① **k-armed bandit**: (making decisions under uncertainty). ⇒ you can't be sure what you would like, but make the best choice you can.

医生在治疗病人时发现一种 treatment 的疗效似乎比另外2种更好.

- 如果他就用这一种 treatment 不用其他2种了, 有可能另外一种 treatment 更好, 只是需要多搭配服用.
  (他就不能收集其他2种 treatment 的数据了.)

- have **an decision maker (an agent)** 如上面例子中的医生 choose between **k actions** (在3种 treatments 里选一种) (each treatment is an action.) and receives a **reward** (如病人在使用该 treatment 后身体变好了) based on its choose.
- 当医生要 decide which action is best, 就要 find the value of taking each action, 即 **Action-Values**. (也叫 **Action-Values function**)

Defined the value of expecting an action as the expected reward: (goal: maximize the expected reward)

If the agent (医生) chose the action has the highest value, he achieve the goal. This procedure is called $\underset{a}{\operatorname{Argmax}}\, q_*(a)$. ($q_*(a)$: the mean of the distribution of each action.)

expected reward (value of an action, not known):

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \quad \forall a \in \{1, \dots, k\}$$

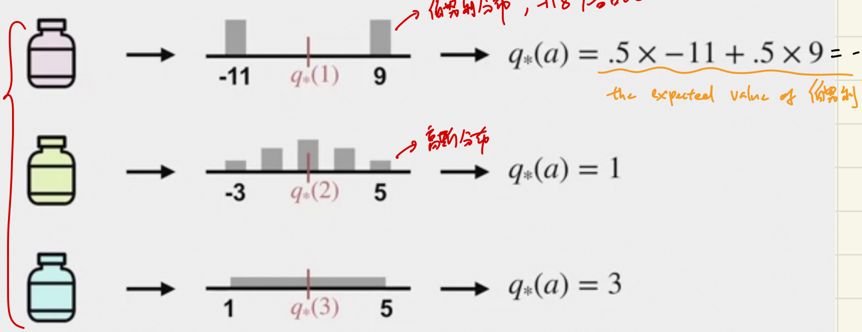
($\doteq$: is defined as; R_t: reward t; $A_t = a$: given selected action a; $\forall a \in \{1,\dots,k\}$: for each possible action $1 \sim k$.)

$$= \sum_r p(r|a)\, r$$

⇒ the sum of all possible rewards (multiply possible reward with the probability of observing that reward).

Calculating $q_*(a)$

对应每一个 action 都有不同的 distribution.



伯努利分布, -11与9各占0.5.

the expected value of 伯努利 distribution

高斯分布

② Sample-Average Method:

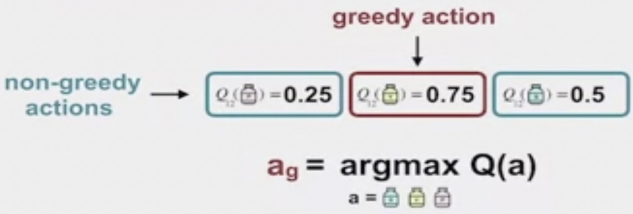
Sample-Average Method

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i}{t-1}$$

($\sum_{i=1}^{t-1} R_i$ ⇒ sum of rewards when taking action a; $t-1$ ⇒ total # of times of action a has been taken)

Action Selection



$$a_g = \underset{a}{\operatorname{argmax}}\ Q(a)$$

⑤

Exploration and Exploitation

- Exploration - improve knowledge for long-term benefit

$q(a) = 0$	$q(a) = 0$	$q(a) = 0$
$N(a) = 0$	$N(a) = 0$	$N(a) = 0$
$q_*(a) = 3$	$q_*(a) = 4$	$q_*(a) = 2$

estimate value of picking that meal ⇐ $q(a) = 0$

$N(a) = 0$ ⇒ # of times you have pick that meal

value of each meal. ⇐ $q_*(a) = 3$

1:24 / 7:36

greedy 方法不一定会导致最优解!! 因为有可能 agent 会得到喜欢某个 meal 然后只要选它,导致他根本没有发现其他 meal 的价值.

Exploration versus Exploitation

∴ 该选择 Explore (选很多不同的菜).

