به نام خدا

Iran University
of Science and
Technology

یادگیری تقویتی در کنترل

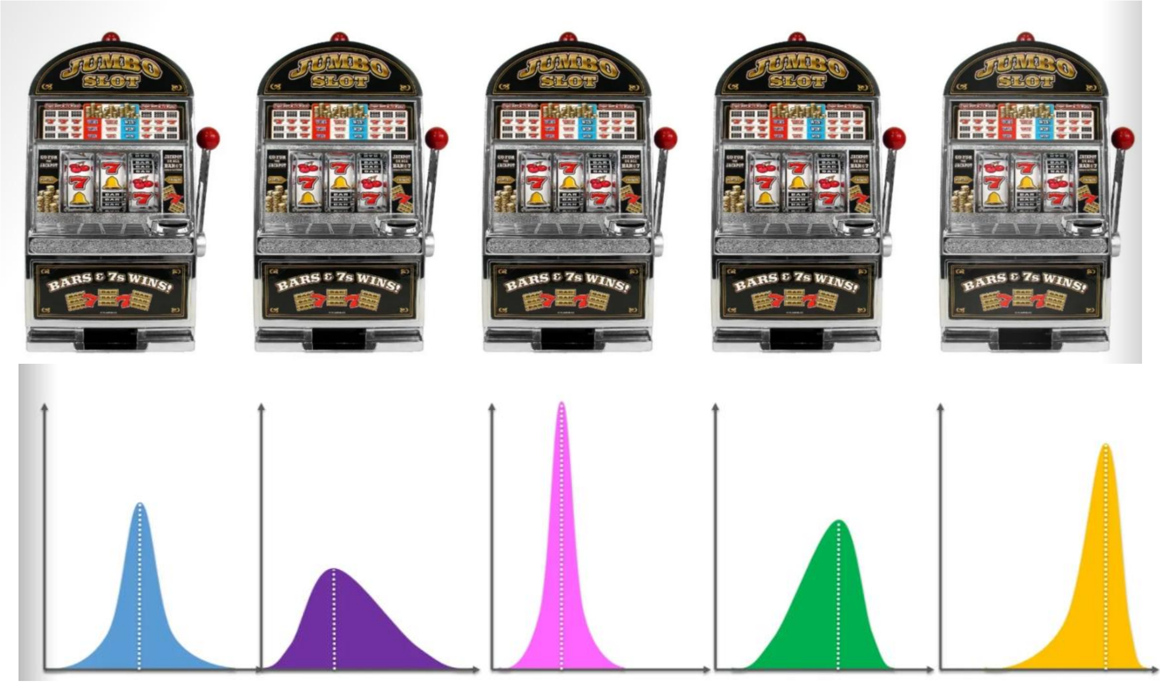
دکتر سعید شمقدری

دانشکده مهندسی برق
گروه کنترل

نیمسال اول ۱۴۰۵–۱۴۰۴

Multi-Armed Bandit

Multi-Armed Bandits مسئله



Multi-Armed Bandits مسئله

Action: $a_t \in A$

Reward: $r_t \sim R \quad R_a(r) = P(r|a)$

Goal: maximize $\Sigma\, r_t$

Value: $Q(a) \approx E[r|a]$

Exact Value: $q_*(a) = E[r|a]$

Optimal Value: $\max_{a \in A} Q(a)$

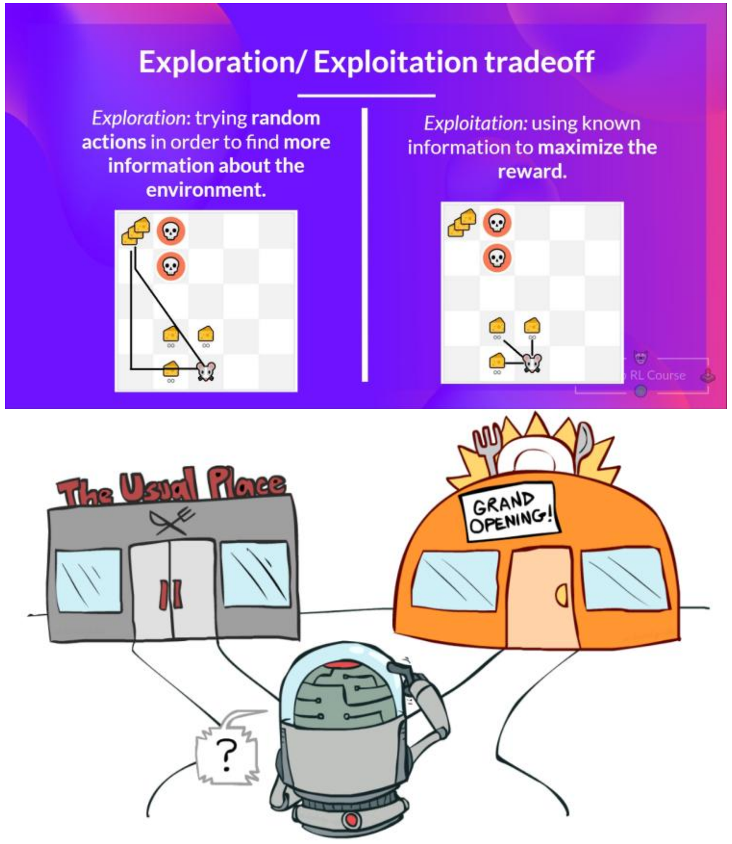
Greedy and ϵ-Greedy Actions

Greedy Action: Exploitation only
Exploration: Selecting a Non-Greedy Action

Exploitation or Exploration??

