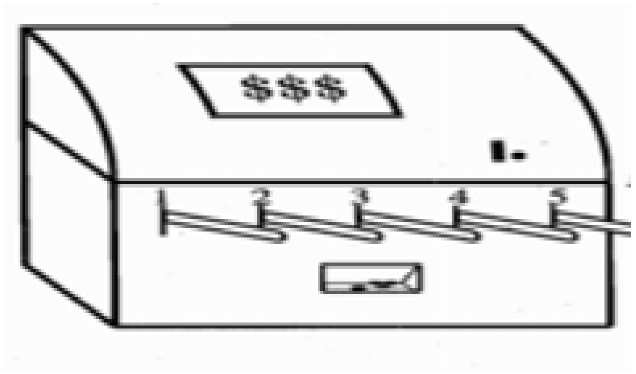
Multi-arm Bandits

presented by Zhenghui Wang

Contents

k-Armed Bandit



notation

- A_t: the action selected on time step t
- R_t: corresponding reward to A_t
- $q_*(a)$: $q_*(a) = \mathbf{E}[\, R_t \mid A_t = a \,]$
- $Q_t(a)$: estimated value of action a at time t $Q_t(a) \approx q_*(a)$

Action-Value Methods

- the *sample-average* method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbf{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbf{1}_{A_i=a}}$$

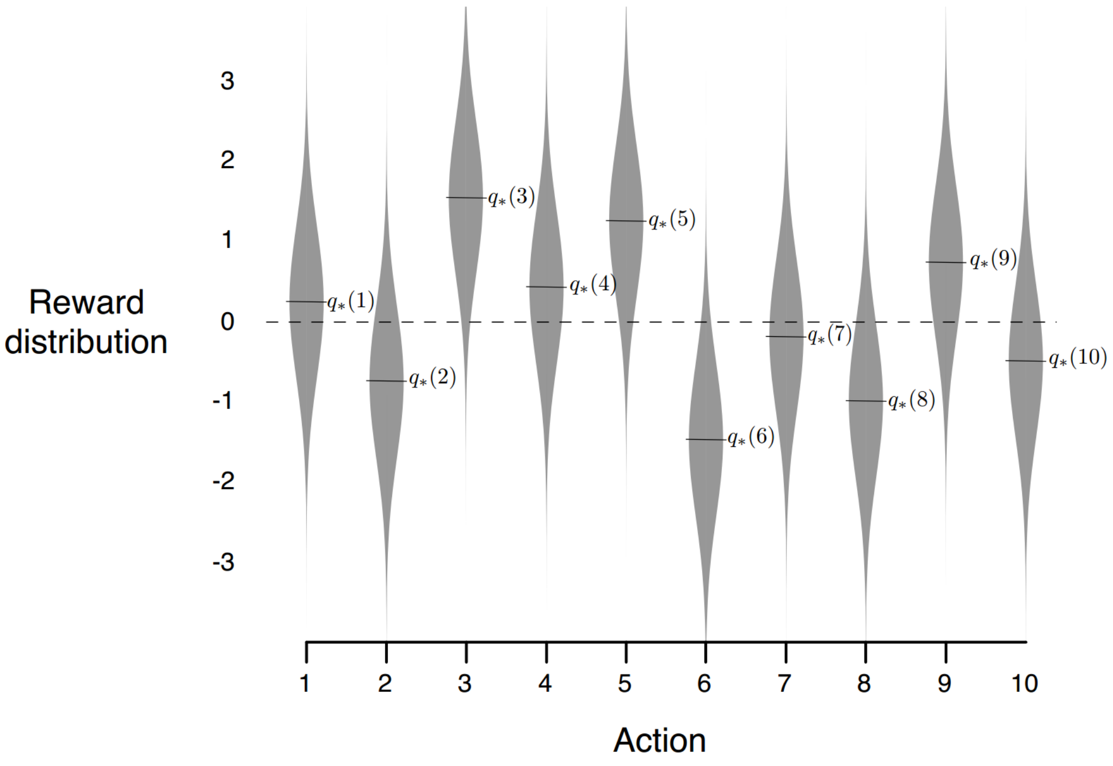
- What if the denominator is **zero?**
- law of large numbers

Action-Value Methods (cont.)

