COMS W4701: Artificial Intelligence

Lecture 14: Multi-armed Bandits

Christopher Lee

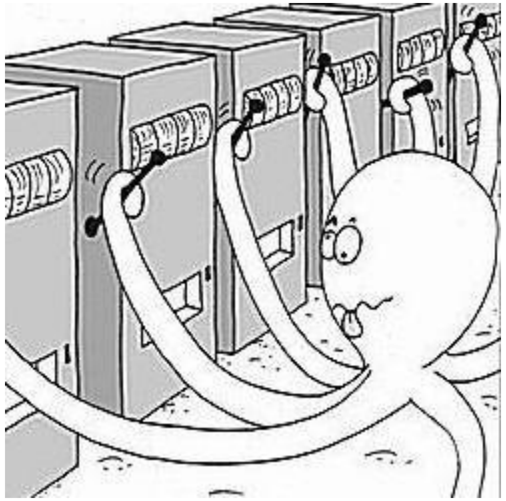
(slides adapted from Tony Dear)

Today

- Multi-armed bandit problems
- Exploration vs exploitation tradeoff
- ε-greedy methods
- Upper confidence bound
- Regret bounds

Multi-Armed Bandits

- Suppose we have K slot machines with different reward distributions
- We can only learn about the machine by trying them (taking actions)
- We want to maximize the overall rewards received



- Tradeoff between **exploration** and **exploitation**
  - Gather more information or maximize best rewards so far?
  - How to determine when current knowledge is good enough?

- Applications: Resource allocation for maximizing productivity, clinical trials to explore different treatments, financial portfolio design, recommendation systems

Action Values

- Suppose action (slot machine) $a \in A$ has unknown mean reward value μ_a
- We can define **action (Q) values** $Q_t(a)$ as *estimates* of each μ_a by averaging the rewards seen by step t

$$Q_t(a) = \frac{\text{sum of rewards from taking } a \text{ prior to } t}{\text{number of times taking } a \text{ prior to } t}$$

- In practice, we can use temporal difference with a fixed or variable learning rate (e.g., $\alpha = \frac{1}{N}$) to update the Q values as we see rewards

$$Q_{t+1}(a) = Q_t(a) + \alpha\big(r - Q_t(a)\big)$$

Initial Values and Exploration

- The choice of initial action values initially *biases* the estimates
- We can set them to reflect prior knowledge about rewards

- *Optimistic* initial values can be used to encourage exploration
- Set all initial Q-values much higher than 0, perhaps even higher than actual rewards

- Agent will initially explore more before action values are brought back down toward more accurate levels, even if we use a greedy policy

ε-greedy Action Selection

- Action selection should balance exploitation (maximizing Q) and exploration

- ***ε*-greedy**: *Exploit* and select $\mathrm{argmax}_a\big(Q(a)\big)$ *most* of the time, but with small probability ε, pick a random action to *explore* instead (may also include greedy action)

- For constant ε, every action will be sampled infinitely often
- In the limit, estimates $Q_t(a)$ will converge to μ_a (though limit may be very large!)

- ***ε*-first**: Set $\varepsilon = 1$ for a fixed number of trials, then set $\varepsilon = 0$ afterward
- ***ε*-decreasing**: Set ε to high initial value (e.g., 1) and decrease it over time

Regret

- We can characterize a bandit algorithm by its **regret**: Difference between cumulative maximum reward μ^* (from best action) and actual rewards received
- We generally want strategies that *minimize expected regret over T timesteps*

$$Regret_T = E\left(T\mu^* - \sum_{t=1}^{T} r_t\right) = \sum_{a \in A} N_a(\mu^* - \mu_a) = \sum_{a \in A} N_a \Delta_a$$

- We can also define regret in terms of the number of times each arm is taken

- Expected regret increases by the *suboptimality gap* Δ_a each time action a is taken

- But what if we don't know μ^*?

