Lecture -02

Immediate Reinforcement Learning :-

Key Characteristics of an RL Problem :-

1. learn to act in many situations.
2. Delayed rewards and credit assignment.
3. Exploration / Exploitation dilemma.

Two types of feedback

- Instructive feedback
- Evaluative feedback.

Instructive feedback	Evaluative feedback.
→ instruct the right action a^*.	→ evaluates the action taken A_t by giving some reward.
→ ignores the action taken.	→ Completely depends on the A_t.
→ SL	→ RL "Exploration Vs. Exploitation"

Multi-armed bandits :-

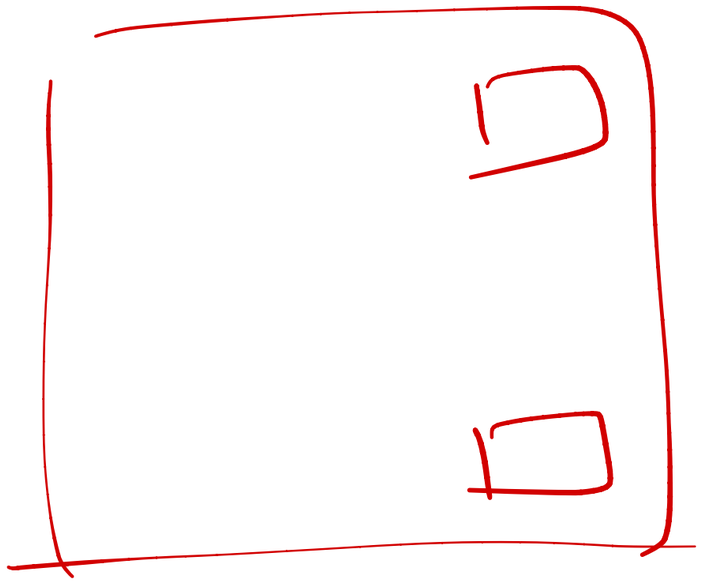
{ - Same state for all 'T'.
- rewards are immediate

"T"

Obj: to maximize the expected total reward over some time period T.

Applications :-

- Medical trials.
- A/B testing
- Ads. [contextual bandits]



Multi-armed bandits :-

A_1 : 1

A_2 : 5

A_3 : 0

A_4 : 0

A_5 : 10

A_6 : 1

- Reward distributions are Stationary.

---

1 2 - - - k

$q_*(1)$ $q_*(2)$ - - - $q_*(k)$

Value of action a = $q_*(a) = \mathbb{E}[R_t \mid A_t = a]$

optimal action $a^* = \underset{a}{\operatorname{argmax}}\ q_*(a)$

$a \in \{a_1, a_2, \ldots a_k\}$

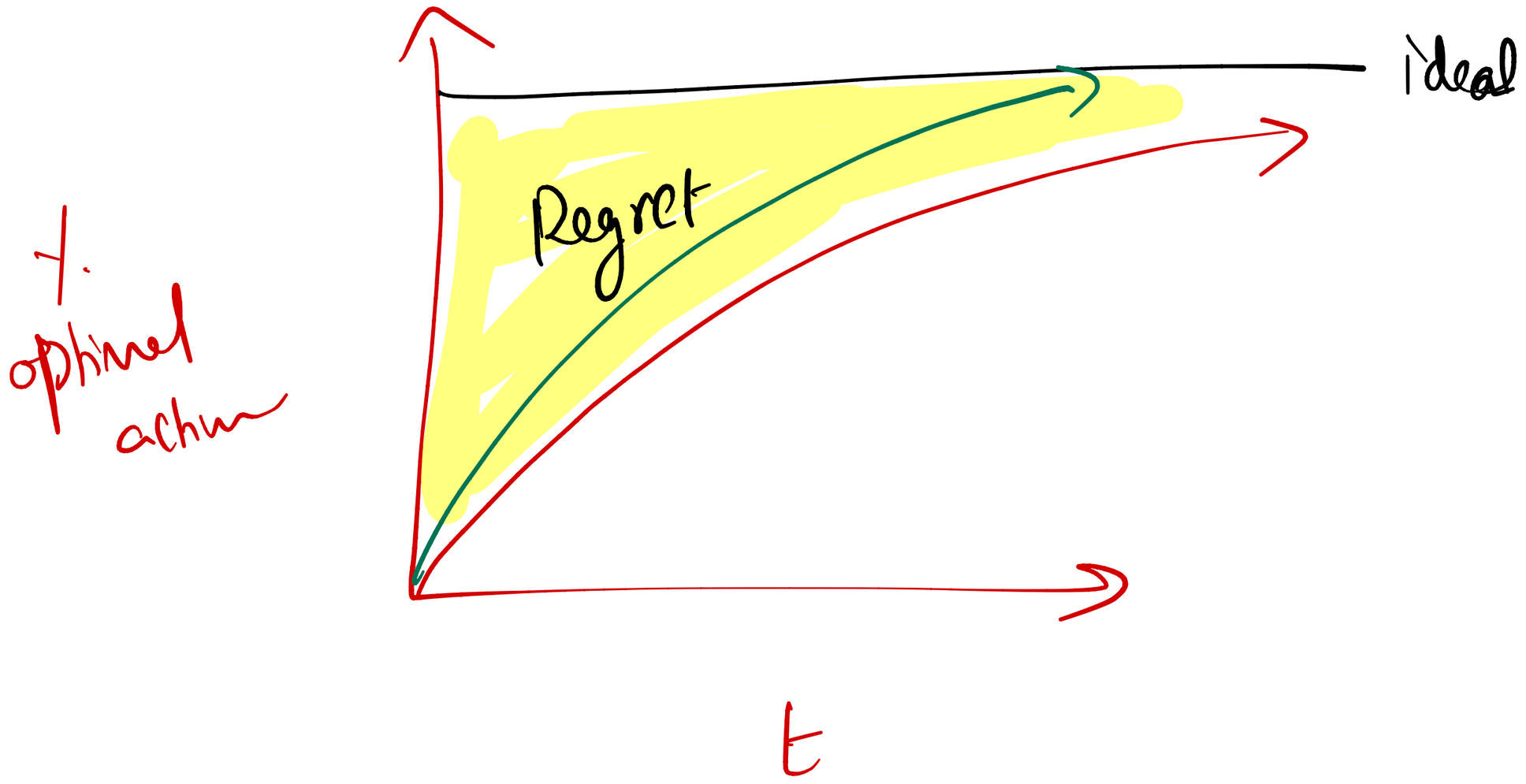
$Q_t(a)$ = estimated value of action 'a' at time 't'.