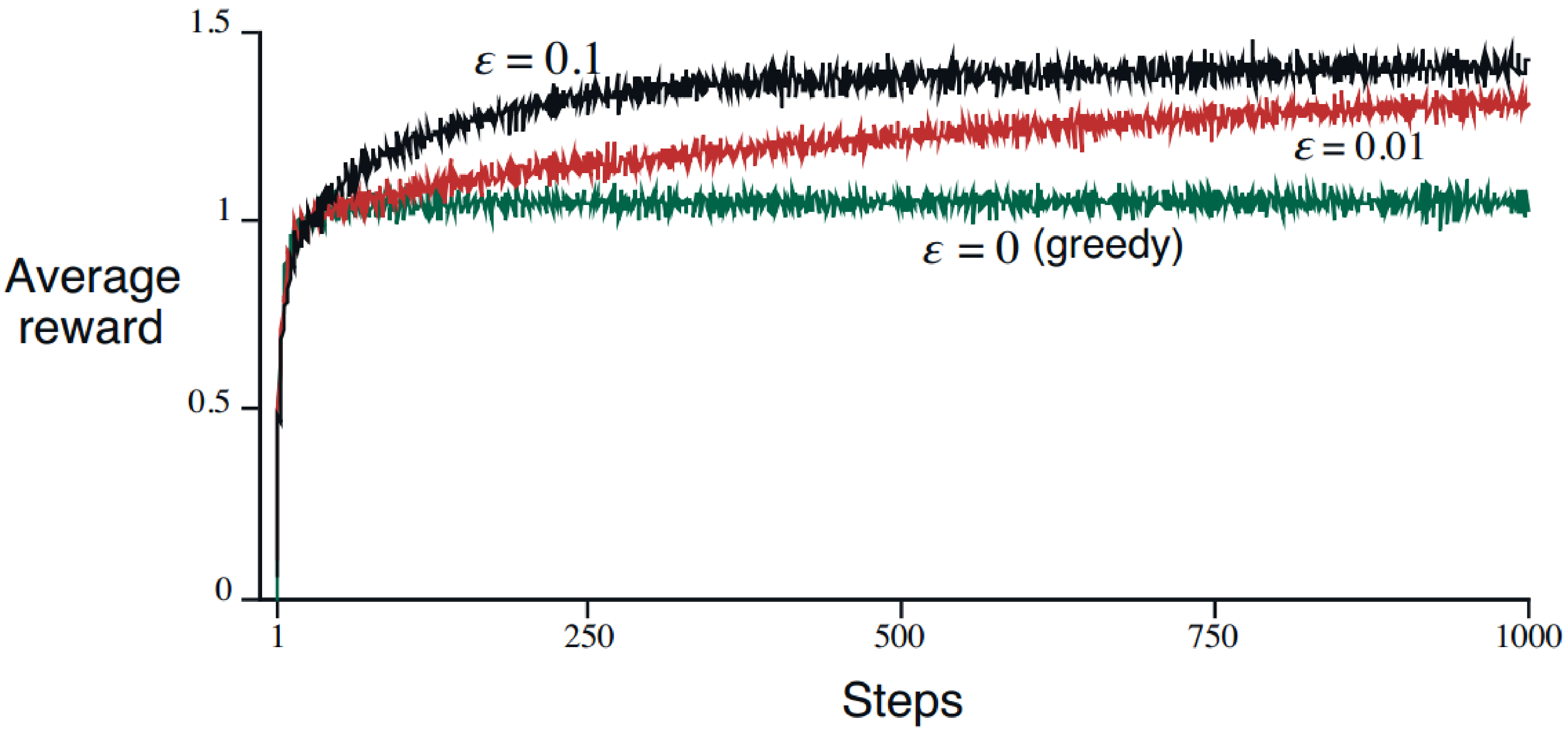
- Methods
  - *greedy* action selection method

