Reinforcement Learning

Multi-Armed Bandits

Stefano V. Albrecht

20 January 2023
Lecture Outline

- Multi-armed bandit problem
- Exploration-exploitation dilemma
- Action-value methods
- Gradient methods
Multi-Armed Bandit Problem

Multi-armed bandit (MAB) problem:

- There are \( k \) actions (“arms”) to choose from
- On each time step \( t = 1, 2, 3, \ldots \), you choose an action \( A_t = a \) and receive a scalar reward sampled from some unknown random variable \( R_t \), where

\[
q_*(a) = \mathbb{E}[R_t | A_t = a]
\]

\( R_t \) are iid (independently and identically distributed)
- Goal: maximise total received rewards over time
Exploration-Exploitation Dilemma

- We can form action-value estimates:
  \[ Q_t(a) \approx q_*(a) \]

- The greedy action at time \( t \) is:
  \[ A^*_t \triangleq \arg \max_a Q_t(a) \]

- **Exploitation:** choose \( A_t = A^*_t \);  **Exploration:** choose \( A_t \neq A^*_t \)

**Exploration-exploitation problem:**
How to balance exploration and exploitation to maximise rewards?
⇒ Can’t exploit or explore all the time (why?)
Action-Value Methods

Action-value methods:

- Learn action-value estimates

- E.g. sample average:

\[
Q_t(a) = \frac{1}{N_t(a)} \sum_{\tau=1}^{t-1} R_\tau \cdot [A_\tau = a]
\]

where \(N_t(a)\) is number of times action \(a\) was selected until before \(t\)

- Sample average converges to true action values in the limit:

\[
\lim_{N_t(a) \to \infty} Q_t(a) = q_*(a)
\]
\( \epsilon \)-Greedy Action Selection

- Greedy action selection:
  \[
  A_t = A_t^* = \arg \max_a Q_t(a)
  \]

- \( \epsilon \)-greedy action selection:
  \[
  A_t = \begin{cases} 
    A_t^* & \text{with probability } 1 - \epsilon \\
    \text{random action} & \text{otherwise}
  \end{cases}
  \]

- Simplest way to balance exploration and exploitation
2000 random MABs each with 10 arms
normal reward dist.
each 1000 time steps
Where is $\epsilon = 0.1$ after 10,000 time steps?
• Sample average (for 1-armed bandit):

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

• Can compute incrementally:

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

• This is a standard form for update rules:

\[ \text{NewEstimate} \leftarrow \text{OldEstimate} + \text{StepSize} [\text{Target} - \text{OldEstimate}] \]
Derivation of Incremental Update

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

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

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

\[ = \frac{1}{n} \left( R_n + (n - 1)Q_n \right) \]

\[ = \frac{1}{n} \left( R_n + nQ_n - Q_n \right) \]

\[ = Q_n + \frac{1}{n} \left[ R_n - Q_n \right], \]
A simple bandit algorithm

Initialize, for $a = 1$ to $k$:
- $Q(a) \leftarrow 0$
- $N(a) \leftarrow 0$

Loop forever:
- $A \leftarrow \begin{cases} \arg \max_a Q(a) & \text{with probability } 1 - \varepsilon \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$ (breaking ties randomly)
- $R \leftarrow \text{bandit}(A)$
- $N(A) \leftarrow N(A) + 1$
- $Q(A) \leftarrow Q(A) + \frac{1}{N(A)} [R - Q(A)]$
Non-Stationary Action Values

Suppose the true action values change slowly over time

- We then say that the problem is non-stationary
- Sample average not appropriate (why?)
- Many RL methods have to deal with non-stationarity (e.g. due to bootstrapping)

Have to “track” action values, e.g. using step size parameter $\alpha \in (0, 1]$

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$
Estimates $Q_t(a)$ will converge to true values $q_*(a)$ with probability 1 if:

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

- e.g. $\alpha_n = \frac{1}{n}$
- not $\alpha_n = \frac{1}{n^2}$
- not $\alpha_n = c$ (constant)
Optimistic Initial Values

All methods so far depend on initial estimates $Q_1$

$\Rightarrow$ Can incentivise exploration by using “optimistic” initial values

e.g. $\forall a : Q_1(a) = 5$ and $\alpha = 0.1$
Instead exploring uniform-randomly ($\epsilon$-greedy), explore “promising” actions first.