$q_*(a)$ = true expected reward.

Regret after 'k' steps!

$$\text{regret} = k\, q_*(a^*) - \underbrace{\sum_{t=1}^{k} R_t}$$

100 — 70


Ideal
Regret
t
optimal
achiev

Action-value methods:-

$Q_t(a)$ = Sum of rewards when 'a' was taken prior to 't'.

$\mathbb{1}_{A_i=a} = \begin{cases} 1 & \text{if 'a' was taken at } i^{th} \text{ step} \\ 0 & \text{o.w} \end{cases}$

# of times 'a' was taken prior to 't'.

$$= \frac{\sum_{i=1}^{t-1} R_i \, \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

$Q_t(a) \longrightarrow q_*(a)$ as $t \to \infty$.

Alg 1: Explore then Commit (ETC):

"m"

---

for each arm

pull that arm `m` times

Q(.) = Average of `m` rewards.

best action = $\underset{a}{\text{argmax}}\ Q(a)$

pull the best arm for the rest of the trials.

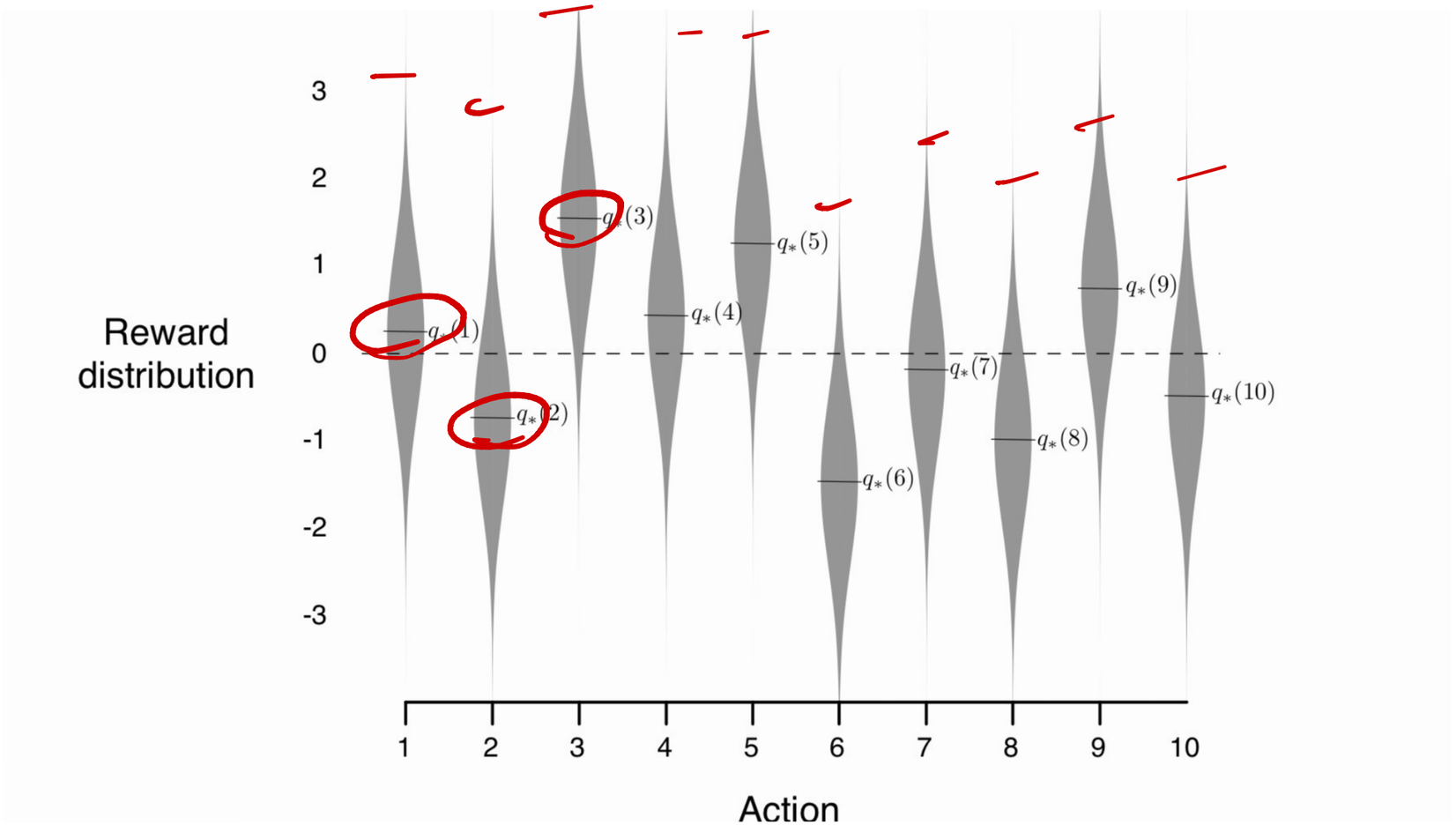
Alg 2: ϵ-greedy

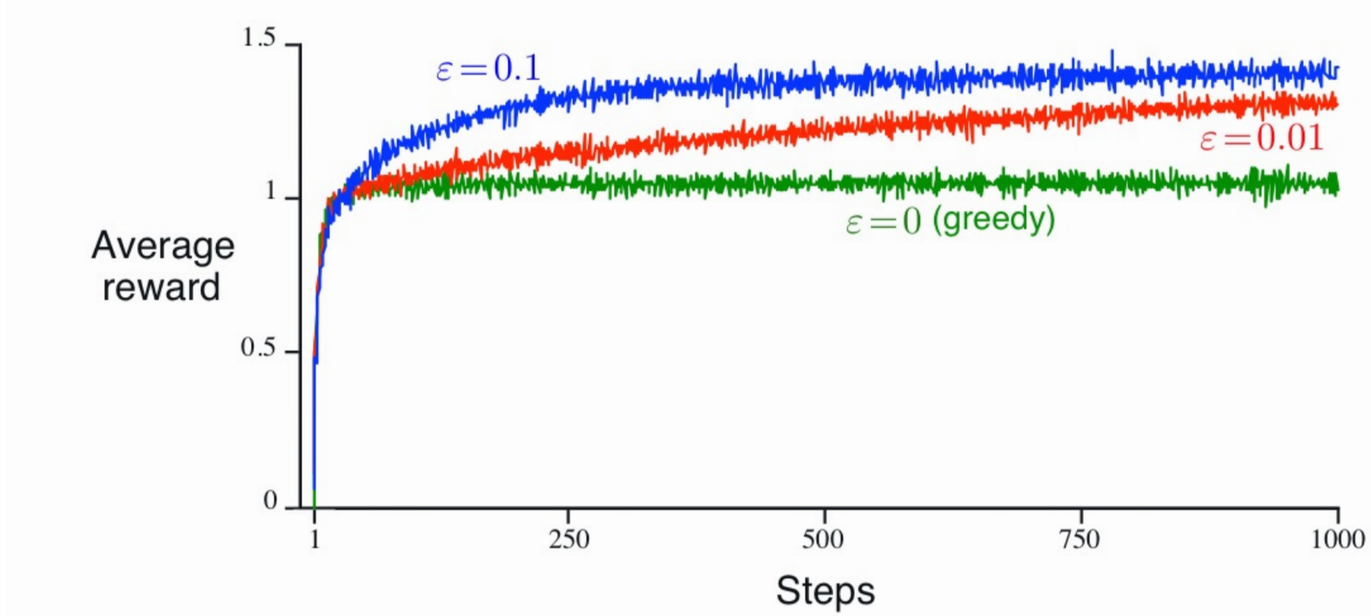
$$A_t = \begin{cases} \underset{a}{\text{argmax}}\ Q_t(a) & \text{w.p. } 1-\epsilon \\ \text{random action} & \text{w.p. } \epsilon \end{cases}$$

