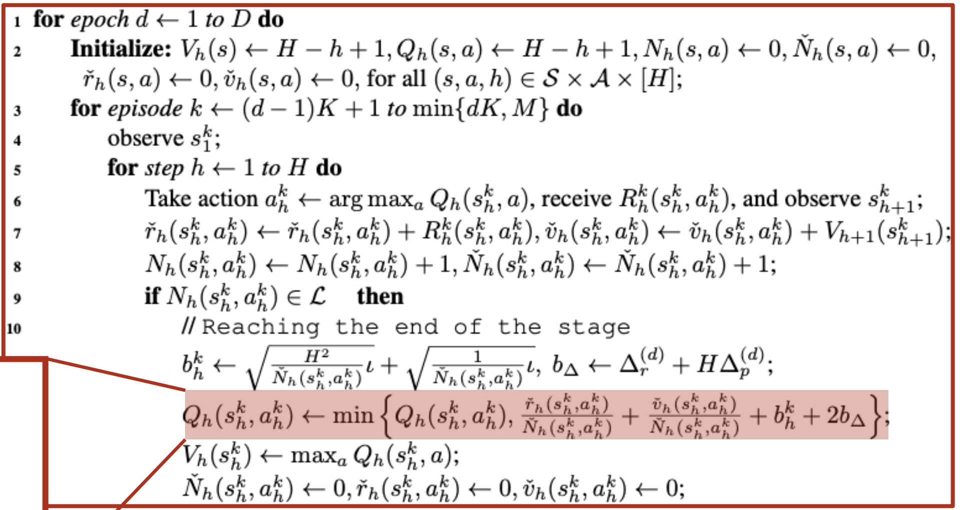
```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0,$
     $\check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1^k$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h^k \leftarrow \arg\max_a Q_h(s_h^k, a)$, receive $R_h^k(s_h^k, a_h^k)$, and observe $s_{h+1}^k$;
            $\check{r}_h(s_h^k, a_h^k) \leftarrow \check{r}_h(s_h^k, a_h^k) + R_h^k(s_h^k, a_h^k), \check{v}_h(s_h^k, a_h^k) \leftarrow \check{v}_h(s_h^k, a_h^k) + V_{h+1}(s_{h+1}^k)$;
            $N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1, \check{N}_h(s_h^k, a_h^k) \leftarrow \check{N}_h(s_h^k, a_h^k) + 1$;
            if $N_h(s_h^k, a_h^k) \in \mathcal{L}$    then
                // Reaching the end of the stage
                $b_h^k \leftarrow \sqrt{\frac{H^2}{\check{N}_h(s_h^k,a_h^k)}\iota} + \sqrt{\frac{1}{\check{N}_h(s_h^k,a_h^k)}\iota},\ b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k), \frac{\check{r}_h(s_h^k,a_h^k)}{\check{N}_h(s_h^k,a_h^k)} + \frac{\check{v}_h(s_h^k,a_h^k)}{\check{N}_h(s_h^k,a_h^k)} + b_h^k + 2b_\Delta\right\}$;
                $V_h(s_h^k) \leftarrow \max_a Q_h(s_h^k, a)$;
                $\check{N}_h(s_h^k, a_h^k) \leftarrow 0, \check{r}_h(s_h^k, a_h^k) \leftarrow 0, \check{v}_h(s_h^k, a_h^k) \leftarrow 0$;
```

$$Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k),\ \overbrace{\underbrace{\frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)}}_{\text{empirical mean reward}} + \underbrace{\frac{\check{v}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)}}_{\text{empirical mean next value}}}^{\text{empirical mean}} + \overbrace{\underbrace{b_h^k}_{\text{sampling / UCB bonus}} + \underbrace{2b_\Delta}_{\text{nonstationarity bonus}}}^{\text{uncertainty bonus}}\right\}$$

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)

- ② **What are UCB(Upper Confidence Bound) terms for?**



```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0,$
     $\check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0,$ for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1^k$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h^k \leftarrow \arg\max_a Q_h(s_h^k, a)$, receive $R_h^k(s_h^k, a_h^k)$, and observe $s_{h+1}^k$;
            $\check{r}_h(s_h^k, a_h^k) \leftarrow \check{r}_h(s_h^k, a_h^k) + R_h^k(s_h^k, a_h^k), \check{v}_h(s_h^k, a_h^k) \leftarrow \check{v}_h(s_h^k, a_h^k) + V_{h+1}(s_{h+1}^k)$;
            $N_h(s_h^k, a_h^k) \leftarrow N_h(s_h^k, a_h^k) + 1, \check{N}_h(s_h^k, a_h^k) \leftarrow \check{N}_h(s_h^k, a_h^k) + 1$;
            if $N_h(s_h^k, a_h^k) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h^k \leftarrow \sqrt{\frac{H^2}{\check{N}_h(s_h^k,a_h^k)}\iota} + \sqrt{\frac{1}{\check{N}_h(s_h^k,a_h^k)}\iota}, \; b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k), \frac{\check{r}_h(s_h^k,a_h^k)}{\check{N}_h(s_h^k,a_h^k)} + \frac{\check{v}_h(s_h^k,a_h^k)}{\check{N}_h(s_h^k,a_h^k)} + b_h^k + 2b_\Delta\right\}$;
                $V_h(s_h^k) \leftarrow \max_a Q_h(s_h^k, a)$;
                $\check{N}_h(s_h^k, a_h^k) \leftarrow 0, \check{r}_h(s_h^k, a_h^k) \leftarrow 0, \check{v}_h(s_h^k, a_h^k) \leftarrow 0$;
```