انتخاب Action:

Greedy Action: $$A_t = \arg\max Q_t(a)$$
ϵ-Greedy Action



یادآوری از Lecture 1:
Exploration در مقابل Exploitation

یک روش تخمین تابع Value

Sample Average:

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

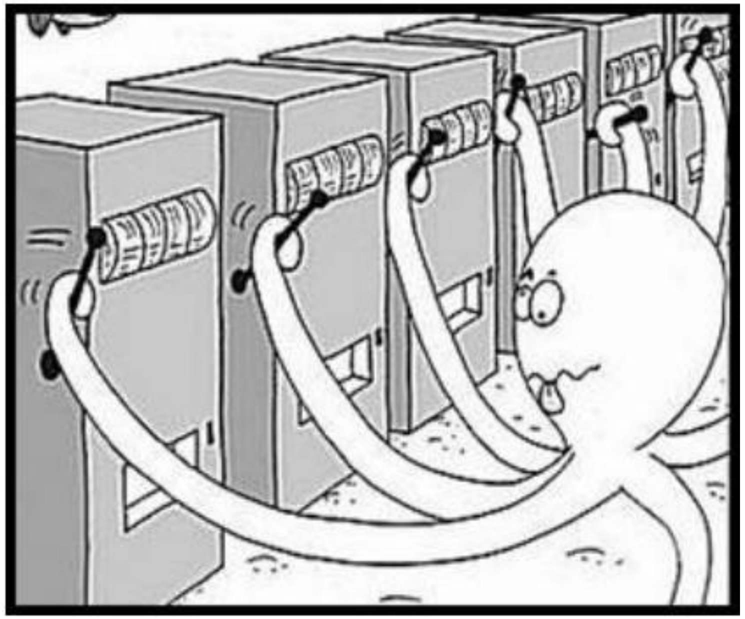
Convergence: (by the law of large numbers…)

$$Q_t(a) \rightarrow q_*(a) \qquad \text{for all } a \in A$$

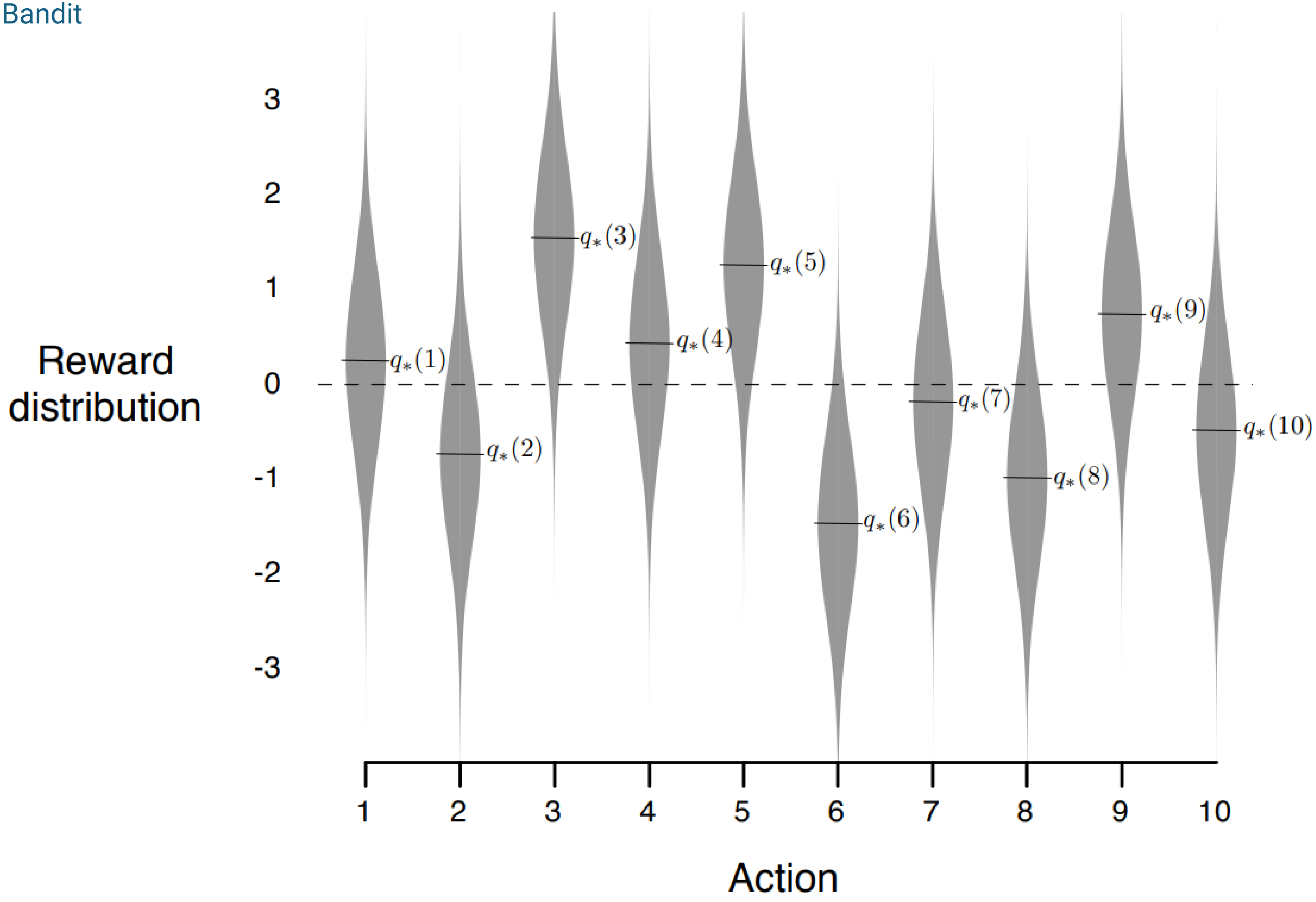
Q: در شرایط همگرایی برای روش ϵ-Greedy Action، احتمال انتخاب Action بهینه؟