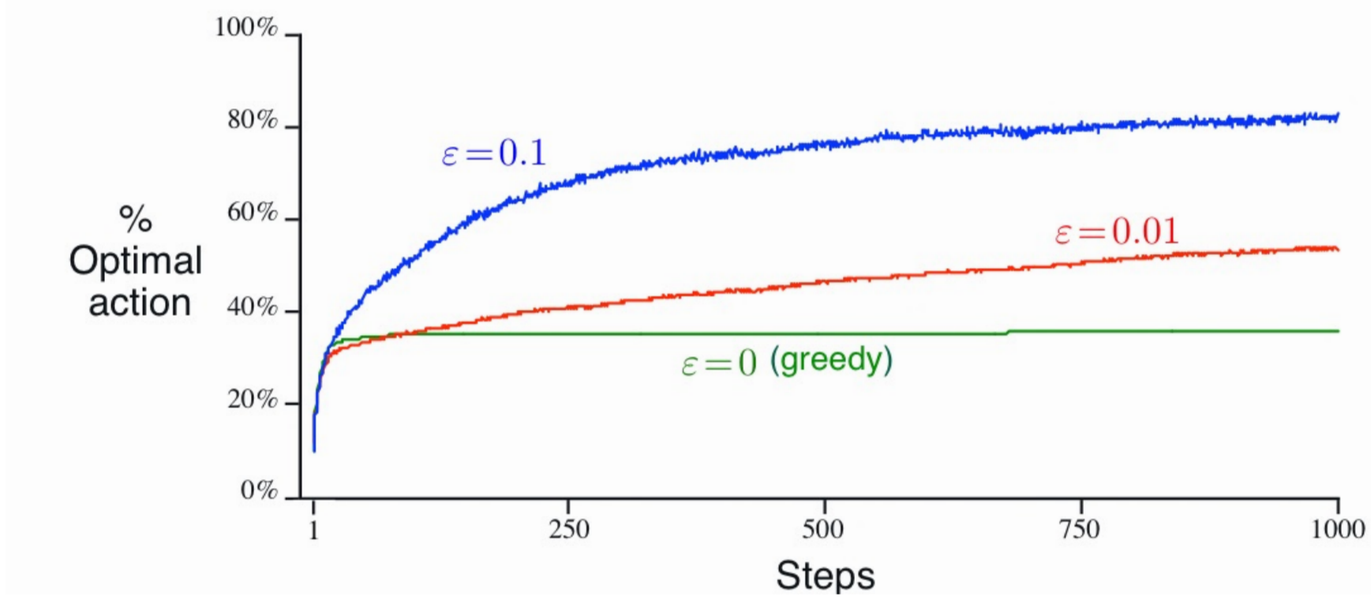
$\epsilon = 0.01$

Note: if denom. is 0, $Q_t(a) = 0$

10-armed testbed :-







ϵ
= hyperparameter for the algo

Incremental estimation of $Q_t(a)$:-

$$Q_n = \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1}$$

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i$$

$$= \frac{1}{n}\left(R_n + \sum_{i=1}^{n-1} R_i\right)$$

$$= \frac{1}{n}\left(R_n + (n-1)\frac{1}{(n-1)} \sum_{i=1}^{n-1} R_i\right)$$

$$= \frac{1}{n}\left(R_n + (n-1)\, Q_n\right)$$

$$Q_{n+1} = \frac{1}{n}(R_n + Q_n - Q_n)$$

$$Q_{n+1} = Q_n + \frac{1}{n}(R_n - Q_n)$$

$$Q_{n+1} = Q_n + \frac{1}{n}\left(R_n - Q_n\right)$$

New estimate = old estimate +

stepsize [target − old est.]

ε- greedy:-

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

$$A \leftarrow \begin{cases} \operatorname{argmax}_a Q(a) & \text{with probability } 1-\varepsilon \quad \text{(breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

$R \leftarrow bandit(A)$

$N(A) \leftarrow N(A) + 1$

$Q(A) \leftarrow Q(A) + \frac{1}{N(A)}\big[R - Q(A)\big]$

Non-stationary problems :-

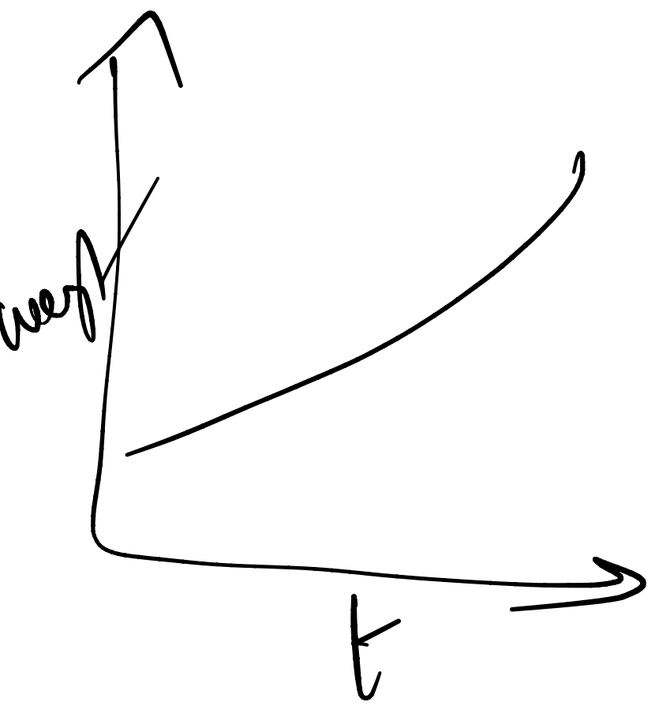
Stationary problems :- $Q_{n+1} = Q_n + \frac{1}{n}\left[R_n - Q_n\right]$

Not ideal in non-stat. settings.

$$Q_{n+1} = Q_n + \alpha\left[R_n - Q_n\right]$$

$\alpha \in (0, 1]$ is constant.