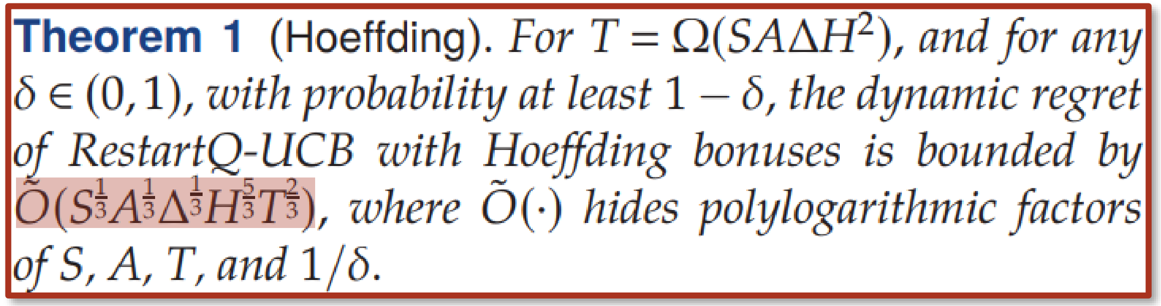
$$Q_h(s_h^k, a_h^k) \leftarrow \min\left\{Q_h(s_h^k, a_h^k), \overbrace{\underbrace{\frac{\check{r}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)}}_{\text{empirical mean reward}} + \underbrace{\frac{\check{v}_h(s_h^k, a_h^k)}{\check{N}_h(s_h^k, a_h^k)}}_{\text{empirical mean next value}}}^{\text{empirical mean}} + \overbrace{\underbrace{b_h^k}_{\text{sampling / UCB bonus}} + \underbrace{2b_\Delta}_{\text{nonstationarity bonus}}}^{\text{uncertainty bonus}}\right\}$$

$$b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$$

```
Nonstationary bonus:
the bigger delta
= the more nonstationary the model is
= the more uncertain the model is
= the more important it is to explore
```

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Hoeffding)



Theorem 1 (Hoeffding). *For $T = \Omega(SA\Delta H^2)$, and for any $\delta \in (0,1)$, with probability at least $1-\delta$, the dynamic regret of RestartQ-UCB with Hoeffding bonuses is bounded by $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{5}{3}}T^{\frac{2}{3}})$, where $\tilde{O}(\cdot)$ hides polylogarithmic factors of S, A, T, and $1/\delta$.*

S : number of states
A : number of actions
Δ : variation budget
H : number of steps per episode
T : total number of timesteps

Our plan

contribution RestartQ-UCB (Hoeffding) $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{5}{3}}T^{\frac{2}{3}})$,

contribution Double RestartQ-UCB $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$ >

contribution RestartQ-UCB (Freedman) $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$ >

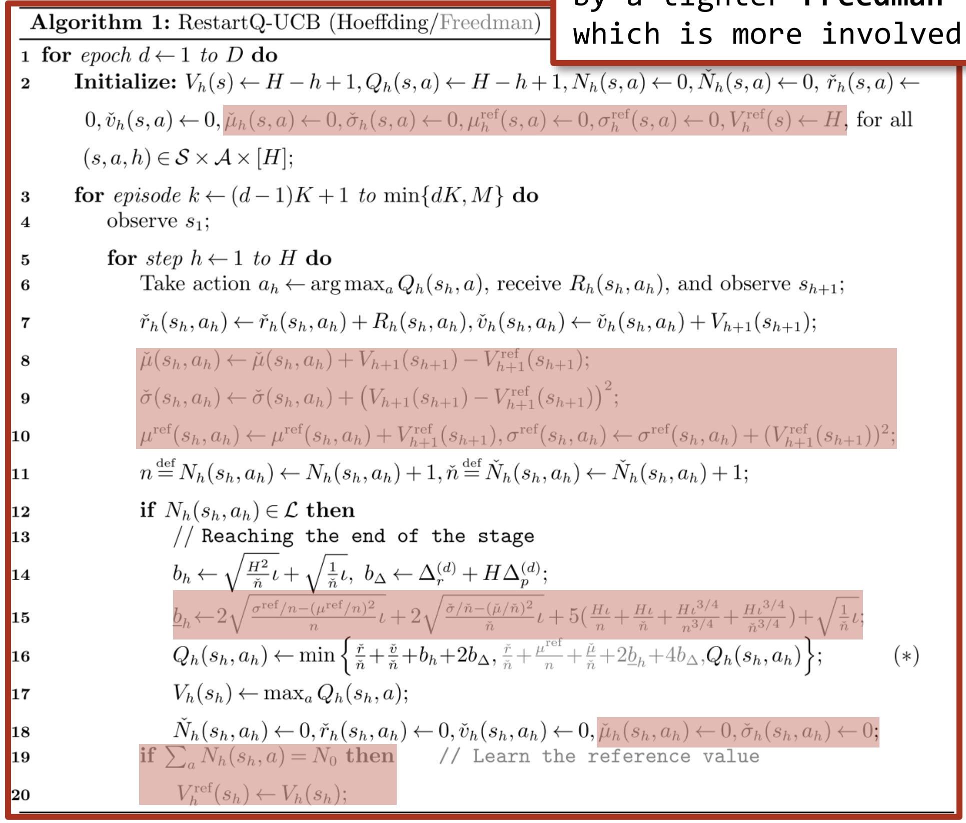
contribution Theoretical Lowerbound $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{2}{3}}T^{\frac{2}{3}})$

↑
Now here

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Freedman)

Hoeffding type UCB is replaced by a tighter **Freedman type UCB**, which is more involved..