The 10-Armed Testbed

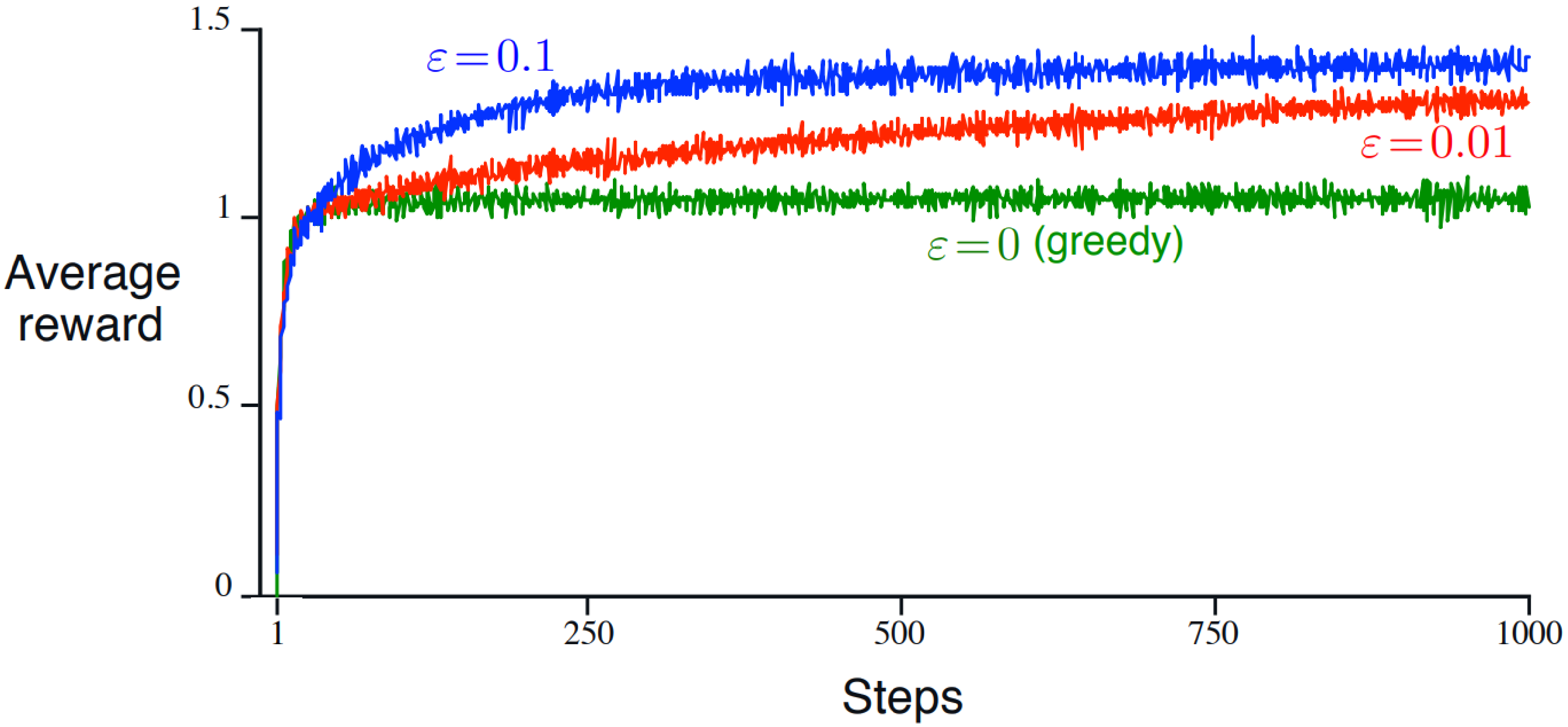
One Run? 1000 sample
Average Behavior?
Average in 2000 problem solution