- Exploration - improve knowledge for long-term benefit ⇒ get more accurate estimate to our value
- Exploitation - exploit knowledge for short-term benefit ⇒ might get more reward

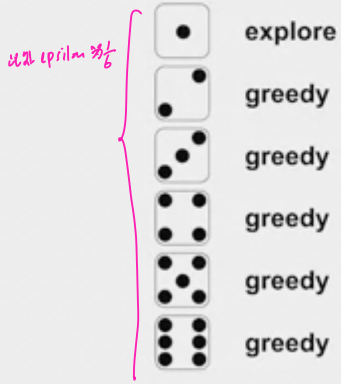
还是 Exploit? (继续选一个菜)

(need to choose when to explore and when to exploit?)

- How do we choose when to explore and when to exploit?

Epslion: 继续 explore 的概率:

Epsilon-Greedy Action Selection



Epsilon-Greedy Action Selection

$$A_t \leftarrow \begin{cases} \underset{a}{\text{argmax}}\ Q_t(a) & \text{with probability } 1-\epsilon \\ a \sim Uniform(\{a_1 \dots a_k\}) & \text{with probability } \epsilon \end{cases}$$

Optimistic Initial Values

$Q_1(\text{P}) = 2.0$

$Q_1(\text{Y}) = 2.0$

$Q_1(\text{B}) = 2.0$

④

Incremental update rule

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

$3.8 = 4 + \frac{1}{n}(2 - 4)$

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

the new reward is our target

$\frac{1}{n}$ (stepsize) produces a number between 0 to 1, 例如 0.51.

$= \frac{1}{n}$

Upper-Confidence Bound (UCB) Action Selection

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

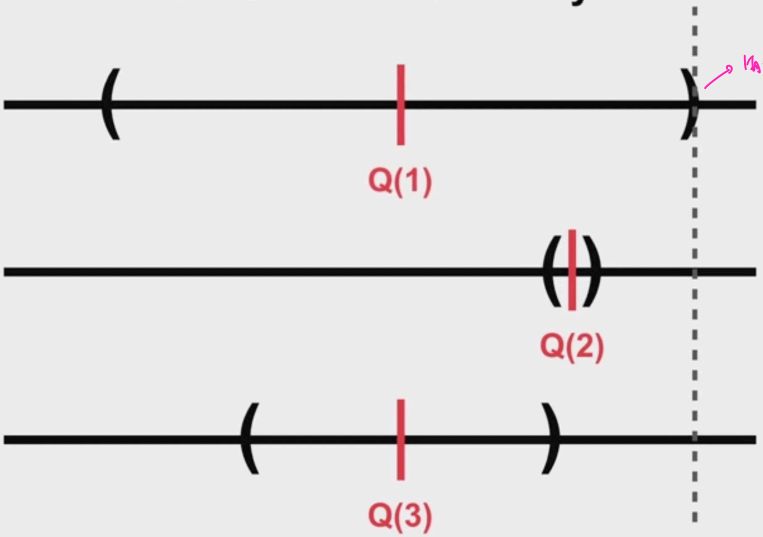
Exploit

Explore

2:37 / 3:10

use the strategy of:

Optimism in the Face of Uncertainty



Sequential Decision-Making

Latest Submission Grade 100%

1. What is the incremental rule (sample average) for action values? **1 / 1 point**

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

Correct
Correct! At each time step the agent moves its prediction in the direction of the error by the step size (here 1/n).

Back

Sequential Decision-Making

Graded Quiz • 45 min

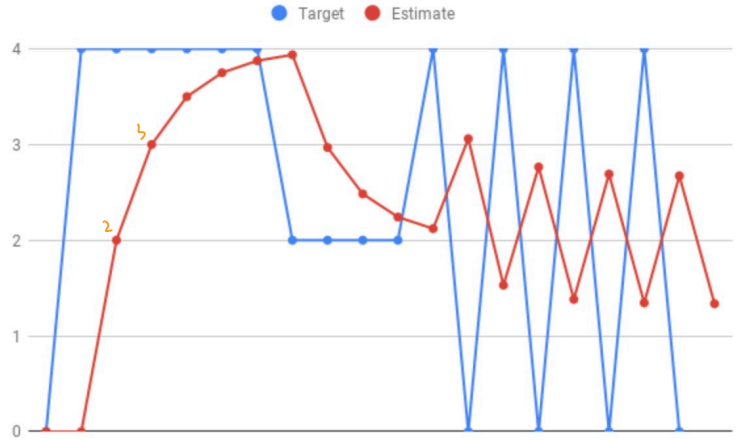
Due Jan 11, 10:59 PM PST

2. Equation 2.5 (from the SB textbook, 2nd edition) is a key update rule we will use throughout the Specialization. We discussed this equation extensively in video. This exercise will give you a better hands-on feel for how it works. The blue line is the target that we might estimate with equation 2.5. The red line is our estimate plotted over time.