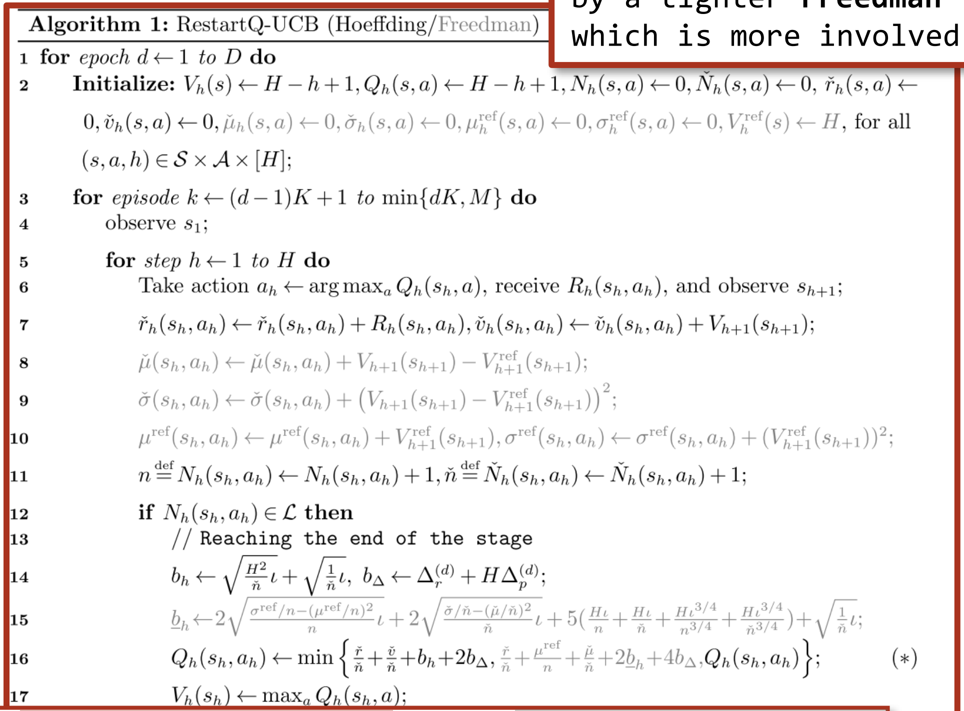


Algorithm 1: RestartQ-UCB (Hoeffding/Freedman)

```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0, \check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0, \check{\mu}_h(s,a) \leftarrow 0, \check{\sigma}_h(s,a) \leftarrow 0, \mu_h^{\mathrm{ref}}(s,a) \leftarrow 0, \sigma_h^{\mathrm{ref}}(s,a) \leftarrow 0, V_h^{\mathrm{ref}}(s) \leftarrow H$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h \leftarrow \arg\max_a Q_h(s_h,a)$, receive $R_h(s_h,a_h)$, and observe $s_{h+1}$;
            $\check{r}_h(s_h,a_h) \leftarrow \check{r}_h(s_h,a_h) + R_h(s_h,a_h), \check{v}_h(s_h,a_h) \leftarrow \check{v}_h(s_h,a_h) + V_{h+1}(s_{h+1})$;
            $\check{\mu}(s_h,a_h) \leftarrow \check{\mu}(s_h,a_h) + V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})$;
            $\check{\sigma}(s_h,a_h) \leftarrow \check{\sigma}(s_h,a_h) + \left(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})\right)^2$;
            $\mu^{\mathrm{ref}}(s_h,a_h) \leftarrow \mu^{\mathrm{ref}}(s_h,a_h) + V_{h+1}^{\mathrm{ref}}(s_{h+1}), \sigma^{\mathrm{ref}}(s_h,a_h) \leftarrow \sigma^{\mathrm{ref}}(s_h,a_h) + (V_{h+1}^{\mathrm{ref}}(s_{h+1}))^2$;
            $n \overset{\mathrm{def}}{=} N_h(s_h,a_h) \leftarrow N_h(s_h,a_h)+1, \check{n} \overset{\mathrm{def}}{=} \check{N}_h(s_h,a_h) \leftarrow \check{N}_h(s_h,a_h)+1$;
            if $N_h(s_h,a_h) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h \leftarrow \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{1}{\check{n}}\iota},\ b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $\underline{b}_h \leftarrow 2\sqrt{\frac{\sigma^{\mathrm{ref}}/n - (\mu^{\mathrm{ref}}/n)^2}{n}\iota} + 2\sqrt{\frac{\check{\sigma}/\check{n} - (\check{\mu}/\check{n})^2}{\check{n}}\iota} + 5\left(\frac{H\iota}{n} + \frac{H\iota}{\check{n}} + \frac{H\iota^{3/4}}{n^{3/4}} + \frac{H\iota^{3/4}}{\check{n}^{3/4}}\right) + \sqrt{\frac{1}{\check{n}}\iota}$;
                $Q_h(s_h,a_h) \leftarrow \min\left\{\frac{\check{r}}{\check{n}} + \frac{\check{v}}{\check{n}} + b_h + 2b_\Delta, \frac{\check{r}}{\check{n}} + \frac{\mu^{\mathrm{ref}}}{n} + \frac{\check{\mu}}{\check{n}} + 2\underline{b}_h + 4b_\Delta, Q_h(s_h,a_h)\right\}$;    (*)
                $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$;
                $\check{N}_h(s_h,a_h) \leftarrow 0, \check{r}_h(s_h,a_h) \leftarrow 0, \check{v}_h(s_h,a_h) \leftarrow 0, \check{\mu}_h(s_h,a_h) \leftarrow 0, \check{\sigma}_h(s_h,a_h) \leftarrow 0$;
            if $\sum_a N_h(s_h,a) = N_0$ then    // Learn the reference value
                $V_h^{\mathrm{ref}}(s_h) \leftarrow V_h(s_h)$;
```

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

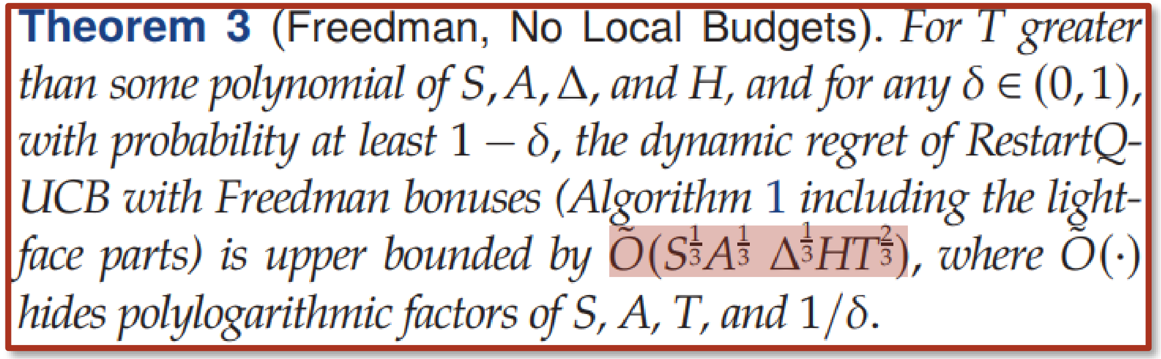
RestartQ-UCB (Freedman)