$$Q_{n+1} = Q_n + \alpha \left[R_n - Q_n \right]$$



$$= \alpha R_n + (1-\alpha) Q_n$$

$$= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha) Q_{n-1}\right]$$

$$= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1}$$

$$= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + (1-\alpha)^3 Q_{n-2}$$

$$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i$$

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

Stationary problems. | Sample average

$$\alpha_n(a) = \alpha$$

Will Not converge

non-stat. prob | recency weighted average

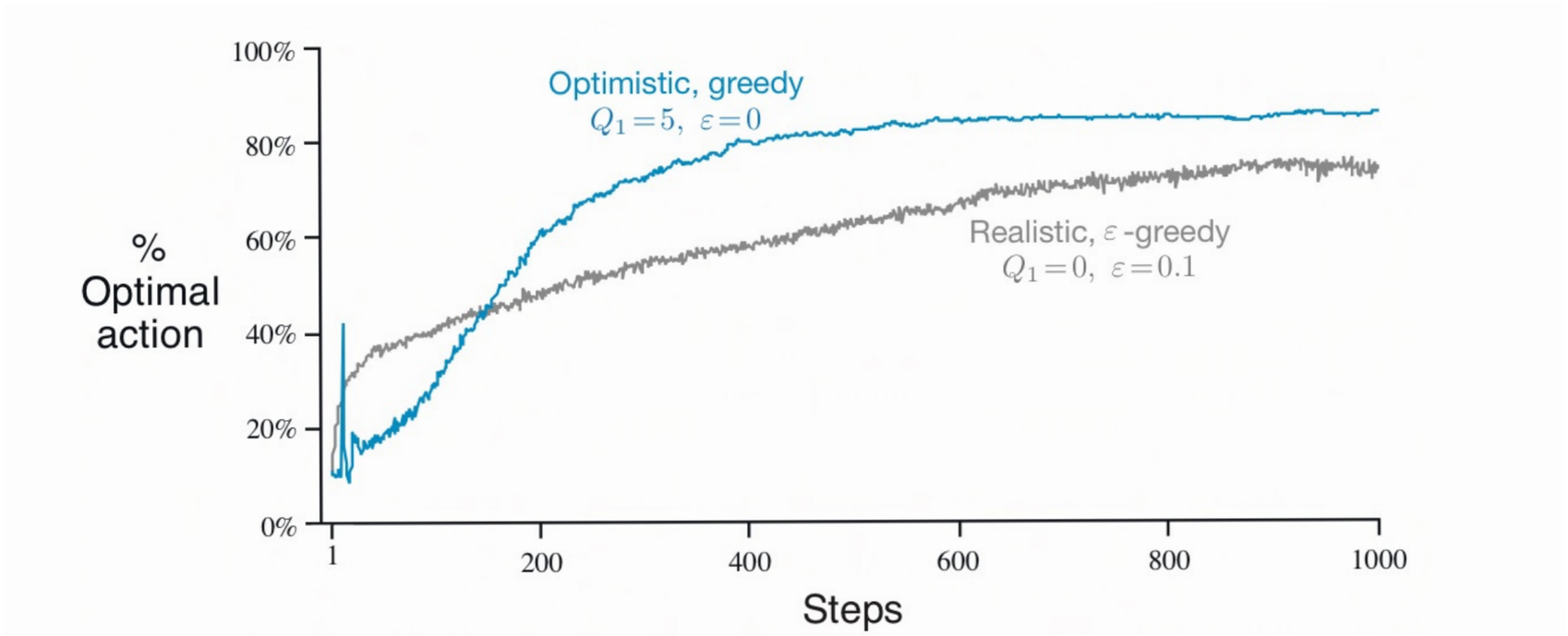
$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \checkmark$$

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

$Q_1(a) = 0$ for all a.

→ Optimistic initial value.

$Q_1(a) = 5$ for all a


100%
80%
60%
40%
20%
0%
% Optimal action
Optimistic, greedy
$Q_1 = 5,\ \varepsilon = 0$
Realistic, ε-greedy
$Q_1 = 0,\ \varepsilon = 0.1$
1
200
400
600
800
1000
Steps