1 / 1 point

$$q_{n+1} = q_n + \alpha_n[R_n - q_n]$$

Given the estimate update in red, what do you think was the value of the step size parameter we used to update the estimate on each time step?



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

∴ 此时 $Q_{n+1} = 3$, $Q_n = 2$, 故 R_n (new reward) 为 estimate 与目标值 Target 值, 即 4.

即有 $3 = 2 + \frac{1}{n}[4-2]$

$= 2 + \frac{1}{n}[2]$

$3 = 2 + 0.5 \times 2$

∴ $\frac{1}{n} = 0.5$

∴ $\alpha_n = 0.5$

∴ Stepsize $= 0.5$

- 1.0
- 1/2
- 1/8
- 1 / (t - 1)

Correct

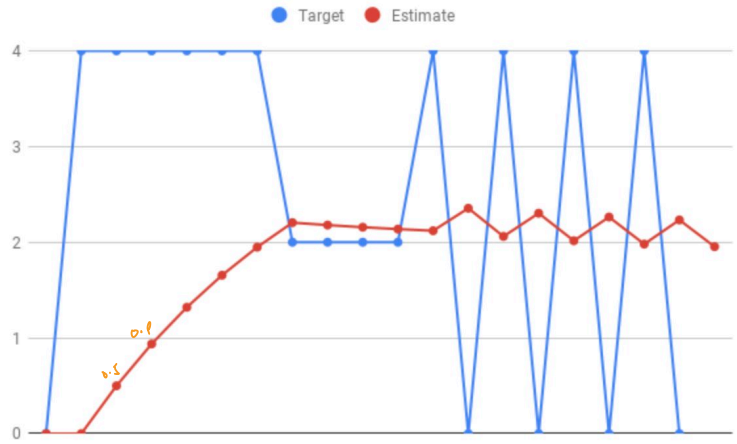
Correct! We can see that the estimate is updated by about half of what the prediction error is.

3. Equation 2.5 (from the SB textbook, 2nd edition) is a key update rule we will use throughout the Specialization. We discussed this equation extensively in video. This exercise will give you a better hands-on feel for how it works. The blue line is the target that we might estimate with equation 2.5. The red line is our estimate plotted over time.

1 / 1 point

$$q_{n+1} = q_n + \alpha_n[R_n - q_n]$$

Given the estimate update in red, what do you think was the value of the step size parameter we used to update the estimate on each time step?



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

∴ 此处设 $Q_{n+1} = 0.9$, $Q_n = 0.5$, 故 R_n (new reward) 为 estimate: 0.9 对应的 Target 值, 即 4.

即有 $0.9 = 0.5 + \frac{1}{n}[4 - 0.5]$

$0.4 = \frac{1}{n}[3.5]$

$\frac{0.4}{3.5} = \frac{1}{n}$

$\therefore \frac{1}{n} = \frac{1}{8}$

$\therefore \alpha_n = \frac{1}{8}$

$\therefore$ Stepsize $= \frac{1}{8}$

- 1 / (t - 1)
- 1.0
- 1/2
- 1/8 ✓

Correct
Correct! We can see that the estimate is updated by ⅛ of the prediction error at each time step.

4. Equation 2.5 (from the SB textbook, 2nd edition) is a key update rule we will use throughout the Specialization. We discussed this equation extensively in video. This exercise will give you a better hands-on feel for how it works. The blue line is the target that we might estimate with equation 2.5. The red line is our estimate plotted over time.

1 / 1 point

$$q_{n+1} = q_n + \alpha_n[R_n - q_n]$$

Given the estimate update in red, what do you think was the value of the step size parameter we used to update the estimate on each time step?

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

∴ 此处设 $Q_{n+1} = 4$, $Q_n = 4$, 故 R_n (new reward) 为 estimate = 4 及处的 Target 值, 即 4.

即有 $4 = 4 + \frac{1}{n}[4-4]$

$0 = 4 + \frac{1}{n}[0]$

∴ $\frac{1}{n} = 1$

∴ $\alpha_n = 1$

∴ Stepsize = 1

- (●) 1.0
- ( ) 1/8
- ( ) 1/2
- ( ) 1 / (t - 1)

Correct
Correct! The estimate is updated to what the previous target was.

5. Equation 2.5 (from the SB textbook, 2nd edition) is a key update rule we will use throughout the Specialization. We discussed this equation extensively in video. This exercise will give you a better hands-on feel for how it works. The blue line is the target that we might estimate with equation 2.5. The red line is our estimate plotted over time.