Algorithm 1: RestartQ-UCB (Hoeffding/Freedman)

```
for epoch $d \leftarrow 1$ to $D$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0, \check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0, \check{\mu}_h(s,a) \leftarrow 0, \check{\sigma}_h(s,a) \leftarrow 0, \mu_h^{\mathrm{ref}}(s,a) \leftarrow 0, \sigma_h^{\mathrm{ref}}(s,a) \leftarrow 0, V_h^{\mathrm{ref}}(s) \leftarrow H$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h \leftarrow \arg\max_a Q_h(s_h,a)$, receive $R_h(s_h,a_h)$, and observe $s_{h+1}$;
            $\check{r}_h(s_h,a_h) \leftarrow \check{r}_h(s_h,a_h) + R_h(s_h,a_h), \check{v}_h(s_h,a_h) \leftarrow \check{v}_h(s_h,a_h) + V_{h+1}(s_{h+1})$;
            $\check{\mu}(s_h,a_h) \leftarrow \check{\mu}(s_h,a_h) + V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})$;
            $\check{\sigma}(s_h,a_h) \leftarrow \check{\sigma}(s_h,a_h) + \left(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})\right)^2$;
            $\mu^{\mathrm{ref}}(s_h,a_h) \leftarrow \mu^{\mathrm{ref}}(s_h,a_h) + V_{h+1}^{\mathrm{ref}}(s_{h+1}), \sigma^{\mathrm{ref}}(s_h,a_h) \leftarrow \sigma^{\mathrm{ref}}(s_h,a_h) + (V_{h+1}^{\mathrm{ref}}(s_{h+1}))^2$;
            $n \overset{\mathrm{def}}{=} N_h(s_h,a_h) \leftarrow N_h(s_h,a_h)+1, \check{n} \overset{\mathrm{def}}{=} \check{N}_h(s_h,a_h) \leftarrow \check{N}_h(s_h,a_h)+1$;
            if $N_h(s_h,a_h) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h \leftarrow \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{1}{\check{n}}\iota},\ b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $\underline{b}_h \leftarrow 2\sqrt{\frac{\sigma^{\mathrm{ref}}/n - (\mu^{\mathrm{ref}}/n)^2}{n}\iota} + 2\sqrt{\frac{\check{\sigma}/\check{n} - (\check{\mu}/\check{n})^2}{\check{n}}\iota} + 5(\frac{H\iota}{n} + \frac{H\iota}{\check{n}} + \frac{H\iota^{3/4}}{n^{3/4}} + \frac{H\iota^{3/4}}{\check{n}^{3/4}}) + \sqrt{\frac{1}{\check{n}}\iota}$;
                $Q_h(s_h,a_h) \leftarrow \min\left\{\frac{\check{r}}{\check{n}} + \frac{\check{v}}{\check{n}} + b_h + 2b_\Delta, \frac{\check{r}}{\check{n}} + \frac{\mu^{\mathrm{ref}}}{n} + \frac{\check{\mu}}{\check{n}} + 2\underline{b}_h + 4b_\Delta, Q_h(s_h,a_h)\right\}$;        (*)
                $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$;
```

Hoeffding type UCB is replaced by a tighter **Freedman type UCB**, which is more involved..



Theorem 3 (Freedman, No Local Budgets). *For T greater than some polynomial of S, A, Δ, and H, and for any $\delta \in (0,1)$, with probability at least $1-\delta$, the dynamic regret of RestartQ-UCB with Freedman bonuses (Algorithm 1 including the light-face parts) is upper bounded by $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$, where $\tilde{O}(\cdot)$ hides polylogarithmic factors of S, A, T, and $1/\delta$.*

RestartQ-UCB (Hoeffding)		Double RestartQ-UCB		RestartQ-UCB (Freedman)		Theoretical Lowerbound
$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{5}{3}}T^{\frac{2}{3}})$	,	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$	>	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$	>	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{2}{3}}T^{\frac{2}{3}})$

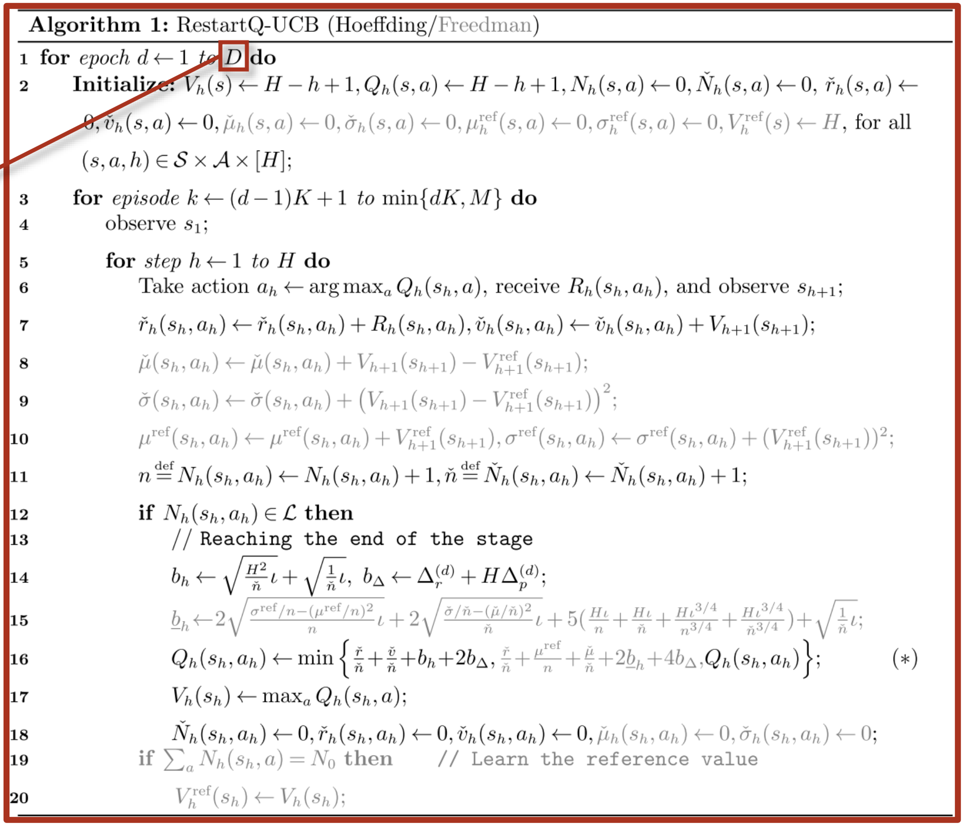
Near-optimal!