Alg 3: Boltzmann exploration :-

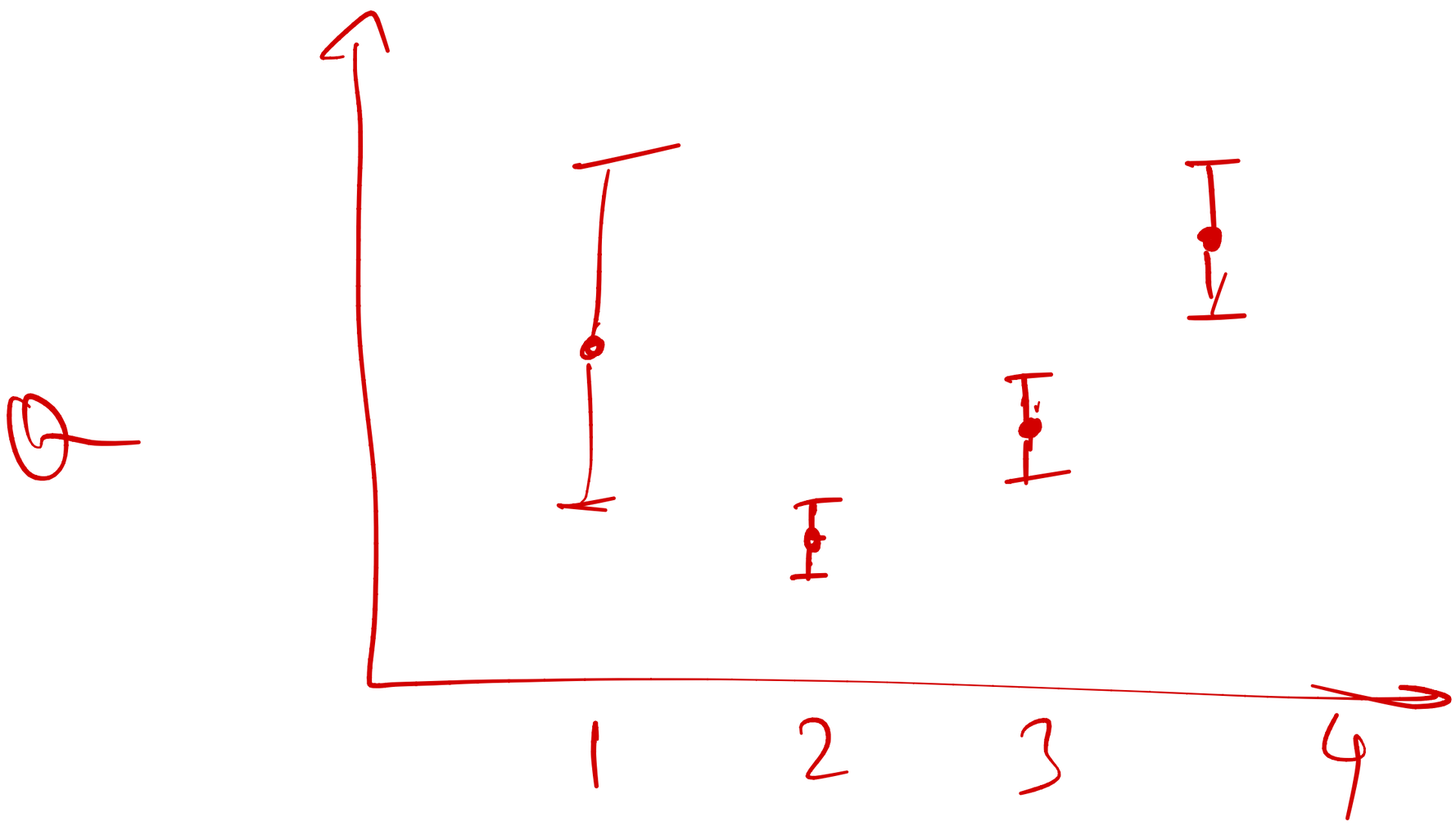
$$Q_t(a_1) \quad Q_t(a_2) \quad \cdots \quad Q_t(a_5)$$

$$Q \quad [10, \ 5, \ 20, \ 15, \ 20]$$

$$\text{Softmax}(Q) \qquad p(a_1) = \frac{e^{Q_t(a_1)/\tau}}{\sum_{i=1}^{k} e^{Q_t(a_i)/\tau}}$$

τ — temp. para.

Alg 4: Upper Confidence Bound (UCB)



UCB1 :

$$A_t = \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c \sqrt{\frac{\ln t}{N_t(a)}} \right]$$