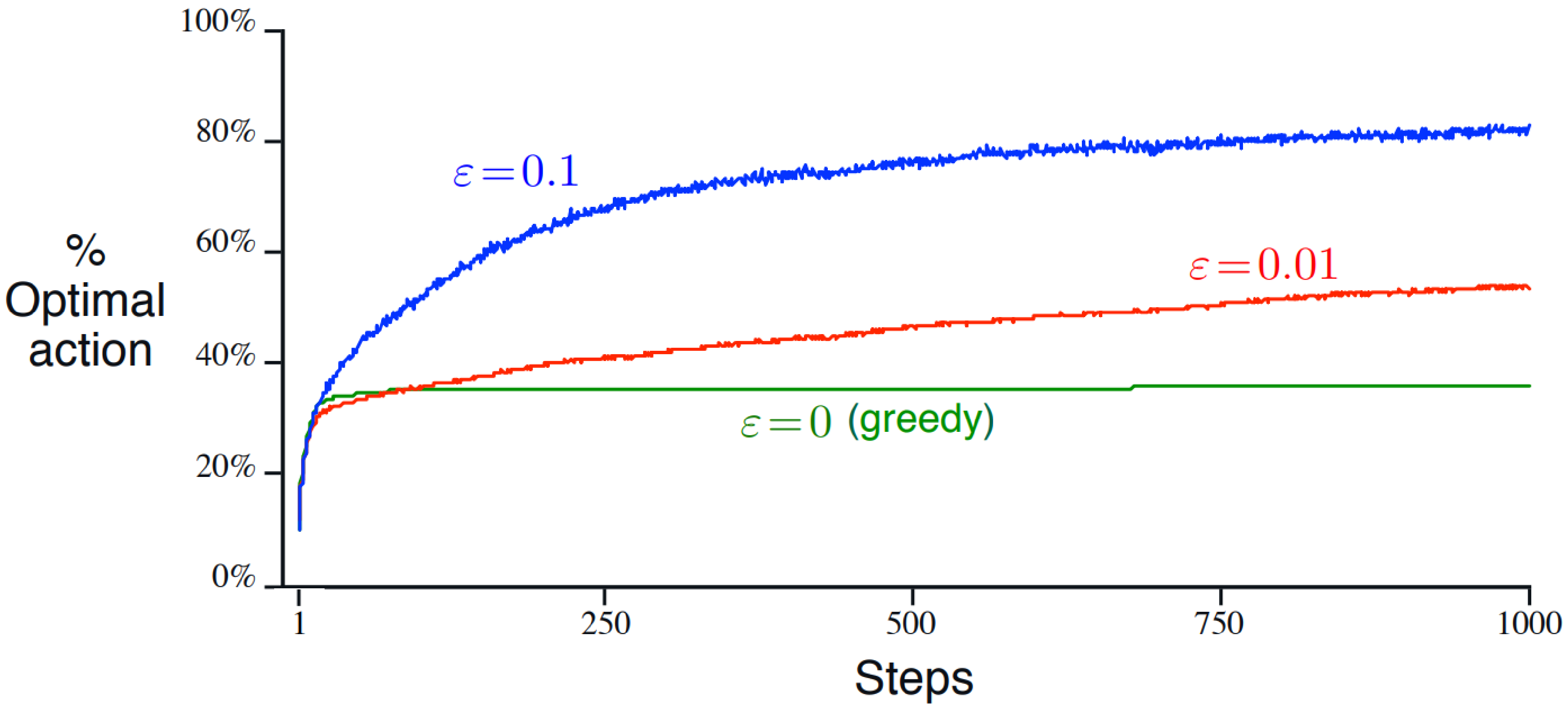


تولید مقادیر $q_*(a)$ برای هر a : توزیع نرمال با میانگین صفر و واریانس ۱
مقادیر پاداش برای هر a: توزیع نرمال با میانگین $q_*(a)$ و واریانس ۱



Reward distribution
3
2
1
0
-1
-2
-3
$q_*(1)$
$q_*(2)$
$q_*(3)$
$q_*(4)$
$q_*(5)$
$q_*(6)$
$q_*(7)$
$q_*(8)$
$q_*(9)$
$q_*(10)$
1
2
3
4
5
6
7
8
9
10
Action


1.5
1
0.5
0
Average reward
$\varepsilon = 0.1$
$\varepsilon = 0.01$
$\varepsilon = 0$ (greedy)
1
250
500
750
1000
Steps


100%
80%
60%
40%
20%
0%
% Optimal action
ε = 0.1
ε = 0.01
ε = 0 (greedy)
1
250
500
750
1000
Steps

Incremental Implementation

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

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

Incremental Implementation

$$NewEstimate \leftarrow OldEstimate + StepSize \left[\underbrace{Target - OldEstimate}_{\text{Error}} \right]$$

α (متناسب با exploration)

حل مسئله RL برای فرایند غیرایستا

وزن بیشتر به پاداش های اخیر

استفاده از Weighted Average بجای Sample Average:

فرض مقدار ثابت برای $\alpha \in (0,1]$:

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

حل مسئله RL برای فرایند غیرایستا

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

حل مسئله RL برای فرایند **غیرایستا**

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

Q: تاثیر افزایش α در واریانس تخمین؟؟

Weighted Average

Unbiased Estimation: $if\ n \to \infty \quad then\ Q = E[R(a)] = q_*(a)$

حل مسئله RL برای فرایند **غیرایستا**

Weighted Average با α متغیر با زمان : $\alpha_n(a)$

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

عدم کوچک شدن سری α_n ها
کاهش اثر شرط اولیه و نوسانات
Exploration کافی

شرط همگرایی

در حالت Sample Average:

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

توجه: برای هر ضریب ثابت برقرار نیست!

حل مسئله RL برای فرایند **غیرایستا**

در حالت Sample Average:

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

توجه: برای هر ضریب ثابت برقرار نیست!

$$1 + \frac{1}{2} + \overbrace{\frac{1}{3} + \frac{1}{4}} + \ \ldots \ \geq 1 + \frac{1}{2} + \overbrace{\frac{1}{4} + \frac{1}{4}} + \underbrace{\left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right)} + \cdots + (\frac{1}{16} + \cdots$$

$$\geq 1 + 1 + 1 + \cdots = \infty$$

حل مسئله RL برای فرایند غیرایستا

تخمینگر Unbiased:

$$Q_{n+1} = Q_n + \alpha_n[R_n - Q_n]$$

$$Q_n = Q_{n-1} + \alpha_{n-1}[R_{n-1} - Q_{n-1}]$$

$$= (1-\alpha_{n-1})Q_{n-1} + \alpha_{n-1}R_{n-1}$$

$$Q_{n+1} = \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\,Q_{n-1}$$

$$= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\cdot\alpha_{n-2}R_{n-2} + \cdots$$

$$q_* = E[R](\alpha_n + (1-\alpha_n)\alpha_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\,\alpha_{n-2} + \cdots + (1-\alpha_n)(1-\alpha_{n-1})\cdots(1-\alpha_1))$$

$$q_* = E[R](\alpha_n + (1-\alpha_n)\underbrace{[\alpha_{n-1} + (1-\alpha_{n-1})\alpha_{n-2} + \cdots]}_{S})$$

حل مسئله RL برای فرایند غیرایستا

تخمینگر Unbiased:

$$q_* = E[R]\big(\alpha_n + (1-\alpha_n)\alpha_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\,\alpha_{n-2} + \cdots + (1-\alpha_n)(1-\alpha_{n-1})\cdots(1-\alpha_1)\big)$$