$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$$

  - $\varepsilon - greedy$ methods

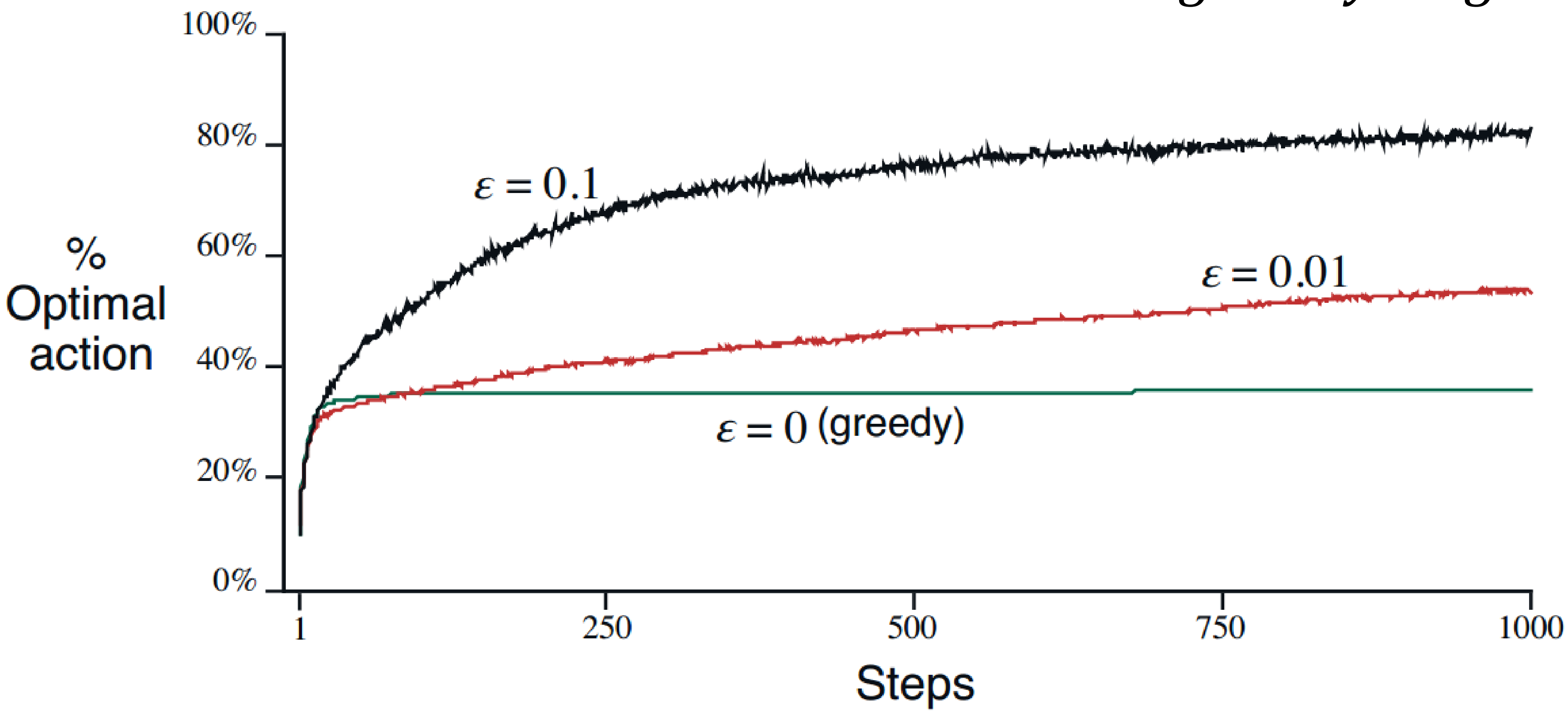


Action-Value Methods (cont.)



Action-Value Methods (cont.)

$\varepsilon - greedy$ or $greedy$?



Incremental Implementation

- Normal method

Recall: $$Q_t(a) \doteq \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbf{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbf{1}_{A_i=a}}$$

$$Q_n \doteq \frac{R_1 + R_2 + \cdots + R_{n-1}}{n-1}.$$

- Incremental Implementation

$$\begin{aligned}
Q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\
&= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\left(R_n + (n-1)\frac{1}{n-1}\sum_{i=1}^{n-1} R_i\right) \\
&= \frac{1}{n}\Big(R_n + (n-1)Q_n\Big) \\
&= \frac{1}{n}\Big(R_n + nQ_n - Q_n\Big) \\
&= Q_n + \frac{1}{n}\Big[R_n - Q_n\Big],
\end{aligned}$$

Incremental Implementation (cont.)

$$NewEstimate \leftarrow OldEstimate \; + \; StepSize \Big[Target - OldEstimate \Big].$$

- $\Big[Target - OldEstimate \Big]$: $error$

Tracking a Nonstationary Problem

$$NewEstimate \leftarrow OldEstimate \ + \ StepSize \Big[Target - OldEstimate \Big].$$

- *exponential, recency-weighted average*

$$\begin{aligned}
Q_{n+1} &\doteq Q_n + \alpha \Big[R_n - Q_n \Big] \\
&= \alpha R_n + (1-\alpha) Q_n \\
&= \alpha R_n + (1-\alpha) \left[\alpha R_{n-1} + (1-\alpha) Q_{n-1} \right] \\
&= \alpha R_n + (1-\alpha) \alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha) \alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\
&\qquad \qquad \cdots + (1-\alpha)^{n-1} \alpha R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i .
\end{aligned}$$