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

RestartQ-UCB (Freedman)

Caveat: For optimality,
$D = S^{-1/3} \cdot A^{-1/3} \cdot \Delta^{2/3} \cdot H^{-2/3} \cdot T^{1/3}$
i.e. restart scheduling
requires prior knowledge of Δ

Can we get rid of this?

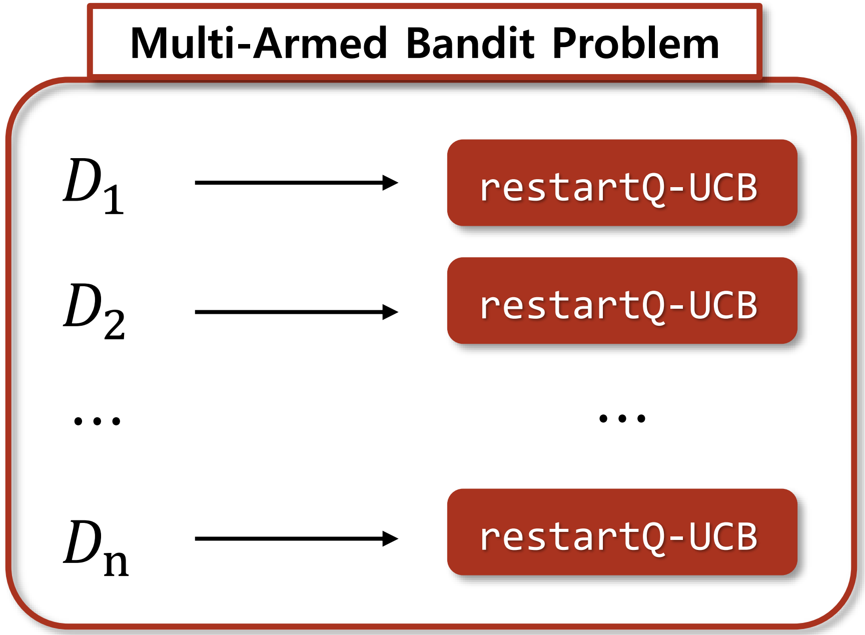


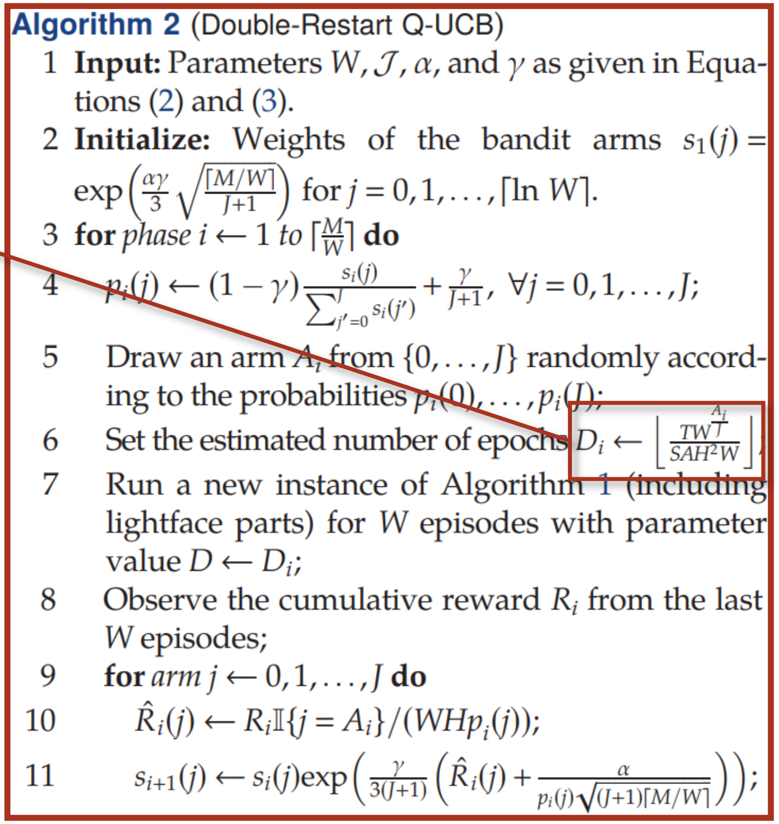
```
Algorithm 1: RestartQ-UCB (Hoeffding/Freedman)
for epoch $d \leftarrow 1$ to $\boxed{D}$ do
    Initialize: $V_h(s) \leftarrow H-h+1, Q_h(s,a) \leftarrow H-h+1, N_h(s,a) \leftarrow 0, \check{N}_h(s,a) \leftarrow 0, \check{r}_h(s,a) \leftarrow 0, \check{v}_h(s,a) \leftarrow 0, \check{\mu}_h(s,a) \leftarrow 0, \check{\sigma}_h(s,a) \leftarrow 0, \mu_h^{\mathrm{ref}}(s,a) \leftarrow 0, \sigma_h^{\mathrm{ref}}(s,a) \leftarrow 0, V_h^{\mathrm{ref}}(s) \leftarrow H$, for all $(s,a,h) \in \mathcal{S} \times \mathcal{A} \times [H]$;
    for episode $k \leftarrow (d-1)K+1$ to $\min\{dK, M\}$ do
        observe $s_1$;
        for step $h \leftarrow 1$ to $H$ do
            Take action $a_h \leftarrow \arg\max_a Q_h(s_h,a)$, receive $R_h(s_h,a_h)$, and observe $s_{h+1}$;
            $\check{r}_h(s_h,a_h) \leftarrow \check{r}_h(s_h,a_h) + R_h(s_h,a_h), \check{v}_h(s_h,a_h) \leftarrow \check{v}_h(s_h,a_h) + V_{h+1}(s_{h+1})$;
            $\check{\mu}(s_h,a_h) \leftarrow \check{\mu}(s_h,a_h) + V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})$;
            $\check{\sigma}(s_h,a_h) \leftarrow \check{\sigma}(s_h,a_h) + \left(V_{h+1}(s_{h+1}) - V_{h+1}^{\mathrm{ref}}(s_{h+1})\right)^2$;
            $\mu^{\mathrm{ref}}(s_h,a_h) \leftarrow \mu^{\mathrm{ref}}(s_h,a_h) + V_{h+1}^{\mathrm{ref}}(s_{h+1}), \sigma^{\mathrm{ref}}(s_h,a_h) \leftarrow \sigma^{\mathrm{ref}}(s_h,a_h) + (V_{h+1}^{\mathrm{ref}}(s_{h+1}))^2$;
            $n \overset{\mathrm{def}}{=} N_h(s_h,a_h) \leftarrow N_h(s_h,a_h) + 1, \check{n} \overset{\mathrm{def}}{=} \check{N}_h(s_h,a_h) \leftarrow \check{N}_h(s_h,a_h) + 1$;
            if $N_h(s_h,a_h) \in \mathcal{L}$ then
                // Reaching the end of the stage
                $b_h \leftarrow \sqrt{\frac{H^2}{\check{n}}\iota} + \sqrt{\frac{1}{\check{n}}\iota},\ b_\Delta \leftarrow \Delta_r^{(d)} + H\Delta_p^{(d)}$;
                $\underline{b}_h \leftarrow 2\sqrt{\frac{\sigma^{\mathrm{ref}}/n - (\mu^{\mathrm{ref}}/n)^2}{n}\iota} + 2\sqrt{\frac{\check{\sigma}/\check{n} - (\check{\mu}/\check{n})^2}{\check{n}}\iota} + 5(\frac{H\iota}{n} + \frac{H\iota}{\check{n}} + \frac{H\iota^{3/4}}{n^{3/4}} + \frac{H\iota^{3/4}}{\check{n}^{3/4}}) + \sqrt{\frac{1}{\check{n}}\iota}$;
                $Q_h(s_h,a_h) \leftarrow \min\left\{\frac{\check{r}}{\check{n}} + \frac{\check{v}}{\check{n}} + b_h + 2b_\Delta, \frac{\check{r}}{\check{n}} + \frac{\mu^{\mathrm{ref}}}{n} + \frac{\check{\mu}}{\check{n}} + 2\underline{b}_h + 4b_\Delta, Q_h(s_h,a_h)\right\}$;   $(*)$
                $V_h(s_h) \leftarrow \max_a Q_h(s_h,a)$;
                $\check{N}_h(s_h,a_h) \leftarrow 0, \check{r}_h(s_h,a_h) \leftarrow 0, \check{v}_h(s_h,a_h) \leftarrow 0, \check{\mu}_h(s_h,a_h) \leftarrow 0, \check{\sigma}_h(s_h,a_h) \leftarrow 0$;
            if $\sum_a N_h(s_h,a) = N_0$ then      // Learn the reference value
                $V_h^{\mathrm{ref}}(s_h) \leftarrow V_h(s_h)$;
```