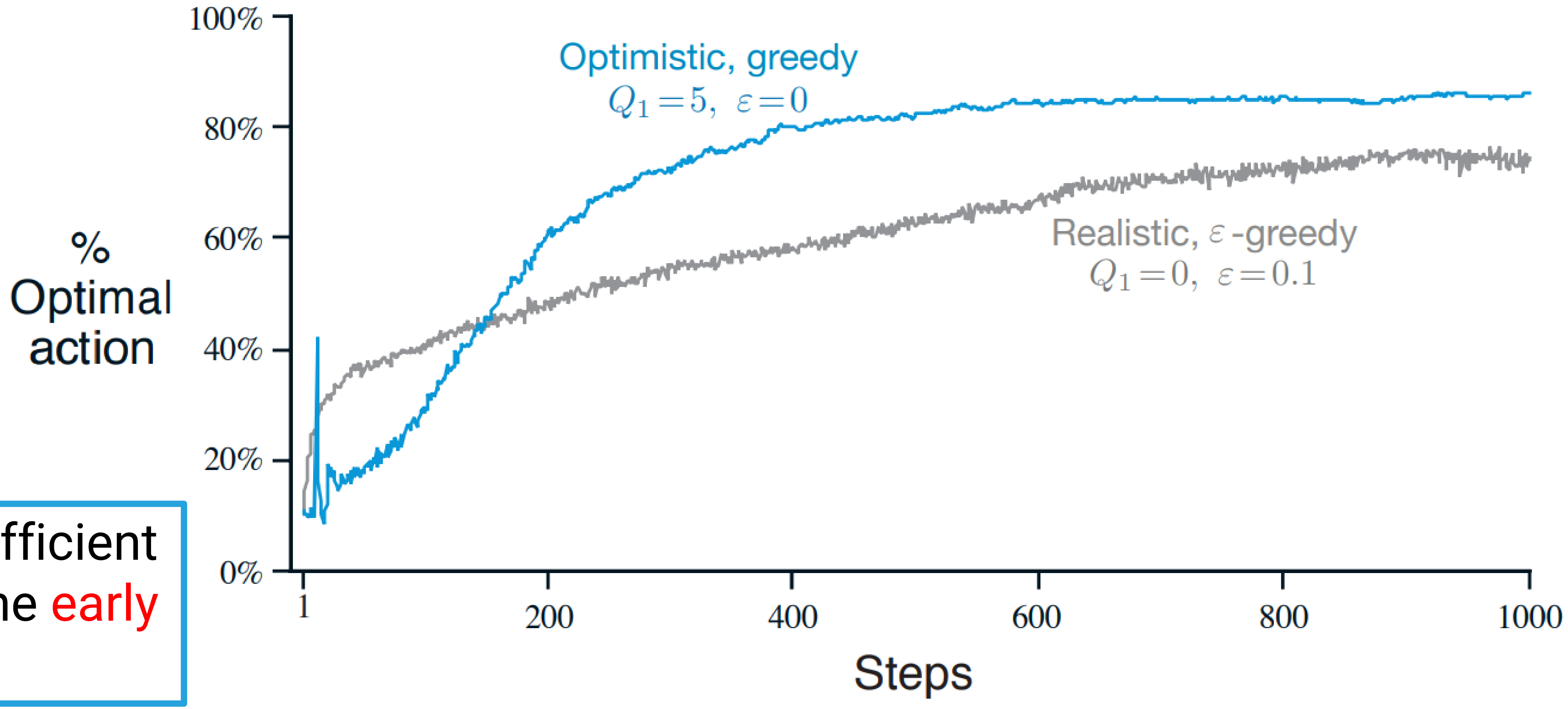
$$q_* = E[R]\big(\alpha_n + (1-\alpha_n)\underbrace{[\alpha_{n-1} + (1-\alpha_{n-1})\alpha_{n-2} + \cdots]}_{S}\big)$$

$$S = \alpha_n + (1-\alpha_n)S$$

$$(1+\alpha_n)S = S + \alpha_n \rightarrow S = 1$$

انتخاب Optimistic مقادیر اولیه

مثال: $Q_1 = 5$



The need for sufficient exploration in the early stages.

Upper Confidence Bound

عدم قطعیت در تخمین Value: نیاز به exploration

Q: ضعف روش ϵ-Greedy؟

در روش ε-greedy اکتشاف از بین عمل‌های غیرحریصانه به صورت تصادفی انجام می‌شود.

مناسب است که میزان **احتمال بهینه بودن** عمل‌های غیرحریصانه نیز در explore لحاظ شود.

Upper Confidence Bound

لحاظ کردن عدم قطعیت تخمین!

$$A_t \doteq \underset{a}{\arg\max}\left[Q_t(a) + c\overbrace{\sqrt{\frac{\ln t}{N_t(a)}}}^{\text{میزان عدم قطعیت}}\right]$$