$$[10,\ 5,\ 20,\ 15,\ 20]$$

softmax: $[0.15\ 0.05\ 0.3\ 0.2\ 0.3]$

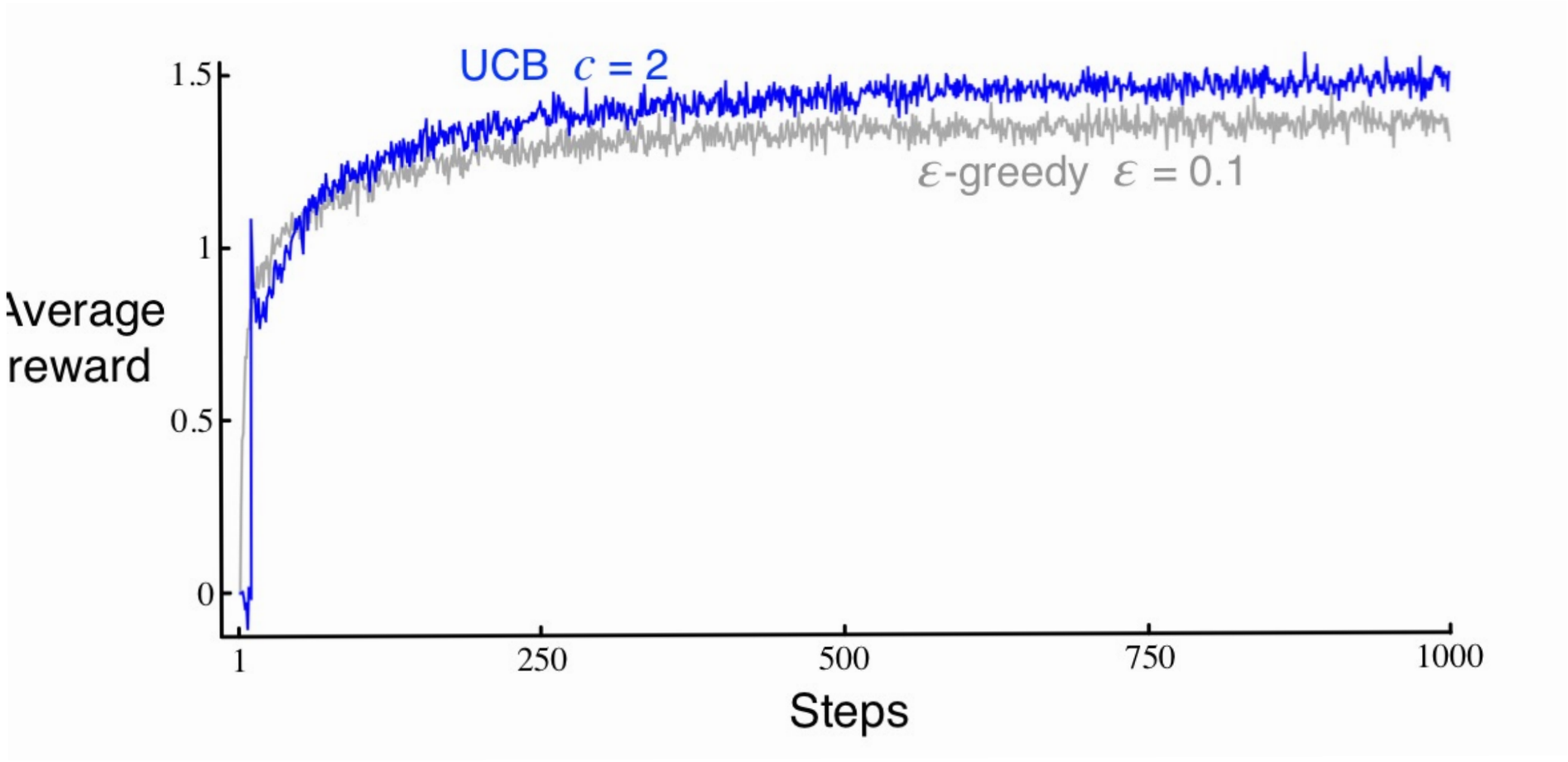
low temp

high temp

$$(0.2\ \ 0.2\ \ 0.2\ \ 0.2\ \ 0.2)$$

0.03 0.02 0.45 0.01 0.45

∧ ∧


UCB $c = 2$
ε-greedy $\varepsilon = 0.1$
Average reward
1.5
1
0.5
0
1
250
500
750
1000
Steps

Alg 5: Gradient bandit.