conditions required to assure convergence

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty.$$

the steps are large enough to eventually overcome any initial conditions

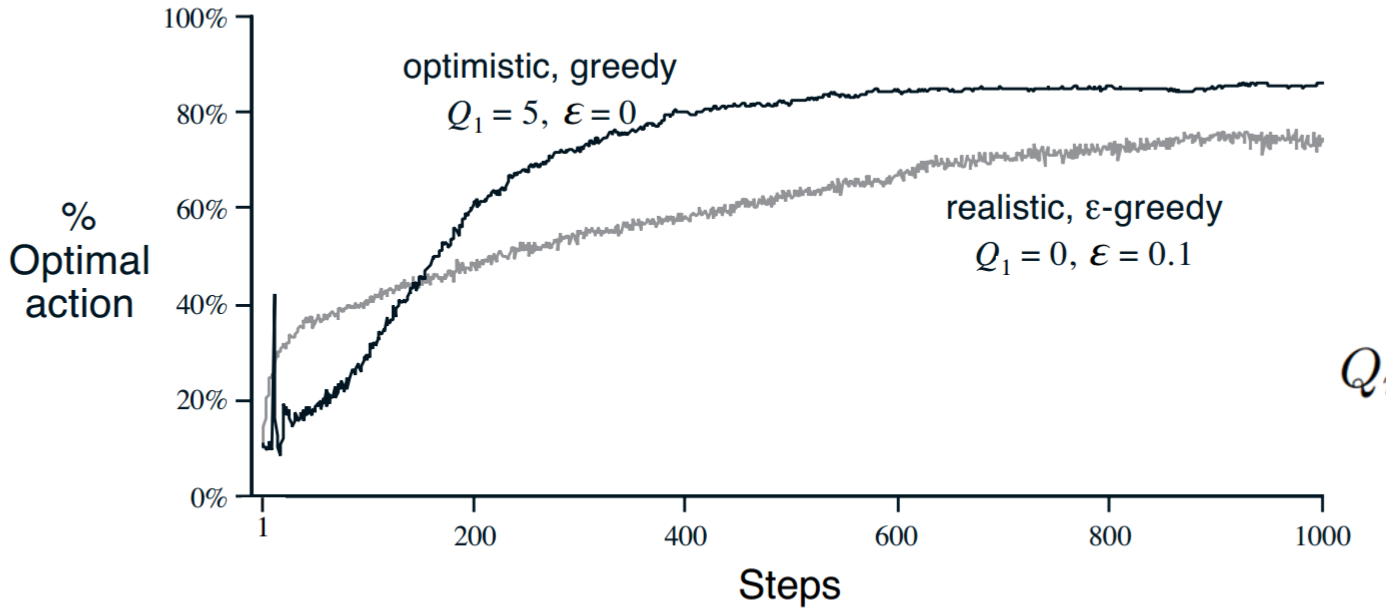
the steps become small enough to assure convergence

- $\alpha_n(a) = \frac{1}{n}$

$\alpha_n(a) = \alpha$

Optimistic Initial Values

- encouraging exploration (not a generally useful approach)
- a simple trick on stationary problems