کنترل میزان exploration

(if zero?) دفعاتی که قبل از t اکشن a انتخاب شده است

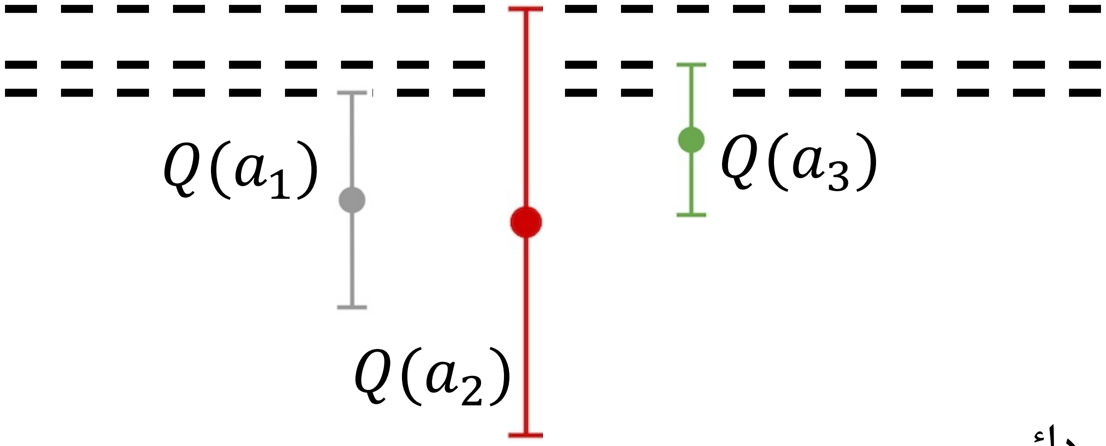
هر بار انتخاب a؟ .... کاهش عدم قطعیت

هر بارانتخاب نشدن a؟ .... افزایش عدم قطعیت

Upper Confidence Bound

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

تخمین نقطه ای ← تخمین بازه ای



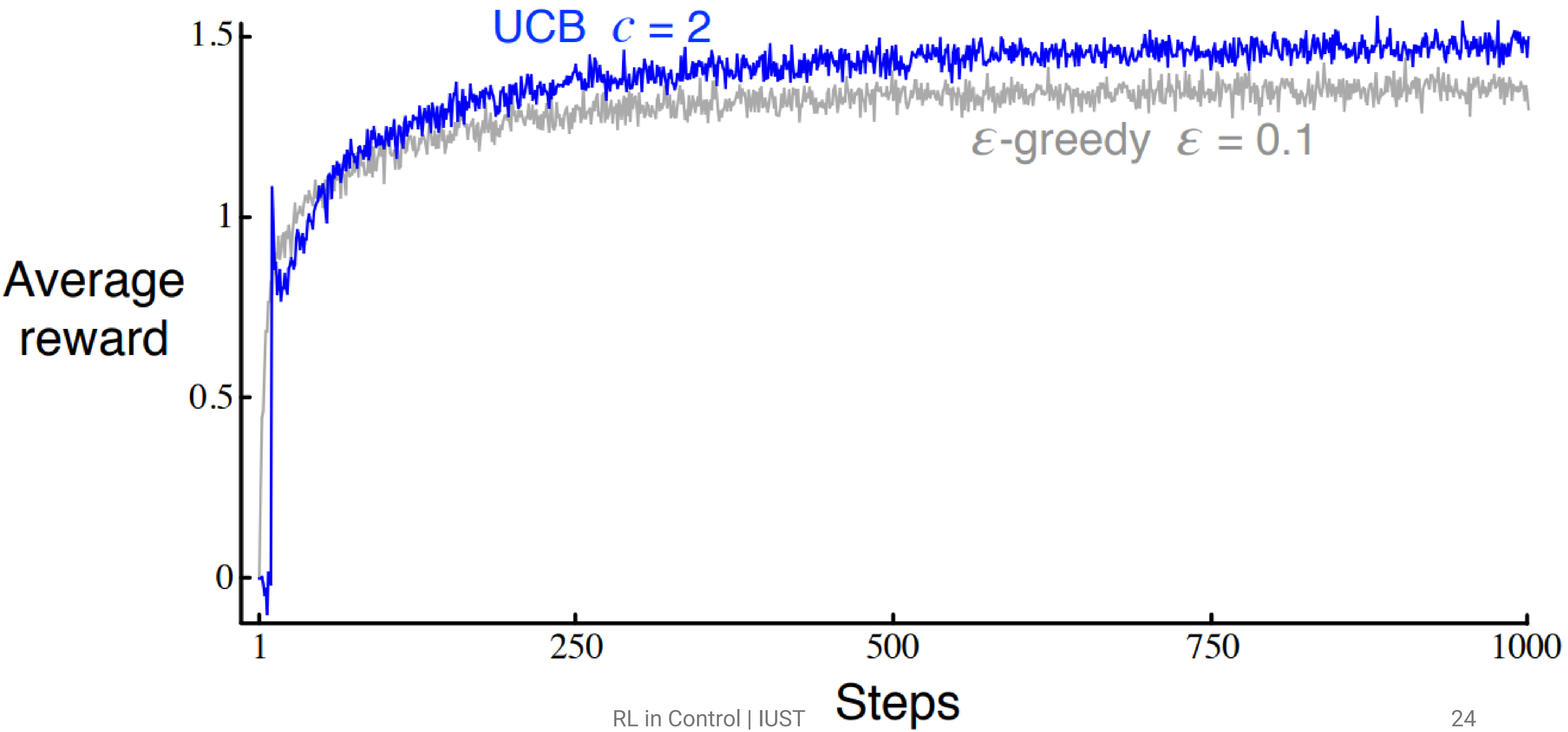
انتخاب تمام اکشن ها در نهایت

انتخاب با تکرار کمتر برای اکشن های با value کم

Exploration واقع بینانه تر

انجام اتوماتیک Exploration و Exploitation به صورت دائم

Upper Confidence Bound



الگوریتم Gradient Bandit

استفاده از یک معیار preference بجای تخمین action value: $H_t(a)$

احتمال انتخاب a مطابق با توزیع soft-max:

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

الگوریتم Gradient Bandit

$H_1(a) = 0 \quad for\ all\ a \in \mathcal{A}$:t=1

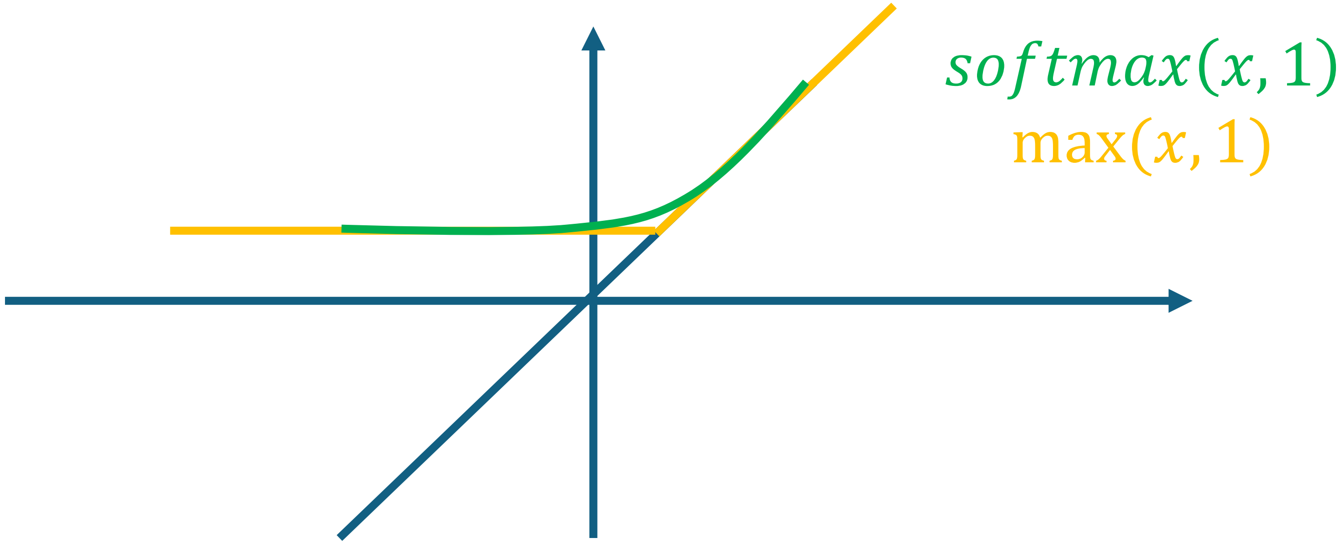
بروزرسانی در t>1:

$$H_{t+1}(A_t) \doteq H_t(A_t) + \alpha\big(R_t - \bar{R}_t\big)\big(1 - \pi_t(A_t)\big), \qquad \text{and}$$
$$H_{t+1}(a) \doteq H_t(a) - \alpha\big(R_t - \bar{R}_t\big)\pi_t(a), \qquad \text{for all } a \neq A_t$$

بروز شدن H_t اکشنی که انتخاب نشده است

متوسط پاداش ها تا زمان t (شامل t)

Baseline $\begin{cases} R_t > \bar{R}_t & \text{افزایش احتمال انتخاب At در آینده} \\ R_t < \bar{R}_t & \text{کاهش احتمال انتخاب At در آینده} \end{cases}$

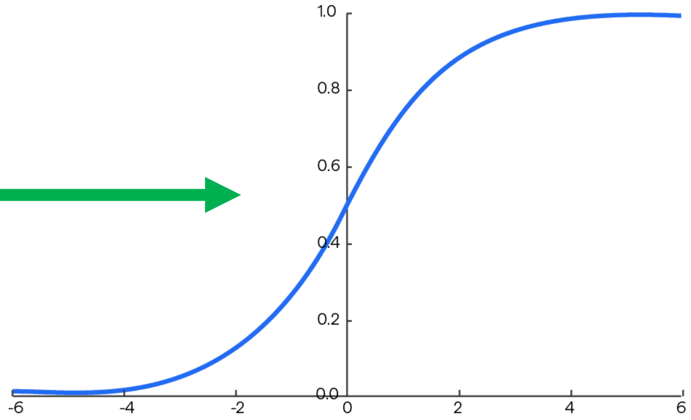


توزیع Soft-max

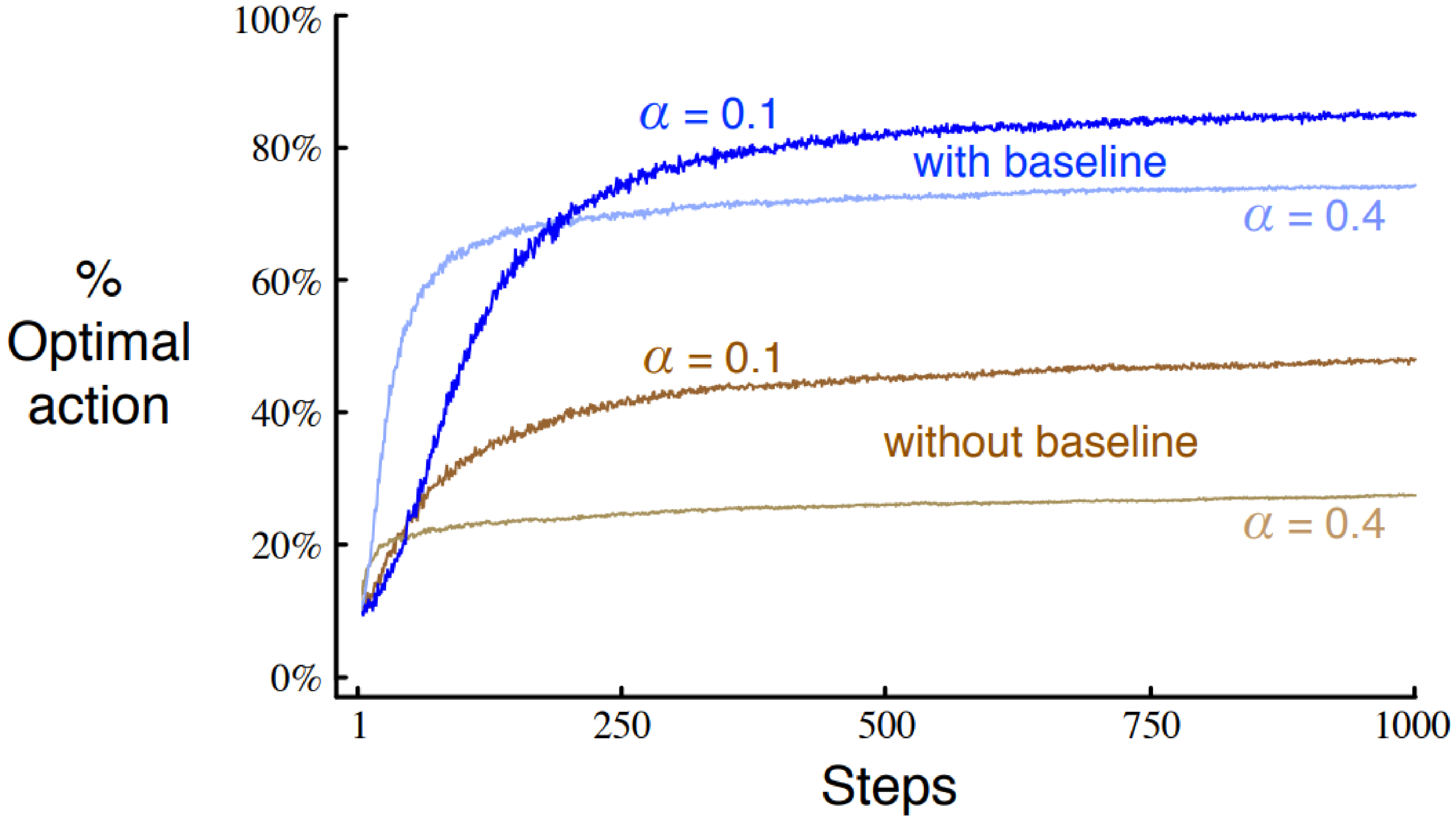
$$softmax(x_1, \dots, x_n) = log \sum_{i=1}^{n} e^{x_i}$$