**Upper Confidence Bound (UCB):** estimate upper confidence bounds on action value estimates and choose action with highest bound:

$$A_t = \begin{cases} a & \text{if } N_t(a) = 0, \text{ else} \\ \arg \max_a \left[ Q_t(a) + c \sqrt{\log t / N_t(a)} \right] \end{cases}$$

(Note: standard UCB assumes rewards in $[0, 1]$)

**Intuition:** second term is size of one-sided confidence interval for average reward
Upper Confidence Bound (UCB) Action Selection

UCB \( c = 2 \)

\( \varepsilon \)-greedy \( \varepsilon = 0.1 \)

Average reward

Steps

See Tutorial 2
Greedy, $\epsilon$-greedy, and UCB use estimates of $q_*(a)$

- Can we select actions without computing estimates of $q_*$?
Gradient Bandit Algorithm

Gradient-based policy optimisation:

- Use differentiable policy $\pi_t(a|\theta)$ with parameter vector $\theta \in \mathbb{R}^d$

  \[ \pi_t(a|\theta) = \Pr\{A_t = a \mid \theta_t = \theta\} \]

- Use gradient ascent on policy parameters to maximise expected reward

  \[ \theta_{t+1} = \theta_t + \alpha \nabla_{\theta_t} \mathbb{E}[R_t] \]
Gradient Bandit Algorithm with Softmax

- Represent $\pi_t$ with softmax distribution:

$$
\pi_t(a) = \frac{e^{H_t(a)}}{\sum_b e^{H_t(b)}}
$$

$H_t(a)$ are preference values (parameters)

- Update policy parameters:

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

$$
= H_t(a) + \alpha (R_t - \bar{R}_t)([a = A_t]_1 - \pi_t(a))
$$

with baseline $\bar{R}_t = \frac{1}{t} \sum_{\tau=1}^t R_\tau$
Gradient Bandit Algorithm

\[ \tilde{R}_t = \frac{1}{t} \sum_{\tau} R_{\tau} \]

Baseline reduces variance in updates
Deterministic Policies

$$H_{t+1}(a) = H_t(a) + \alpha (R_t - \bar{R}_t) ([a = A_t]_1 - \pi_t(a))$$

Bonus questions:

- What if some actions have zero probability?
- E.g. what if initial policy is deterministic?

$$\pi_1(a) = 1 \text{ for some } a$$
Summary Comparison of Bandit Algorithms

Average reward over first 1000 steps

- $\varepsilon$-greedy
- UCB
- Greedy with optimistic initialization ($\alpha = 0.1$)
- Gradient bandit
Multi-armed bandit problem is simplest type of RL problem

- Bandit algorithms seek to maximise total reward over extended time
- Must balance *exploration and exploitation* – a key problem in RL
- First building block for more complex RL algorithms
Required:

- RL book, Chapter 2 (2.1–2.8)
  (Box “The Bandit Gradient Algorithm as Stochastic Gradient Ascent” in Sec 2.8 is not examined)

Optional:

- UCB paper:

- Bandit Algorithms
  by Tor Lattimore and Csaba Szepesvári
In *exact* gradient ascent:

\[
H_{t+1} = H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} \quad \text{where} \quad \mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x)
\]
In **exact** gradient ascent:

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

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

\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \left[ \sum_x \pi_t(x) q_*(x) \right]
\]
In exact gradient ascent:

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

where

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

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

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

(product derivative rule)
In **exact** gradient ascent:

\[
H_{t+1} = H_t(a) + \alpha \frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} \quad \text{where} \quad \mathbb{E}[R_t] = \sum_x \pi_t(x) q_*(x)
\]