$$Q_{n+1} \doteq Q_n + \alpha \Big[R_n - Q_n \Big]$$

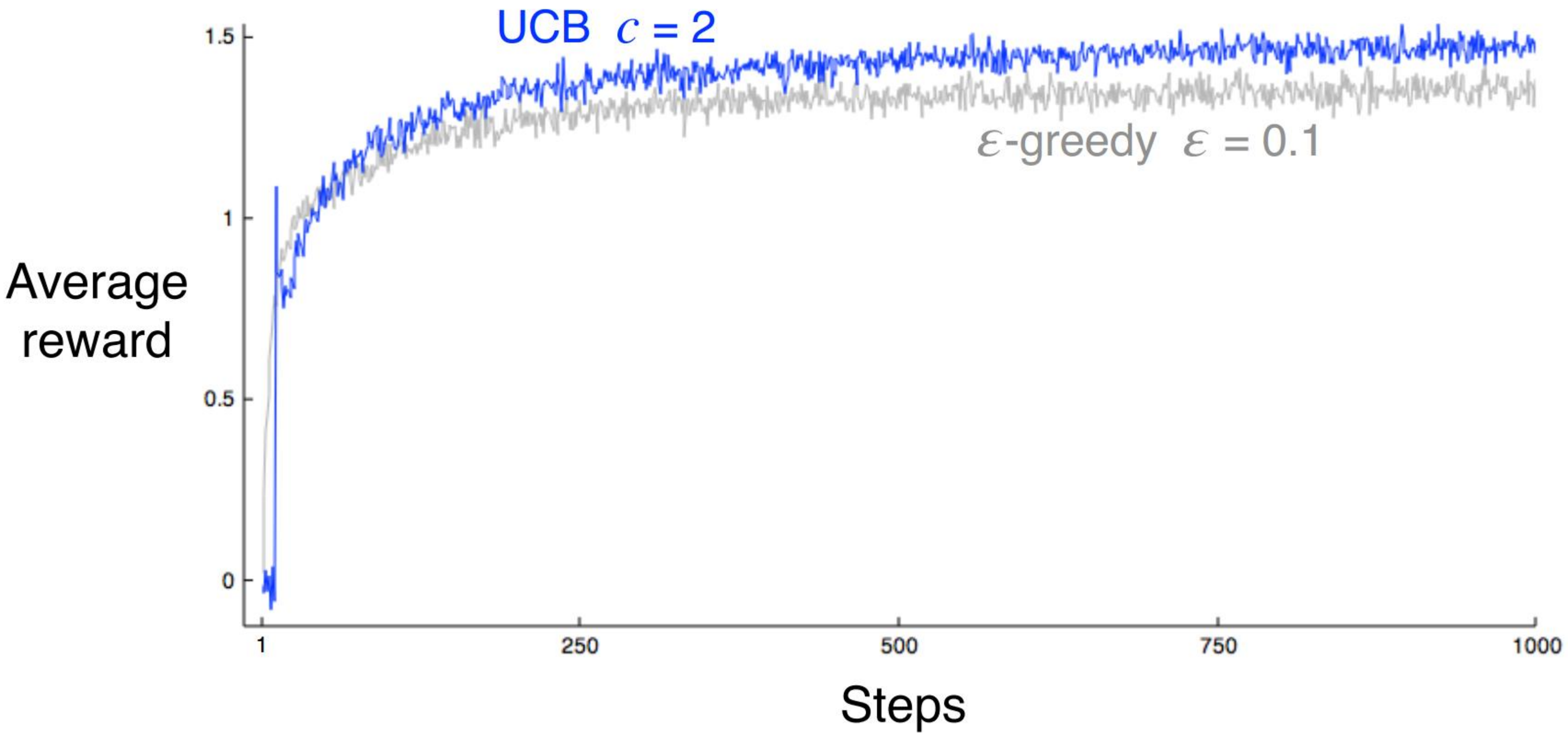
Upper-Confidence-Bound Action Selection

- $\varepsilon - greedy$ method's problem
- UCB Action Selection

$$A_t \doteq \underset{a}{\arg\max} \left[Q_t(a) + c\sqrt{\frac{\log t}{N_t(a)}} \right]$$

- Disadvantage：nonstationary problems

Upper-Confidence-Bound Action Selection （cont.）



Gradient Bandit Algorithms

- $H_t(a)$: numerical *preference* for each action a
- $\pi_t(a)$: the probability of taking action a at time t

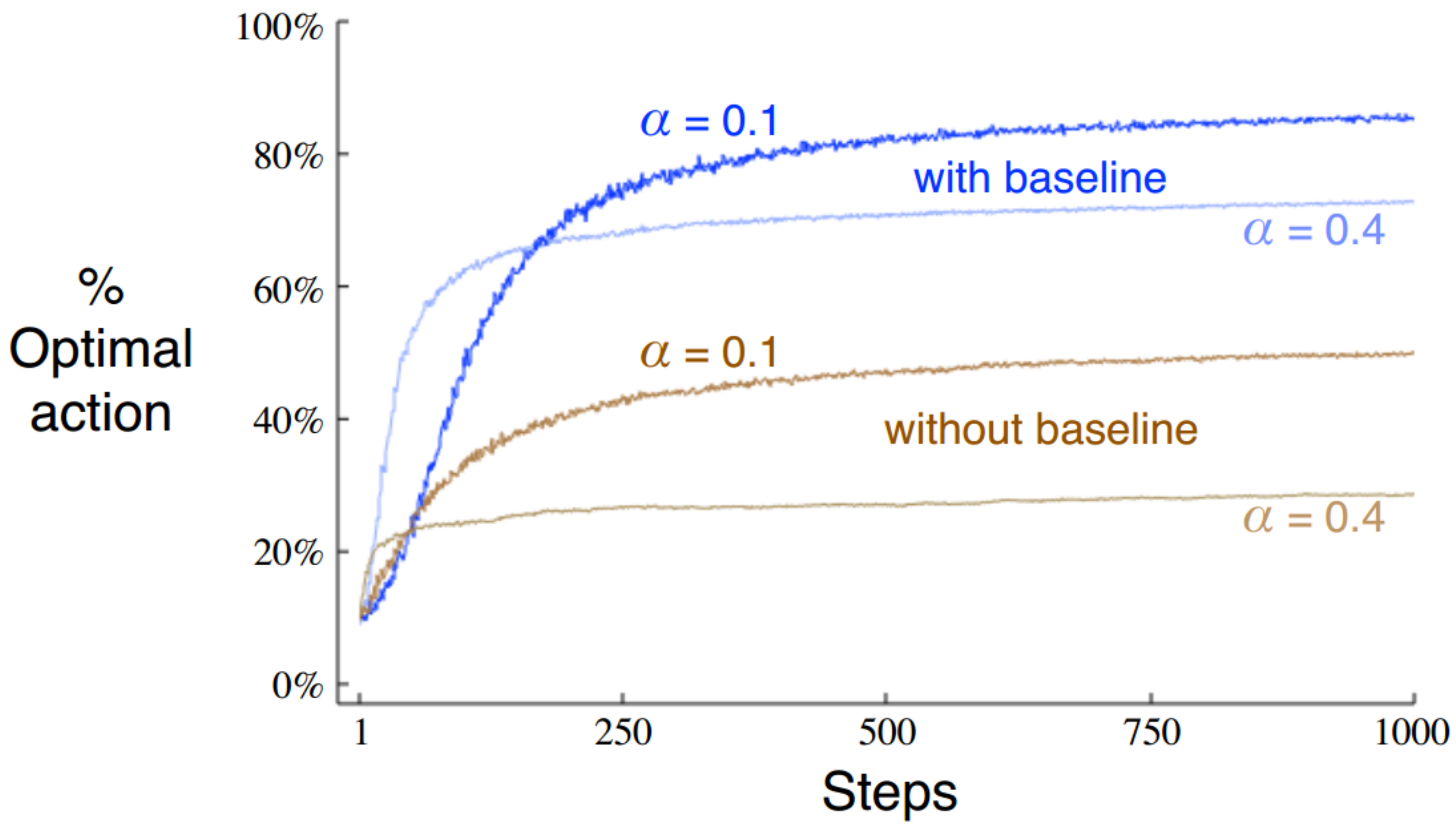
- $$\Pr\{A_t = a\} \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{k} e^{H_t(b)}} \doteq \pi_t(a)$$