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Double-Restart Q-UCB

Here, D is “learned” in an online manner, rather than “given” as a parameter



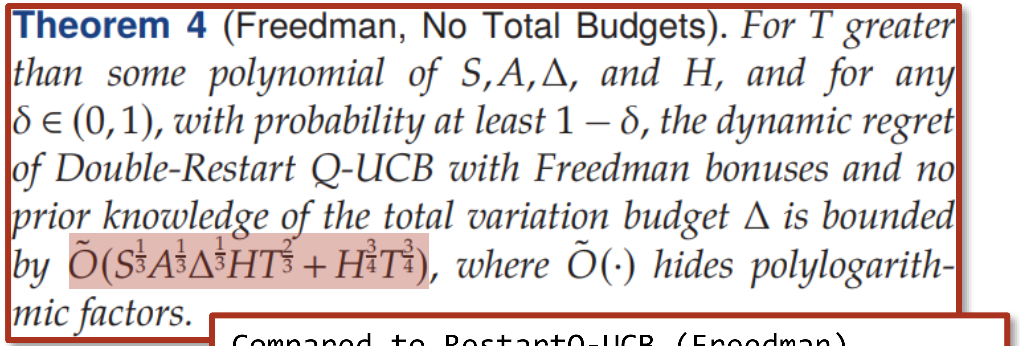


Algorithm 2 (Double-Restart Q-UCB)

```
Input: Parameters $W, \mathcal{J}, \alpha$, and $\gamma$ as given in Equations (2) and (3).
Initialize: Weights of the bandit arms $s_1(j) = \exp\left(\frac{\alpha\gamma}{3}\sqrt{\frac{\lceil M/W \rceil}{J+1}}\right)$ for $j = 0, 1, \ldots, \lceil \ln W \rceil$.
for phase $i \leftarrow 1$ to $\lceil \frac{M}{W} \rceil$ do
    $p_i(j) \leftarrow (1-\gamma)\frac{s_i(j)}{\sum_{j'=0}^{J} s_i(j')} + \frac{\gamma}{J+1}, \forall j = 0, 1, \ldots, J$;
    Draw an arm $A_i$ from $\{0, \ldots, J\}$ randomly according to the probabilities $p_i(0), \ldots, p_i(J)$;
    Set the estimated number of epochs $D_i \leftarrow \left\lfloor \frac{TW^{\frac{A_i}{J}}}{SAH^2W} \right\rfloor$
    Run a new instance of Algorithm 1 (including lightface parts) for $W$ episodes with parameter value $D \leftarrow D_i$;
    Observe the cumulative reward $R_i$ from the last $W$ episodes;
    for arm $j \leftarrow 0, 1, \ldots, J$ do
        $\hat{R}_i(j) \leftarrow R_i \mathbb{I}\{j = A_i\}/(WHp_i(j))$;
        $s_{i+1}(j) \leftarrow s_i(j)\exp\left(\frac{\gamma}{3(J+1)}\left(\hat{R}_i(j) + \frac{\alpha}{p_i(j)\sqrt{(J+1)\lceil M/W \rceil}}\right)\right)$;
```