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

\[
= \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} \quad \text{(product derivative rule)}
\]

\[
= \sum_x \left( q_*(x) - B_t \right) \frac{\partial \pi_t(x)}{\partial H_t(a)} \quad \text{(} B_t \text{ is “baseline”)}
\]
$$\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x)$$
\[
\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x) \\
= \frac{\partial}{\partial H_t(a)} \left[ e^{H_t(x)} \sum_y e^{H_t(y)} \right]
\]
\[
\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x) = \frac{\partial}{\partial H_t(a)} \left[ \frac{e^{H_t(x)}}{\sum_y e^{H_t(y)}} \right] = \frac{\frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_y e^{H_t(y)} - e^{H_t(x)} \frac{\partial}{\partial H_t(a)} \sum_y e^{H_t(y)}}{(\sum_y e^{H_t(y)})^2} \quad \text{(quotient derivative rule)}
\]
\[
\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x)
\]

\[
= \frac{\partial}{\partial H_t(a)} \left[ \frac{e^{H_t(x)}}{\sum_y e^{H_t(y)}} \right]
\]

\[
= \frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_y e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_y e^{H_t(y)}}{\partial H_t(a)}
\]

\[
= \frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_y e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_y e^{H_t(y)}}{\partial H_t(a)}
\]

\[
= \frac{[a = x]_1 e^{H_t(x)} \sum_y e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_y e^{H_t(y)})^2}
\]

(quotient derivative rule)

\[
= \frac{[a = x]_1 e^{H_t(x)} \sum_y e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_y e^{H_t(y)})^2}
\]

\[
([b]_1 = 1 \text{ iff } b \text{ is true, else } 0)
\]

\[
\left( \frac{\partial e^x}{\partial x} = e^x \right)
\]
\[
\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x)
\]
\[
= \frac{\partial}{\partial H_t(a)} \left[ \frac{e^{H_t(x)}}{\sum_y e^{H_t(y)}} \right]
\]
\[
= \frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_y e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_y e^{H_t(y)}}{\partial H_t(a)}
\]
\[
= \frac{[a = x] e^{H_t(x)} \sum_y e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_y e^{H_t(y)})^2}
\]
\[
= \frac{[a = x] e^{H_t(x)} \sum_y e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{(\sum_y e^{H_t(y)})^2}
\]
\[
= [a = x] \frac{e^{H_t(x)}}{\sum_y e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{(\sum_y e^{H_t(y)})^2}
\]

(quotient derivative rule)

\[
[b]_1 = 1 \text{ iff } b \text{ is true, else 0}
\]

\[
(\frac{\partial e^x}{\partial x} = e^x)
\]
\[
\frac{\partial \pi_t(x)}{\partial H_t(a)} = \frac{\partial}{\partial H_t(a)} \pi_t(x) \\
= \frac{\partial}{\partial H_t(a)} \left[ \frac{e^{H_t(x)}}{\sum_y e^{H_t(y)}} \right] \\
= \frac{\partial e^{H_t(x)}}{\partial H_t(a)} \sum_y e^{H_t(y)} - e^{H_t(x)} \frac{\partial \sum_y e^{H_t(y)}}{\partial H_t(a)} \\
= \frac{[a = x]_1 e^{H_t(x)} \sum_y e^{H_t(y)} - e^{H_t(x)} e^{H_t(a)}}{\left( \sum_y e^{H_t(y)} \right)^2} \\
= \frac{[a = x]_1 e^{H_t(x)}}{\sum_y e^{H_t(y)}} - \frac{e^{H_t(x)} e^{H_t(a)}}{\left( \sum_y e^{H_t(y)} \right)^2} \\
= [a = x]_1 \pi_t(x) - \pi_t(x) \pi_t(a) \\
= \pi_t(x) \left( [a = x]_1 - \pi_t(a) \right)
\]
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x) \quad \text{(multiply by } \pi_t(x)/\pi_t(x)\text{)}
\]
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x) (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x)
\]  
\[
= \mathbb{E} \left[ (q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right]
\]

(multiply by \(\pi_t(x)/\pi_t(x)\))

(write as expectation over \(x\))
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x)
\]