1 / 1 point

$$q_{n+1} = q_n + \alpha_n[R_n - q_n]$$

Given the estimate update in red, what do you think was the value of the step size parameter we used to update the estimate on each time step?

- 1.0
- 1/2
- 1/8
- 1 / (t - 1)

Correct

Correct! We can see that the estimate is updated fully to the target initially, and then over time the amount that the estimate updates is reduced. This indicates that our step size is reducing over time.

6. What is the exploration/exploitation tradeoff? **1 / 1 point**

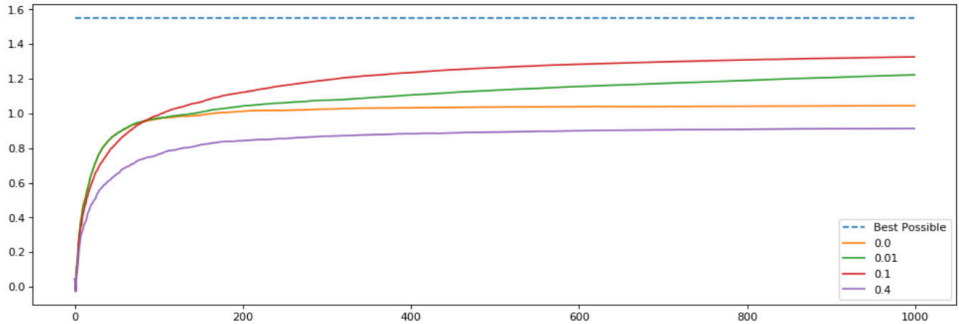
- (●) The agent wants to explore to get more accurate estimates of its values. The agent also wants to exploit to get more reward. The agent cannot, however, choose to do both simultaneously.
- ( ) The agent wants to explore the environment to learn as much about it as possible about the various actions. That way once it knows every arm's true value it can choose the best one for the rest of the time.
- ( ) The agent wants to maximize the amount of reward it receives over its lifetime. To do so it needs to avoid the action it believes is worst to exploit what it knows about the environment. However to discover which arm is truly worst it needs to explore different actions which potentially will lead it to take the worst action at times.

Correct

Correct! The agent wants to maximize the amount of reward it receives over time, but needs to explore to find the right action.

7. Why did epsilon of 0.1 perform better over 1000 steps than epsilon of 0.01?

1 / 1 point

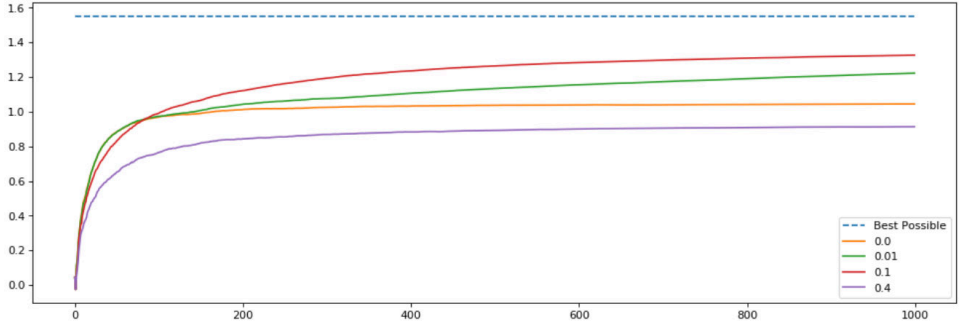


- (●) The 0.01 agent did not explore enough. Thus it ended up selecting a suboptimal arm for longer.
- ( ) The 0.01 agent explored too much causing the arm to choose a bad action too often.
- ( ) Epsilon of 0.1 is the optimal value for epsilon in general.

Correct

Correct! The agent needs to be able to explore enough to be able to find the best arm to pull over time. Here epsilon of 0.01 does not allow for enough exploration in the time allotted.

8. If exploration is so great why did epsilon of 0.0 (a greedy agent) perform better than epsilon of 0.4? **1 / 1 point**



- ○ Epsilon of 0.4 doesn't explore often enough to find the optimal action.
- ○ Epsilon of 0.0 is greedy, thus it will always choose the optimal arm.
- ◉ Epsilon of 0.4 explores too often that it takes many sub-optimal actions causing it to do worse over the long term.

Correct

Correct! While we want to explore to find the best arm, if we explore too much we can spend too much time choosing bad actions even when we know the correct one. In this case the action-value estimates are likely correct, however the policy does not always choose the action with the highest value.