Chapter 2

- N-armed bandit problem → nonassociative

Exploit with greedy actions
Explore with nongreedy actions

- $q(a)$ is the actual value of action a

$Q_t(a)$ is the estimated value of action a at time t

True value of an action is the mean reward received when taking that action.

(2.1)
- $$Q_t(a) = \frac{R_1 + R_2 + \ldots + R_{N_t(a)}}{N_t(a)}$$

$N_t(a)$ = number of times action a was selected prior to t

$Q_t(a)$ converges to $q(a)$ as $N_t(a) \to \infty$

↳ sample average method

Chapter 2

- $\underset{x}{\operatorname{argmax}} f(x) = \{ x \mid \forall_y : f(x) \geq f(y) \}$

  $\max_x f(x) = \{ f(x) \mid \forall y : f(x) \geq f(y) \}$

- Greedy action selection

(2.2) $A_t = \underset{a}{\operatorname{argmax}} Q_t(a)$

Select the action with the highest estimated action-value

- 10-armed bandit - $q(a)$ for each one is chosen according to a normal (Gaussian) distribution with mean = 0 and variance = 1. This is the mean reward received when a given action is selected

The actual reward per step is the $q(a)$ plus a random number chosen according to normal distribution with mean = 0 and variance = 1

- Nongreedy action chosen randomly according to discrete uniform distribution

Chapter 2

- Incrementally update $Q(a)$

(2.3) $$Q_{k+1} = \frac{1}{k}\sum_{i=1}^{k} R_i \implies Q_k = \frac{1}{k-1}\sum_{i=1}^{k-1} R_i$$

$$= \frac{1}{k}\left[R_k + \sum_{i=1}^{k-1} R_i\right]$$

$$= \frac{1}{k}\left[R_k + (k-1)\left(\frac{1}{k-1}\sum_{i=1}^{k-1} R_i\right)\right]$$

$$= \frac{1}{k}\left[R_k + (k-1)\,Q_k\right]$$

$$= \frac{1}{k}\left[R_k + kQ_k - Q_k\right]$$

$$= Q_k + \frac{1}{k}\left[R_k - Q_k\right]$$

(2.4) - New $\leftarrow$ Old + StepSize $\underbrace{[\text{Target} - \text{Old}]}_{\text{error}}$

- $\alpha = \frac{1}{k}$, more generally $\alpha_t(a)$

(2.5)

- $Q_{k+1} = Q_k + \frac{1}{k}(R_k - Q_k)$

$= Q_k + \alpha(R_k - Q_k)$

- $\alpha \in (0, 1] \Rightarrow 0 < \alpha \leq 1$

α is the step size parameter

α is a fixed value for nonstationary problems

- Example

$R_1 = 1$

$R_2 = 10$

$Q_1 = 0 \quad R_3 = 7$

$Q_2 = \frac{1}{1} = 1$

$Q_3 = \frac{1+10}{2} = 5.5$

$Q_4 = \frac{1+10+7}{3} = 6$

$Q_2 = Q_1 + \frac{1}{k}(R_1 - Q_1) = 0 + \frac{1}{1}(1-0) = 1$

$Q_3 = Q_2 + \frac{1}{k}(R_2 - Q_2) = 1 + \frac{1}{2}(10-1) = 5.5$

$Q_4 = Q_3 + \frac{1}{k}(R_3 - Q_3) = 5.5 + \frac{1}{3}(7-5.5) = 6$

- α is a fixed value for nonstationary problems

(2.6) - $Q_{k+1} = Q_k + \alpha(R_k - Q_k) \Rightarrow Q_k = Q_{k-1} + \alpha(R_{k-1} - Q_{k-1})$

$$= Q_k + \alpha R_k - \alpha Q_k \qquad = Q_{k-1} + \alpha R_{k-1} - \alpha Q_{k-1}$$

$$= \alpha R_k + (1-\alpha) Q_k \qquad = \alpha R_{k-1} + (1-\alpha) Q_{k-1}$$

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

$$= \alpha R_k + (1-\alpha)\alpha R_{k-1} + (1-\alpha)^2 Q_{k-1}$$

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

$$\begin{aligned} = \;& (1-\alpha)^0 \alpha R_k \\ &+ (1-\alpha)^1 \alpha R_{k-1} \\ &+ (1-\alpha)^2 \alpha R_{k-2} \\ &+ \dots \\ &+ (1-\alpha)^{k-1} \alpha R_1 \\ &+ (1-\alpha)^k Q_1 \end{aligned}$$

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

- $(1-\alpha)^k + \sum_{i=1}^{k} \alpha (1-\alpha)^{k-i} = 1$

weighted average of past rewards

- $(1-\alpha)^k + \sum_{i=1}^{k} \alpha(1-\alpha)^{k-i} = 1$

let $k=3$

$$(1-\alpha)^3 + \alpha(1-\alpha)^2 + \alpha(1-\alpha)^1 + \alpha(1-\alpha)^0$$

$$= (1-\alpha)(1-2\alpha+\alpha^2) + \alpha(1-2\alpha+\alpha^2) + \alpha - \alpha^2 + \alpha$$

$$= 1-2\alpha+\alpha^2 - \alpha(1-2\alpha+\alpha^2) + \alpha - 2\alpha^2 + \alpha^3 + 2\alpha - \alpha^2$$

$$= 1-2\alpha+\alpha^2 - \alpha + 2\alpha^2 - \alpha^3 + 3\alpha - 3\alpha^2 + \alpha^3$$

$= 1 \longrightarrow$ all the α cancel out

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

(2.5) $= Q_k + \alpha(R_k - Q_k)$, for nonstationary

$\alpha_k(a) = \frac{1}{k}$ for stationary problems

Q_1 = initial estimated value (not Q_0 in this notation)

- Optimistic initial values encourage exploration at the beginning of the run. Not suitable for nonstationary problems

- Upper confidence bound - select nongreedy actions according to potential for actually being optimal

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

where $c > 0$, controls the degree of exploration

$\sqrt{\frac{\ln t}{N_t(a)}}$ is a measure of the uncertainty in the estimate of the action value

- UCB is difficult to use in nonstationary problems, but gives good results for simple bandit problems

- Why a spike at $t=11$ in Figure 2.3? Note that when $N_t(a)=0$, then it is considered a maximizing action. The first 10 plays cycle through the actions that haven't been picked yet. On the 11th play, it's going to choose the action with the highest initial estimate (more likely to be one of the best actions). On the 12th play the action chosen at $t=11$ doesn't have as high a weight as the others

- Gradient ascent bandit uses soft-max distribution (aka Gibbs or Boltzmann distribution)

(2.9) $$P_r\{A_t = a\} = \frac{e^{H_t(a)}}{\sum_{b=1}^{n} e^{H_t(b)}} = \pi_t(a)$$

$\pi_t(a)$ = Probability of taking action a at time t

$H_1(a) = 0$ for all a $\longrightarrow$ initial value

$H_t(a)$ is a numerical preference for action a

- Learning algorithm based on stochastic gradient ascent. Preferences updated each step of selecting A_t and receiving reward R_t

(2.10) $$H_{t+1}(A_t) = H_t(A_t) + \alpha(R_t - \bar{R}_T)(1 - \pi_t(A_t))$$

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

α is the step size parameter

$\bar{R}_t$ is the average of all rewards up to and including t. This value can be computed incrementally.

Chapter 2

- Examples so far have been nonassociative. Associative search tasks make policy decisions based on clues from the environment. Associative search tasks are also called contextual bandits.

If actions are allowed to affect the next situation as well as the reward, then we have the full reinforcemt learning problem.