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

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

\[
= \mathbb{E} \left[ \frac{\partial \pi_t(A_t)}{\partial H_t(a)} \right]
\]

\[
(\mathbb{E}[R_t|A_t] = q_*(A_t) \text{ and } B_t \equiv \tilde{R}_t)
\]
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x)
\]

(multiply by \(\pi_t(x)/\pi_t(x)\))

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

(write as expectation over \(x\))

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

(\(\mathbb{E}[R_t|A_t] = q_*(A_t)\) and \(B_t \doteq \bar{R}_t\))

\[
= \mathbb{E} \left[ (R_t - \bar{R}_t) \pi_t(A_t)([a = A_t]_1 - \pi_t(a)) / \pi_t(A_t) \right]
\]
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x)
\]

(multiply by \(\pi_t(x)/\pi_t(x)\))

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

(write as expectation over \(x\))

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

(\(\mathbb{E}[R_t|A_t] = q_*(A_t)\) and \(B_t \doteq \bar{R}_t\))

\[
= \mathbb{E}\left[\left(R_t - \bar{R}_t\right) \pi_t(A_t)([a = A_t]_1 - \pi_t(a))/\pi_t(A_t)\right]
\]

\[
= \mathbb{E}\left[\left(R_t - \bar{R}_t\right)([a = A_t]_1 - \pi_t(a))\right]
\]
\[
\frac{\partial \mathbb{E}[R_t]}{\partial H_t(a)} = \sum_x \pi_t(x)(q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} / \pi_t(x) \\
(\text{multiply by } \pi_t(x)/\pi_t(x))
\]
\[
= \mathbb{E}\left[ (q_*(A_t) - B_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right] \\
(\text{write as expectation over } x)
\]
\[
= \mathbb{E}\left[ (R_t - \bar{R}_t) \frac{\partial \pi_t(A_t)}{\partial H_t(a)} / \pi_t(A_t) \right] \\
(\mathbb{E}[R_t|A_t] = q_*(A_t) \text{ and } B_t = \bar{R}_t)
\]
\[
= \mathbb{E}\left[ (R_t - \bar{R}_t) \pi_t(A_t) ([a = A_t]_1 - \pi_t(a)) / \pi_t(A_t) \right]
\]
\[
= \mathbb{E}\left[ (R_t - \bar{R}_t) ([a = A_t]_1 - \pi_t(a)) \right]
\]
Thus:
\[
H_{t+1}(a) = H_t(a) + \alpha(R_t - \bar{R}_t) ([a = A_t]_1 - \pi_t(a)) \quad \Box
\]
$$\sum_x (q_\star(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} = \sum_x (q_\star(x) - B_t) \pi_t(x) ([a = x]_1 - \pi_t(a))$$

Baseline $B_t$ does not change expectation because:
\[
\sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} = \sum_x (q_*(x) - B_t) \pi_t(x) ([a = x]_1 - \pi_t(a))
\]

Baseline \( B_t \) does not change expectation because:

\[
\sum_x \left( q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} - B_t \frac{\partial \pi_t(x)}{\partial H_t(a)} \right) = \sum_x q_*(x) \frac{\partial \pi_t(x)}{\partial H_t(a)} - \sum_x B_t \frac{\partial \pi_t(x)}{\partial H_t(a)} = \ldots - B_t \sum_x \frac{\partial \pi_t(x)}{\partial H_t(a)} = 0 \text{ because...}
\]
$$\sum_x (q_*(x) - B_t) \frac{\partial \pi_t(x)}{\partial H_t(a)} = \sum_x (q_*(x) - B_t) \pi_t(x) ([a = x]_1 - \pi_t(a))$$

Baseline $B_t$ does not change expectation because:

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

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

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

$$= 0 \quad \text{because...}$$

$$\sum_x \pi_t(x) ([a = x]_1 - \pi_t(a))$$

$$= \sum_x \pi_t(x) [a = x]_1 - \sum_x \pi_t(x) \pi_t(a)$$

$$= \pi_t(a) - \sum_x \pi_t(x) \pi_t(a)$$

$$= \pi_t(a) - \pi_t(a) \sum_x \pi_t(x) = 0 \quad \text{because...}$$

$$= 1$$