- $$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big)$$

- $$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \quad \forall a \neq A_t$$

- $H_1(a) = 0, \forall a$

Gradient Bandit Algorithms (cont.)



mean of +4 instead of zero

Gradient Bandit Algorithms (cont.)

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

$$\mathbb{E}[R_t] \doteq \sum_b \pi_t(b) q_*(b)$$

$$\begin{aligned}
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \left[\sum_b \pi_t(b) q_*(b) \right] \\
&= \sum_b q_*(b) \frac{\partial \pi_t(b)}{\partial H_t(a)} \\
&= \sum_b \left(q_*(b) - X_t\right) \frac{\partial \pi_t(b)}{\partial H_t(a)}
\end{aligned}$$

X_t can be any scalar that does not depend on b

$\sum_b \frac{\partial \pi_t(b)}{\partial H_t(a)} = 0$

Gradient Bandit Algorithms (cont.)

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_b \pi_t(b)\big(q_*(b) - X_t\big)\frac{\partial \pi_t(b)}{\partial H_t(a)}/\pi_t(b)$$

$$= \mathbb{E}\left[\big(q_*(A_t) - X_t\big)\frac{\partial \pi_t(A_t)}{\partial H_t(a)}/\pi_t(A_t)\right]$$

$$= \mathbb{E}\left[\big(R_t - \bar{R}_t\big)\frac{\partial \pi_t(A_t)}{\partial H_t(a)}/\pi_t(A_t)\right],$$

$$\mathbb{E}[R_t|A_t] = q_*(A_t)$$

$$= \mathbb{E}\big[\big(R_t - \bar{R}_t\big)\pi_t(A_t)\big(\mathbf{1}_{a=A_t} - \pi_t(a)\big)/\pi_t(A_t)\big]$$

$$= \mathbb{E}\big[\big(R_t - \bar{R}_t\big)\big(\mathbf{1}_{a=A_t} - \pi_t(a)\big)\big].$$

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \pi_t(b)\big(\mathbf{1}_{a=b} - \pi_t(a)\big)$$

Gradient Bandit Algorithms (cont.)

$$H_{t+1}(a) \doteq H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)}$$

$$\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \mathbb{E}\Big[\big(R_t - \bar{R}_t\big)\big(\mathbf{1}_{a=A_t} - \pi_t(a)\big)\Big] .$$

$$H_{t+1}(a) = H_t(a) + \alpha\big(R_t - \bar{R}_t\big)\big(\mathbf{1}_{a=A_t} - \pi_t(a)\big), \qquad \forall a$$

Gradient Bandit Algorithms (cont.)