در یادگیری ماشین و شبکه‌های عصبی:
برای Classification در طبقه‌ی خروجی مقادیر عددی تولید می‌شود. احتمال اینکه ورودی شبکه به کدام کلاس متعلق است از رابطه‌ی زیر بدست می‌آید.

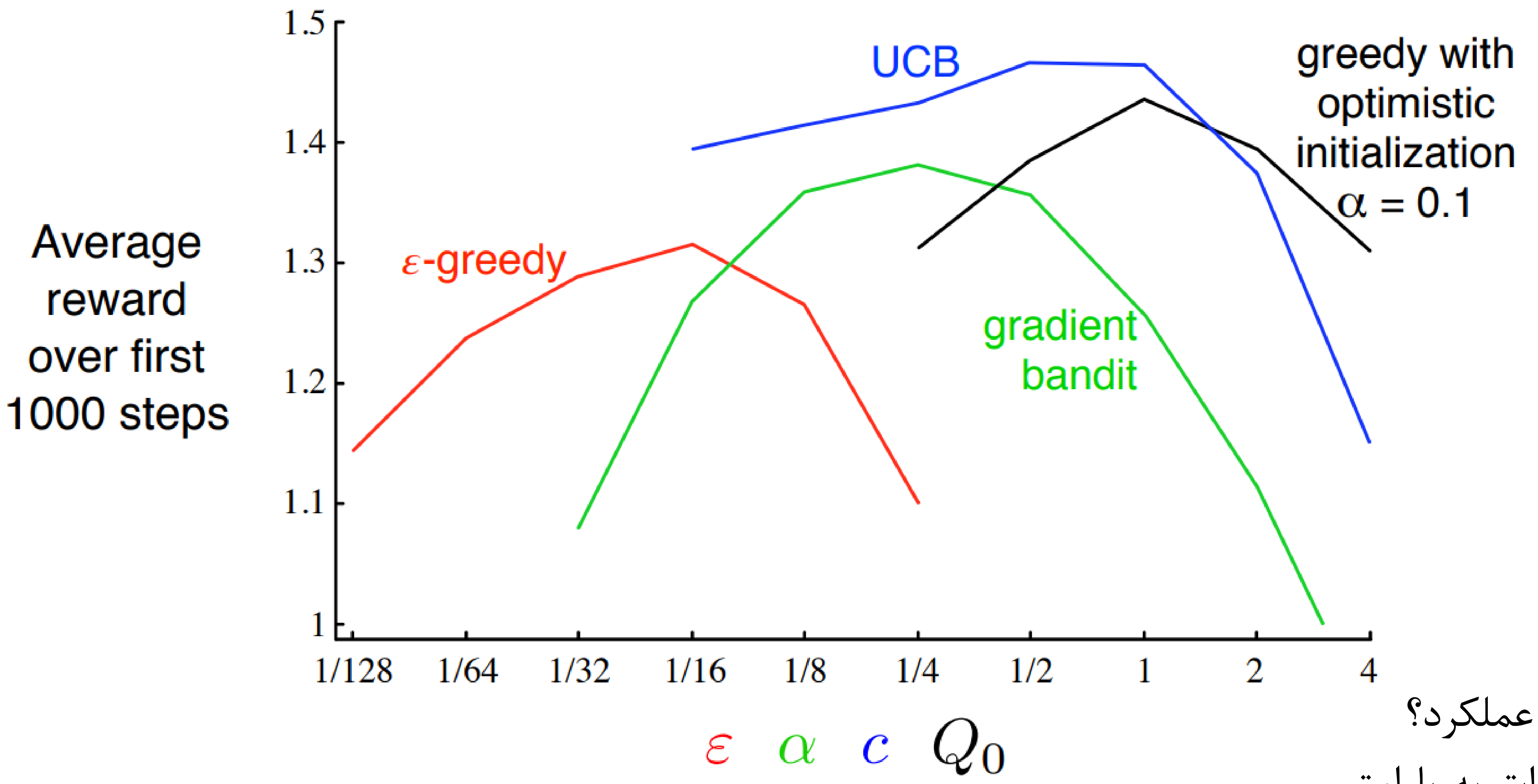
$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$



الگوریتم Gradient Bandit



مقایسه



Q: ⬅️ محدوده مناسب عملکرد؟

توجه به حساسیت تغییرات به پارامتر