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

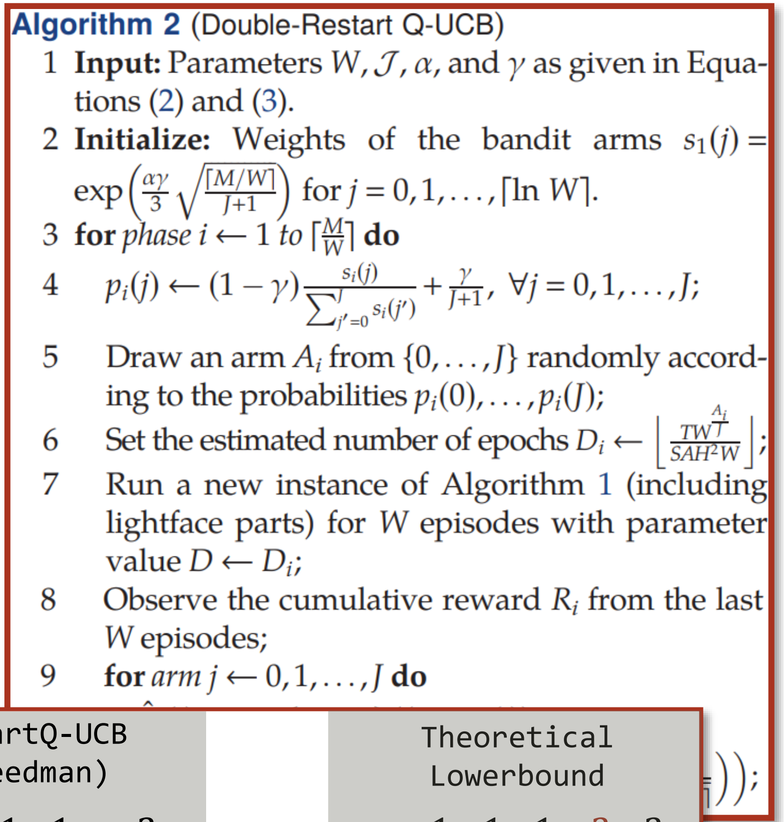
Double-Restart Q-UCB



Theorem 4 (Freedman, No Total Budgets). *For T greater than some polynomial of S, A, Δ, and H, and for any $\delta \in (0,1)$, with probability at least $1-\delta$, the dynamic regret of Double-Restart Q-UCB with Freedman bonuses and no prior knowledge of the total variation budget Δ is bounded by $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$, where $\tilde{O}(\cdot)$ hides polylogarithmic factors.*

Compared to RestartQ-UCB (Freedman),

- more overhead for learning D
- but more robust to unknown, irregular, and abrupt environment changes



Algorithm 2 (Double-Restart Q-UCB)

```
Input: Parameters $W, \mathcal{J}, \alpha$, and $\gamma$ as given in Equations (2) and (3).
Initialize: Weights of the bandit arms $s_1(j) = \exp\left(\frac{\alpha\gamma}{3}\sqrt{\frac{\lceil M/W \rceil}{J+1}}\right)$ for $j = 0, 1, \ldots, \lceil \ln W \rceil$.
for phase $i \leftarrow 1$ to $\lceil \frac{M}{W} \rceil$ do
    $p_i(j) \leftarrow (1-\gamma)\frac{s_i(j)}{\sum_{j'=0}^{J} s_i(j')} + \frac{\gamma}{J+1}, \forall j = 0, 1, \ldots, J$;
    Draw an arm $A_i$ from $\{0, \ldots, J\}$ randomly according to the probabilities $p_i(0), \ldots, p_i(J)$;
    Set the estimated number of epochs $D_i \leftarrow \left\lfloor \frac{TW^{\frac{A_i}{J}}}{SAH^2W} \right\rfloor$;
    Run a new instance of Algorithm 1 (including lightface parts) for $W$ episodes with parameter value $D \leftarrow D_i$;
    Observe the cumulative reward $R_i$ from the last $W$ episodes;
    for arm $j \leftarrow 0, 1, \ldots, J$ do
```

RestartQ-UCB (Hoeffding)		Double RestartQ-UCB		RestartQ-UCB (Freedman)		Theoretical Lowerbound
$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{5}{3}}T^{\frac{2}{3}})$	,	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$	>	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$	>	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{2}{3}}T^{\frac{2}{3}})$

Still near-optimal!

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Experiment

- ## Compared Algorithms

		Restart?	Exploration	Framework	Time Complexity
	RestartQ-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}})$
	Double-Restart Q-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}} + \mathrm{H}^{\frac{3}{4}}\mathrm{T}^{\frac{3}{4}})$
Baseline	LSVI-UCB-Restart	O	UCB	Least-Squares Value Iteration	$\tilde{O}(S^{\frac{4}{3}}\mathrm{A}^{\frac{4}{3}}\Delta^{\frac{1}{3}}\mathrm{H}^{\frac{4}{3}}\mathrm{T}^{\frac{2}{3}})$ **(SOTA)**
	Q-Learning UCB	**X**	UCB	Q-Learning	-
	Epsilon-Greedy	O	**ϵ-greedy**	Q-Learning	-

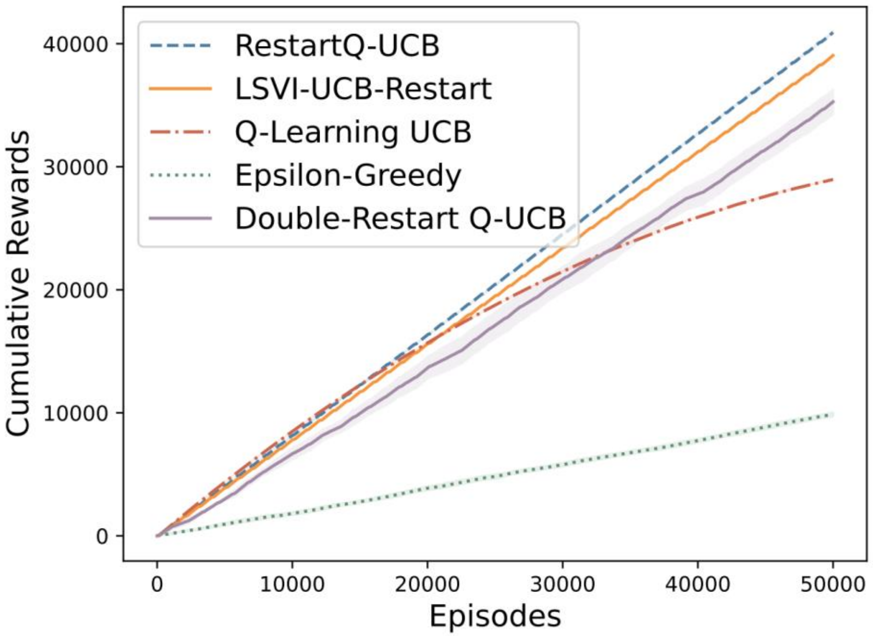
Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Experiment