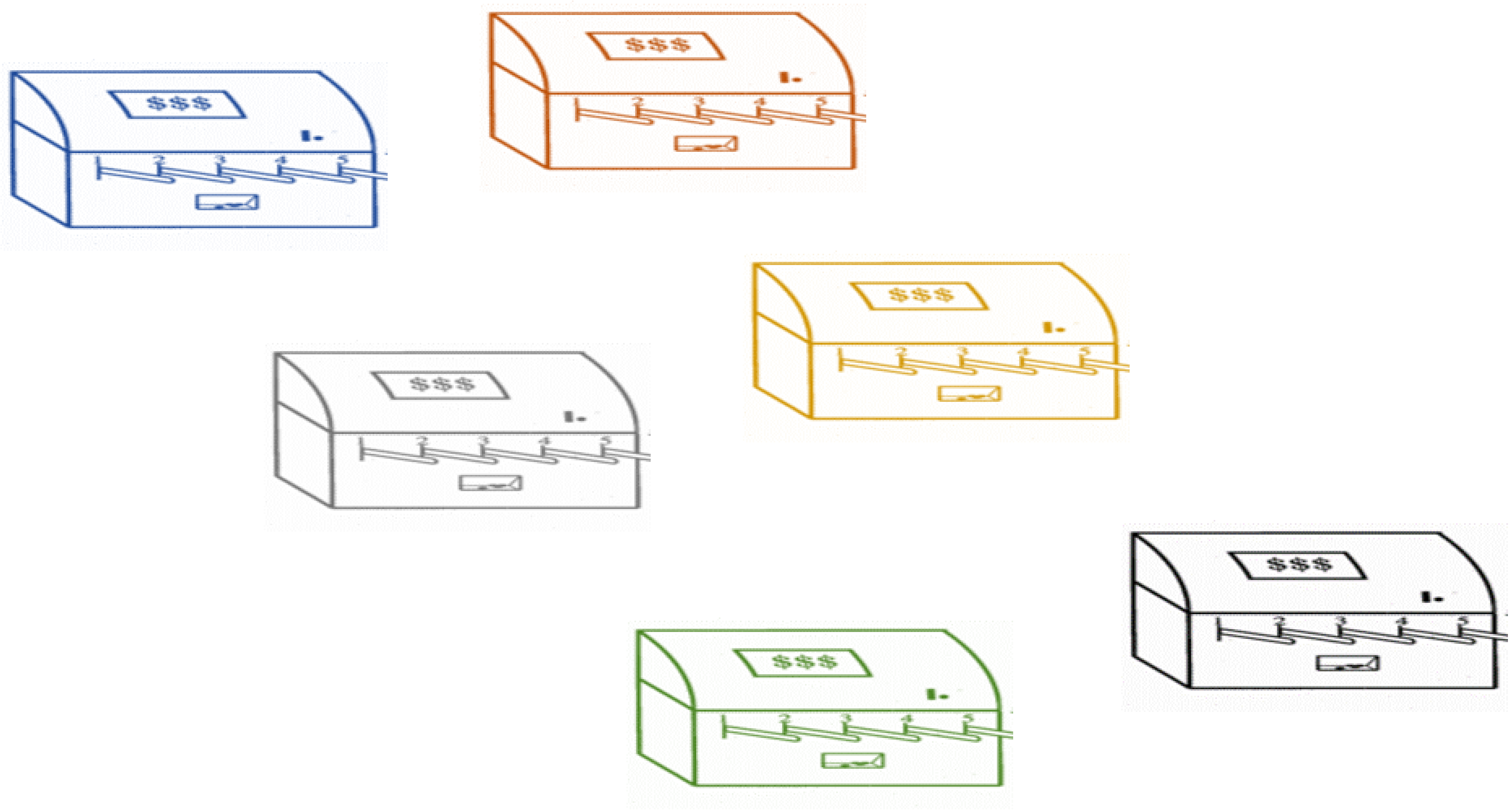
$$\frac{\partial}{\partial x}\left[\frac{f(x)}{g(x)}\right] = \frac{\frac{\partial f(x)}{\partial x} g(x) - f(x)\frac{\partial g(x)}{\partial x}}{g(x)^2}$$

$$\begin{aligned}
\frac{\partial \pi_t(b)}{\partial H_t(a)} &= \frac{\partial}{\partial H_t(a)} \pi_t(b) \\
&= \frac{\partial}{\partial H_t(a)}\left[\frac{e^{H_t(b)}}{\sum_{c=1}^{k} e^{H_t(c)}}\right] \\
&= \frac{\frac{\partial e^{H_t(b)}}{\partial H_t(a)} \sum_{c=1}^{k} e^{H_t(c)} - e^{H_t(b)} \frac{\partial \sum_{c=1}^{k} e^{H_t(c)}}{\partial H_t(a)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)^2} \\
&= \frac{\mathbf{1}_{a=b} e^{H_t(a)} \sum_{c=1}^{k} e^{H_t(c)} - e^{H_t(b)} e^{H_t(a)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)^2}
\end{aligned}$$

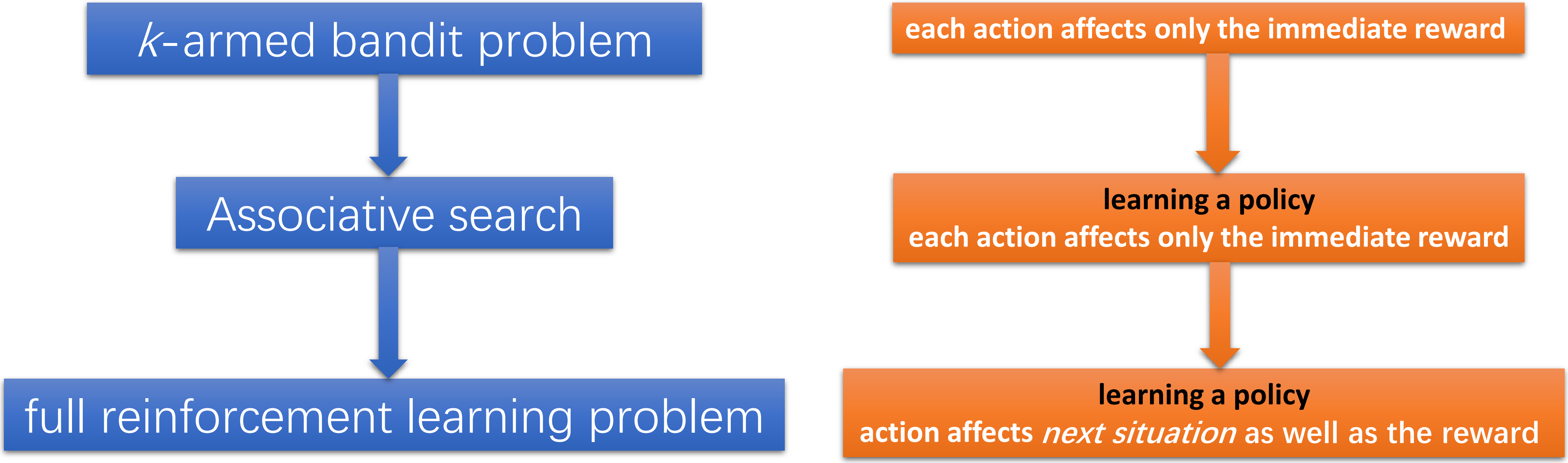
$$\begin{aligned}
&= \frac{\mathbf{1}_{a=b} e^{H_t(b)}}{\sum_{c=1}^{k} e^{H_t(c)}} - \frac{e^{H_t(b)} e^{H_t(a)}}{\left(\sum_{c=1}^{k} e^{H_t(c)}\right)^2} \\
&= \mathbf{1}_{a=b} \pi_t(b) - \pi_t(b)\pi_t(a) \\
&= \pi_t(b)\big(\mathbf{1}_{a=b} - \pi_t(a)\big).
\end{aligned}$$