$H_t(a)$ pref. for taking action "a"

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

$H_1(a) = 0$ for all a.

$$H_{t+1}(A_t) = H_t(A_t) + \alpha\left(R_t - \bar{R}_t\right)\left(1 - \bar{\pi}_t(A_t)\right)$$

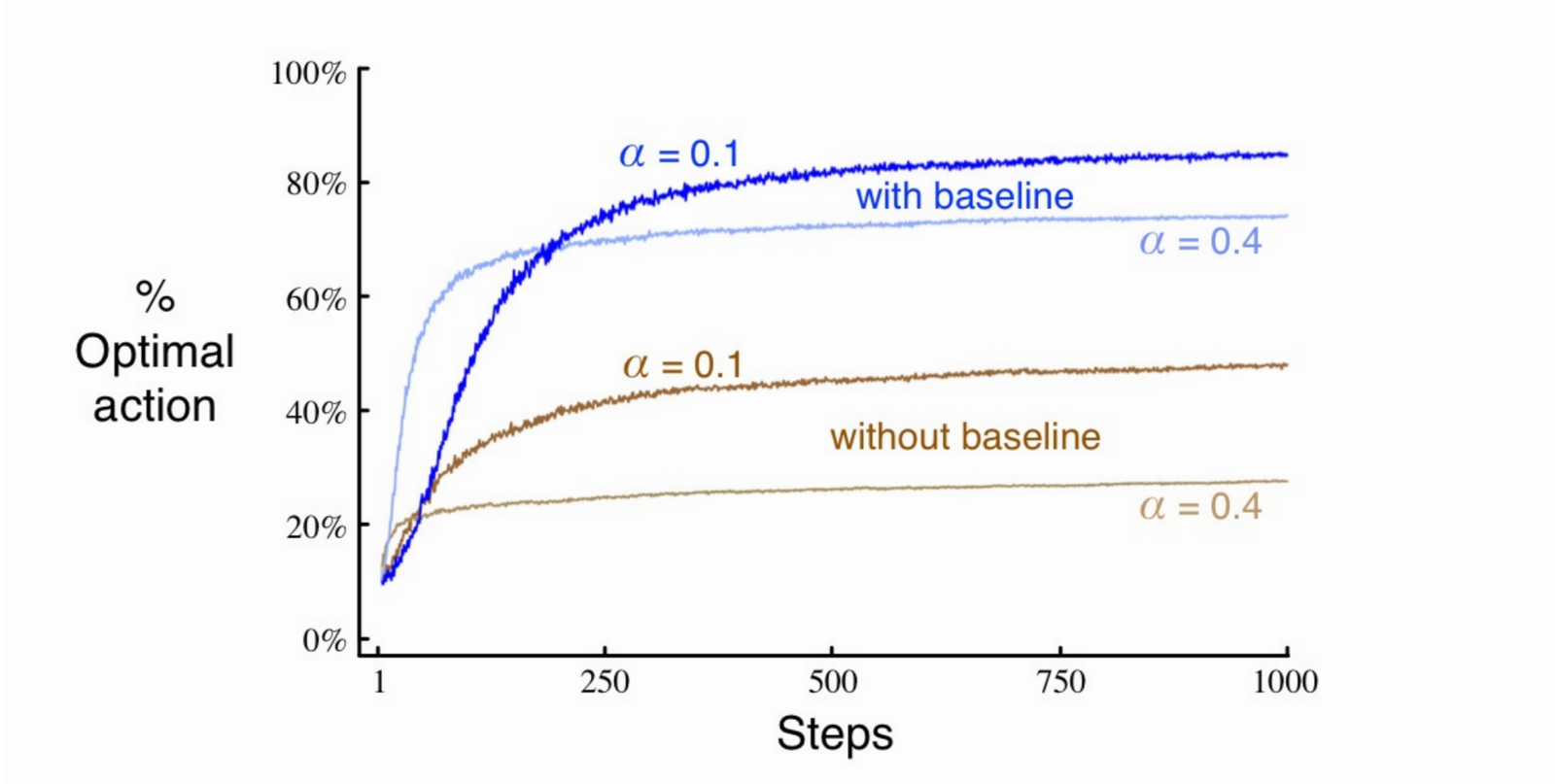
$$H_{t+1}(a) = H_t(a) - \alpha\left(R_t - \bar{R}_t\right)\pi_t(a)$$

