Multi-armed Bandits

Rupam Mahmood

January 15, 2020

Expectations review

- ✓ Expectation: $E[X] = \sum x\ P(X=x)$
- ✓ summarizes the outcomes of an experiments
- ✓ Conditional expectation w.r.t. an event: $E[X \mid Y=y] = \sum x\ P(X=x \mid Y=y)$
- ✓ Conditional expectation w.r.t a random variable: $E[X \mid Y]$
- ✓ … is a random variable that maps outcomes to numbers:
- ✓ $E[X \mid Y](y) = E[X \mid Y=y]$
- ✓ … is a function of Y that “best approximates” X

Worksheet question 1

Action values

$$q_*(a) \doteq E[R \,|\, A = a]$$

- ✓ What is the sample space?
- ✓ What is the optimal behavior?

Estimating action values

$$
\begin{aligned}
Q_{n+1} = \frac{1}{n}\sum_{i=1}^{n} R_i \quad &= \frac{1}{n}\sum_{i=1}^{n-1} R_i + \frac{1}{n} R_n \\
&= \frac{n-1}{n}\,\underbrace{\frac{1}{n-1}\sum_{i=1}^{n-1} R_i}_{?} + \frac{1}{n} R_n \\
&= \left(1 - \frac{1}{n}\right) Q_n + \frac{1}{n} R_n \\
&= Q_n + \frac{1}{n}\left(R_n - Q_n\right)
\end{aligned}
$$

Life of a bandit agent

Initialize N and Q

Loop forever:

take an action A based on Q and an action-selection strategy

Observe reward R

Update estimates N and Q

Pseudocode

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

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

$R \leftarrow bandit(A)$

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

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

Life of an RL-glue bandit agent

…

```
for i in range(num_steps):

    reward, _, action, _ = rl_glue.rl_step()
```

A simple bandit algorithm

Initialize, for $a = 1$ to k:

$Q(a) \leftarrow 0$

$N(a) \leftarrow 0$

Loop forever:

agent step

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

$R \leftarrow bandit(A)$ environment step

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

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

Review & clarifications

- ✓ Is the action which has highest expected reward(value) the optimal action?
- ✓ What exactly the stepsize is? The meaning and influence of it? and why it could be constant?
- ✓ Is it possible that in e-greedy, with probability e, the action taken by agent randomly could be the greedy action again?