$$\frac{\partial \pi_t(b)}{\partial H_t(a)} = \pi_t(b)\big(\mathbf{1}_{a=b} - \pi_t(a)\big)$$

Associative Search (Contextual Bandits)

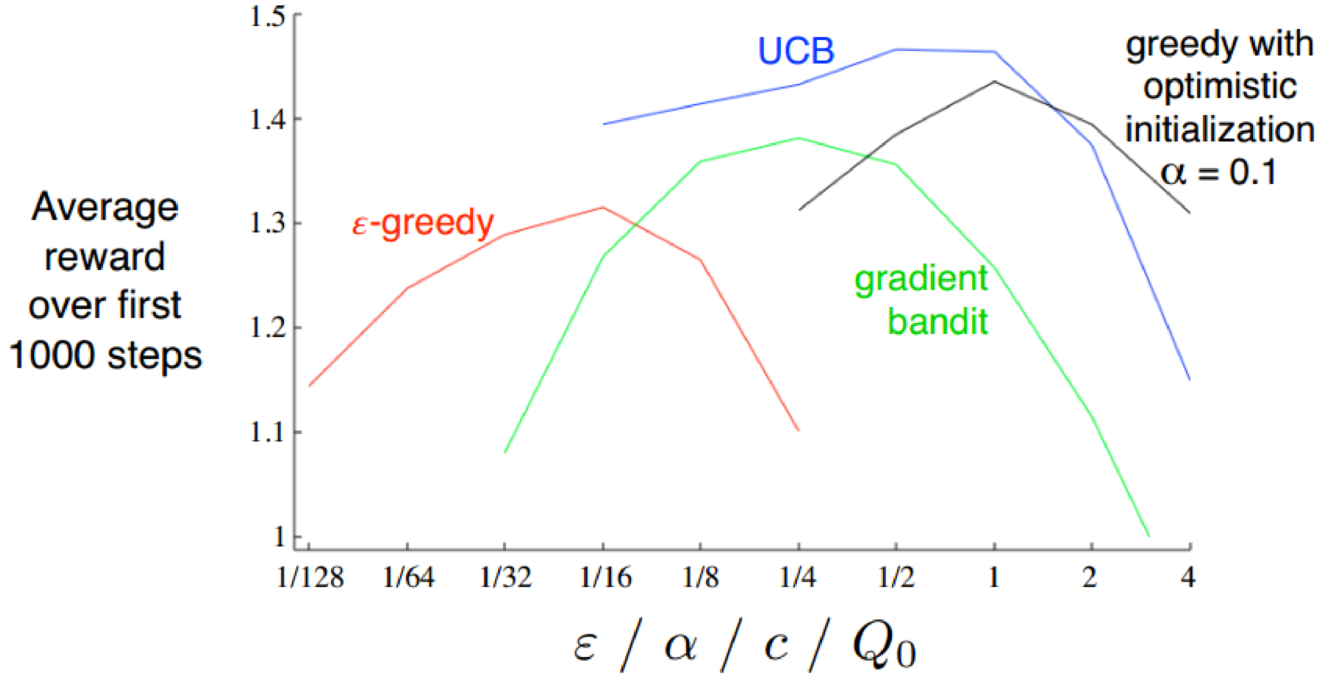


Associative Search/Contextual Bandits (cont.)



Summary

- ε −greedy methods
- Upper-Confidence-Bound Action Selection
- Optimistic Initial Values
- Gradient Bandit Algorithms



Thanks