$\forall a \neq A_t$.

$\bar{R}_t$ = average reward so far.

$R_t > \bar{R}_t$: increase the prob of A_t

$R_t < \bar{R}_t$: decrease " " "


100%
80%
60%
40%
20%
0%
% Optimal action
α = 0.1
with baseline
α = 0.4
α = 0.1
without baseline
α = 0.4
1
250
500
750
1000
Steps

Gradient bandit → Stochastic gradient ascent.

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

$$\mathbb{E}[R_t] = \sum_x \pi_t(x)\, q_*(x)$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[\sum_x \pi_t(x)\, q_*(x) \right]$$

$$= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$\frac{\partial E[R_t]}{\partial H_t(a)} = \sum_x \left(q_{t*}(x) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)}$$

$$= \sum_x \pi_t(x) \cdot \left(q_{t*}(x) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)} \Big/ \pi_t(x)$$

$$= \mathbb{E}\left[\left(q_{t*}(A_t) - B_t\right) \frac{\partial \pi_t(x)}{\partial H_t(a)} \Big/ \pi_t(A_t) \right]$$

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

$$\frac{\partial \pi_t(x)}{\partial H_t(c)} = \frac{\partial}{\partial H_t(c)} \pi_t(x)$$

$$= \frac{\partial}{\partial H_t(c)} \left[\frac{e^{H_t(x)}}{\sum_{y=1}^{k} e^{H_t(y)}} \right]$$

$$= \frac{\sum_{y=1}^{k} e^{H_t(y)} \frac{\partial e^{H_t(x)}}{\partial H_t(c)} - e^{H_t(x)} \frac{\partial \sum_{y=1}^{k} e^{H_t(y)}}{\partial H_t(c)}}{\left(\sum_{y=1}^{k} e^{H_t(y)} \right)^2}$$

$$= \frac{\sum_{y=1}^{k} e^{H_t(y)} \dfrac{\partial e^{H_t(n)}}{\partial H_t(a)} - e^{H_t(n)} \dfrac{\partial \sum_{y=1}^{k} e^{H_t(y)}}{\partial H_t(a)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)^2}$$

$$= \frac{\mathbb{1}_{a=n}\, e^{H_t(n)} \sum_{y=1}^{k} e^{H_t(y)} - e^{H_t(n)} e^{H_t(a)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)^2}$$

$$= \frac{\mathbb{1}_{a=n}\, e^{H_t(n)}}{\sum_{y=1}^{k} e^{H_t(y)}} - \frac{e^{H_t(n)} e^{H_t(a)}}{\left(\sum_{y=1}^{k} e^{H_t(y)}\right)^2}$$

$$= \mathbb{1}_{a=x} \pi_t(x) - \pi_t(x)\,\pi_t(a)$$

$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \pi_t(x)\left(\mathbb{1}_{a=x} - \pi_t(a)\right)$$

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

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

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

$$= \mathbb{E}\left[(R_t - \bar{R}_t)\left(\mathbb{1}_{a=A_t} - \pi_t(a)\right)\right)$$

$$H_{t+1}(a) = H_t(a) + \alpha\left(R_t - \bar{R}_t\right)\left(\mathbb{1}_{a=A_t} - \pi_t(a)\right) \quad \forall a$$