ε Regret Bounds

- No strategy can achieve zero regret on a bandit problem; some exploration is always required to learn and become confident about the reward distributions

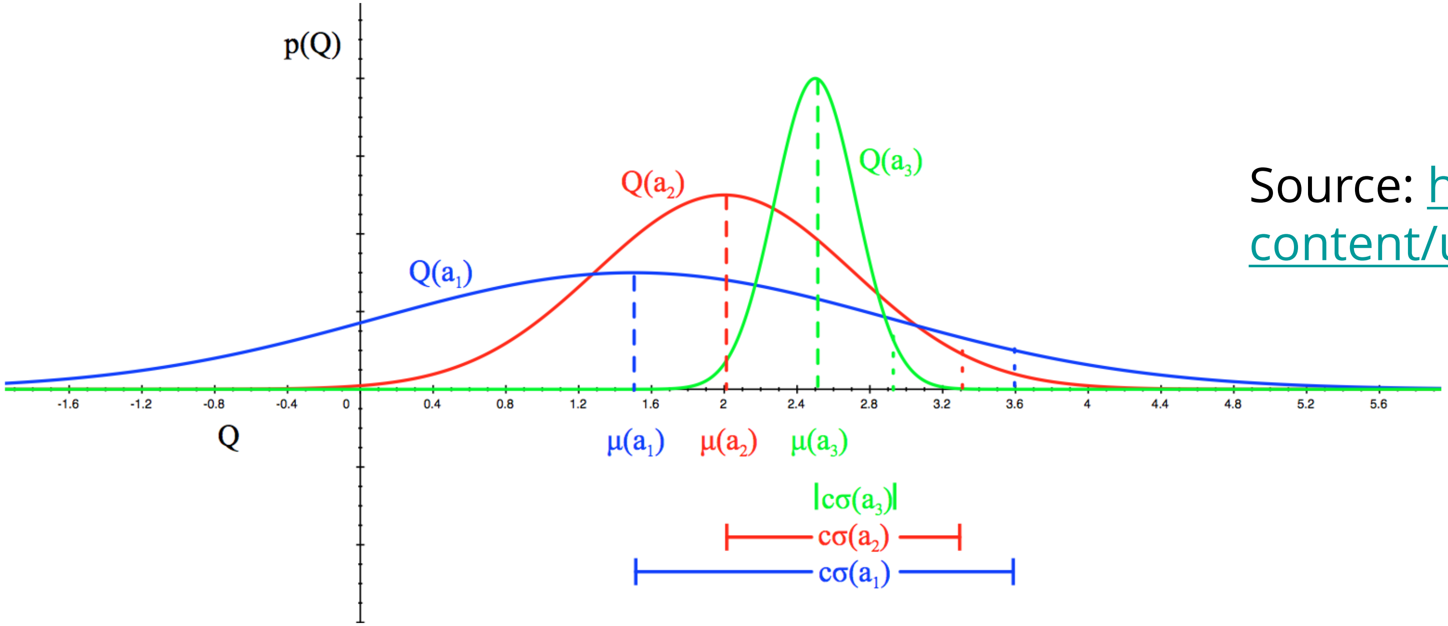
- In ε-greedy, probability of taking a suboptimal action in each time step is (at least) $\frac{\varepsilon}{|A|}$
- May be higher due to exploitation of a suboptimal action

- Expected regret in every time step is $\frac{\varepsilon}{|A|}\sum_a \Delta_a$—linear growth over time!

- With other methods, best case regret can grow more slowly on the order of $O(\log t)$ (Lai and Robbins, 1985)

Estimate Uncertainty

- ε methods only estimate value means, but not *uncertainty* (variance)
- Instead of exploring randomly, we can measure the uncertainty $U(a)$ of each action value estimate to perform "targeted" exploration



Source: https://www.davidsilver.uk/wp-content/uploads/2020/03/XX.pdf

- Exploitation-exploration tradeoff: Pick action that maximizes $Q(a) + U(a)$

Upper Confidence Bound

- **UCB1 algorithm** defines $U_t(a)$ as follows:

$$U_t(a) = c\sqrt{\frac{\ln t}{N_t(a)}}$$

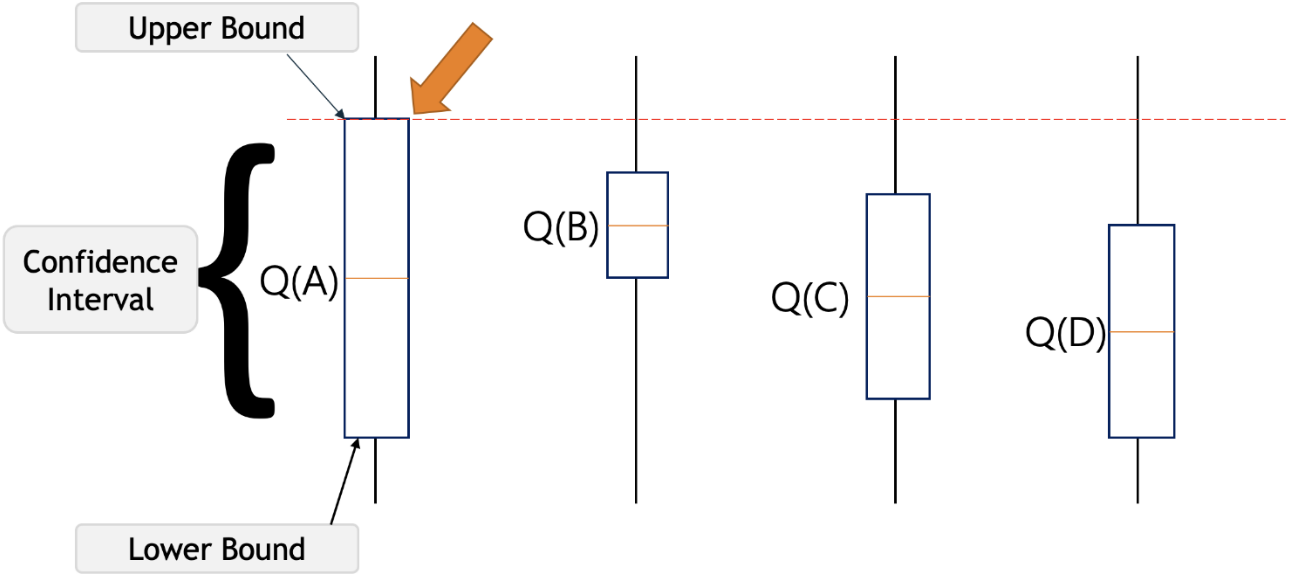
- At each step, pick action $\text{argmax}_a\big(Q(a) + U(a)\big)$
- $c \geq 0$: Tunable hyperparameter controlling exploration
- $N_t(a)$: Number of times action a taken prior to time t

- $1/\sqrt{N(a)}$ is proportional to standard deviation of $Q(a)$
- Initially large; decreases as a is repeatedly tried and we become confident
- $\ln t$ increases (slowly) over time; all actions tried infinitely often as $t \to \infty$

Optimism Under Uncertainty

- Maximizing $Q + U$ means that we are *optimistic under uncertainty*
- Higher uncertainty gives an action value a larger "bonus" for selection
- For UCB1, Hoeffding's inequality shows that the probability of the "error" being greater than $U(a)$ shrinks over time

$$\Pr[\mu_a - Q_t(a) > U_t(a)] \leq t^{-2c^2}$$



https://www.geeksforgeeks.org/upper-confidence-bound-algorithm-in-reinforcement-learning/

UCB1 Regret Bounds

- Can show that for UCB, suboptimal arm frequency $N_a(t)$ grows as $O(\log t)$
- Actual value of N_a is proportional to exploration parameter c and inversely proportional to suboptimality gap Δ_a

- Since number of tries of suboptimal arms grows as $\log t$, regret bound of UCB1 is also $O(\log t)$—better than ε-greedy!

- In practice, performance depends on c and problem difficulty
- UCB performs worse with more arms and/or smaller suboptimality gaps

General Bandit Algorithm Outline

Algorithm 1: General Bandit Algorithm Procedure

Initialize, for $i = 1$ to k:

$\quad Q_0(a_i) \leftarrow 0$

$\quad N_0(a_i) \leftarrow 0$

for $t = 1, 2, \ldots, \infty$ **do**

$\quad A_t \leftarrow \textsc{Choose-Action}\big(Q_{t-1}(a_1), Q_{t-1}(a_2), \ldots, Q_{t-1}(a_k)\big)$

$\quad R_t \leftarrow \textsc{Pull-Arm}\big(A_t\big)$

$\quad Q_t(A_t), N_t(A_t) \leftarrow \textsc{Update}\big(N_{t-1}(A_t), Q_{t-1}(A_t), R_t\big)$

end

Adapted from *Reinforcement Learning: An Introduction, 2nd ed. (Richard Sutton & Andrew Barto, 2020)*

Summary

- MAB problems model decision making in stochastic environments
- Fundamental tradeoff of exploration vs exploitation

- We can keep track of rewards and observations so far
- We can weight this info alongside uncertainty to determine our actions

- ε-greedy methods explore randomly with fixed or varying probability
- UCB1 is optimistic under uncertainty, choosing actions using a weighted balance between exploitation and exploration