- **Compared Algorithms**

		Restart?	Exploration	Framework	Time Complexity
	RestartQ-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}})$
	Double-Restart Q-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}} + \mathrm{H}^{\frac{3}{4}}\mathrm{T}^{\frac{3}{4}})$
Baseline	LSVI-UCB-Restart	O	UCB	Least-Squares Value Iteration	$\tilde{O}(S^{\frac{4}{3}}\mathrm{A}^{\frac{4}{3}}\Delta^{\frac{1}{3}}\mathrm{H}^{\frac{4}{3}}\mathrm{T}^{\frac{2}{3}})$ (SOTA)
	Q-Learning UCB	X	UCB	Q-Learning	-
	Epsilon-Greedy	O	ϵ-greedy	Q-Learning	-

- **Simulation**
  - Benchmark
    - Bidirectional Diabolical Combination Lock
    - particularly difficult for **exploration**



Q-Learning UCB
: no restart

Epsilon-Greedy
: no UCB

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

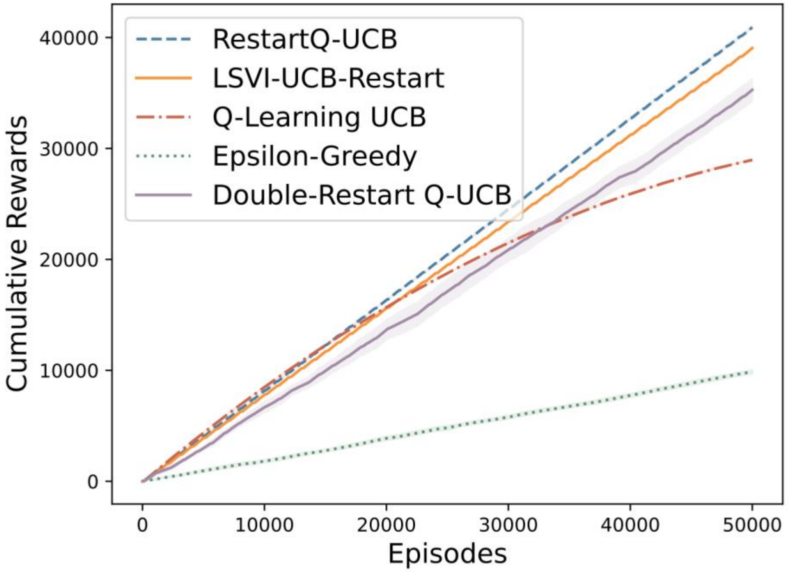
Experiment

- **Compared Algorithms**

	Restart?	Exploration	Framework	Time Complexity
RestartQ-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$
Double-Restart Q-UCB	O	UCB	Q-Learning	$\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$
LSVI-UCB-Restart	O	UCB	Least-Squares Value Iteration	$\tilde{O}(S^{\frac{4}{3}}A^{\frac{4}{3}}\Delta^{\frac{1}{3}}H^{\frac{4}{3}}T^{\frac{2}{3}})$ (SOTA)
Q-Learning UCB	X	UCB	Q-Learning	-
Epsilon-Greedy	O	ϵ-greedy	Q-Learning	-

Baseline

- **Simulation**



Algorithm	Time per episode
RestartQ-UCB	0.102 ms
Double-Restart Q-UCB	0.105 ms
LSVI-UCB-Restart	57.65 ms
Q-Learning UCB	0.098 ms
Epsilon-Greedy	0.123 ms

only little overhead!

very slow!

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Concluding Remark

- **Proposed two model-free learning algorithms for nonstationary MDPs**

  - RestartQ-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}})$

  - Double-Restart Q-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}} + \mathrm{H}^{\frac{3}{4}}\mathrm{T}^{\frac{3}{4}})$

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Concluding Remark

- **Proposed two model-free learning algorithms for nonstationary MDPs**
  - RestartQ-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}})$
  - Double-Restart Q-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}} + \mathrm{H}^{\frac{3}{4}}\mathrm{T}^{\frac{3}{4}})$
- **And showed that they are near-optimal**
  - w.r.t. the information-theoretical lowerbound $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}^{\frac{2}{3}}\mathrm{T}^{\frac{2}{3}})$

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Concluding Remark

- **Proposed two model-free learning algorithms for nonstationary MDPs**
  - RestartQ-UCB : $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}})$
  - Double-Restart Q-UCB : $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}HT^{\frac{2}{3}} + H^{\frac{3}{4}}T^{\frac{3}{4}})$
- **And showed that they are near-optimal**
  - w.r.t. the information-theoretical lowerbound $\tilde{O}(S^{\frac{1}{3}}A^{\frac{1}{3}}\Delta^{\frac{1}{3}}H^{\frac{2}{3}}T^{\frac{2}{3}})$
- **With supporting empirical results**
  - competitive in both rewards & time
  - justification of Restart & UCB in their design

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Concluding Remark

- **Proposed two model-free learning algorithms for nonstationary MDPs**
  - RestartQ-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}})$
  - Double-Restart Q-UCB : $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}\mathrm{T}^{\frac{2}{3}} + \mathrm{H}^{\frac{3}{4}}\mathrm{T}^{\frac{3}{4}})$
- **And showed that they are near-optimal**
  - w.r.t. the information-theoretical lowerbound $\tilde{O}(S^{\frac{1}{3}}\mathrm{A}^{\frac{1}{3}}\Delta^{\frac{1}{3}}\mathrm{H}^{\frac{2}{3}}\mathrm{T}^{\frac{2}{3}})$
- **With supporting empirical results**
  - competitive in both rewards & time
  - justification of Restart & UCB in their design
- **Future Direction**
  - Close the remaining $\tilde{O}(H^{\frac{1}{3}})$ gap

Mao, W., Zhang, K., Zhu, R., Simchi-Levi, D., & Başar, T. (2025). Model-Free Nonstationary Reinforcement Learning: Near-Optimal Regret and Applications in Multiagent Reinforcement Learning and Inventory Control. *Management Science*, *71*(2), 1564-1580.

Thank You!