Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 18

Solution by : Benyamin Naderi

Q1

Consider a stochastic multi-armed bandit (MAB) problem with 3 arms. The true (unknown) reward distributions are:

- Arm 1: Bernoulli($\mu_1 = 0.4$)
- Arm 2: Bernoulli($\mu_2 = 0.6$)
- Arm 3: Bernoulli($\mu_3 = 0.8$)

An algorithm runs for $T = 1000$ rounds and incurs an expected cumulative regret

$$R(T) = 120.$$

A) What does it mean for a bandit problem to be *stochastic*? How does this differ from an *adversarial* bandit?
B) Compute the per-round regret Δ_i for each suboptimal arm i. What is the theoretical lower bound for $R(T)$ in this setting?

Stochastic bandit: Each arm i has a fixed but unknown reward distribution (here Bernoulli(μ_i)). When an arm is pulled, the reward is an i.i.d. sample from that arm's distribution. The randomness is *statistical*, independent of the learner's past actions (other than the choice of arm).
Adversarial bandit: The rewards can be chosen by an adversary, possibly based on the learner's past actions. There is no fixed distribution; the adversary can sequence rewards to frustrate the algorithm. Stochastic algorithms aim for logarithmic regret, while adversarial algorithms guarantee $O(\sqrt{T})$ worst-case regret.

B) Per-Round Gaps and Theoretical Lower Bound

The best mean reward is

$$\mu^* = \max_i \mu_i = 0.8.$$

The per-round regret (gap) for each arm i is

$$\Delta_i = \mu^* - \mu_i.$$

$$\begin{aligned}\Delta_1 &= 0.8 - 0.4 = 0.4,\\ \Delta_2 &= 0.8 - 0.6 = 0.2,\\ \Delta_3 &= 0.8 - 0.8 = 0 \quad \text{(optimal arm)}.\end{aligned}$$

Instance-dependent asymptotic lower bound (Lai & Robbins):

$$\liminf_{T\to\infty} \frac{R(T)}{\ln T} \geq \sum_{i:\mu_i<\mu^*} \frac{\Delta_i}{\mathrm{KL}(\mathrm{Bern}(\mu_i)\|\mathrm{Bern}(\mu^*))},$$

where the KL divergence for Bernoulli distributions is

$$\mathrm{KL}(\mathrm{Bern}(p)\|\mathrm{Bern}(q)) = p\ln\frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}.$$

Compute KL divergences:
For $p = 0.4$, $q = 0.8$:

$$\mathrm{KL}(0.4\|0.8) = 0.4\ln(0.4/0.8) + 0.6\ln(0.6/0.2) \approx 0.382.$$

For $p = 0.6$, $q = 0.8$:

$$\mathrm{KL}(0.6\|0.8) = 0.6\ln(0.6/0.8) + 0.4\ln(0.4/0.2) \approx 0.105.$$

Coefficient in front of $\ln T$:

$$\frac{\Delta_1}{\mathrm{KL}(0.4\|0.8)} + \frac{\Delta_2}{\mathrm{KL}(0.6\|0.8)} \approx \frac{0.4}{0.382} + \frac{0.2}{0.105} \approx 2.958.$$

Asymptotic lower bound for $T = 1000$:

$$R(1000) \gtrsim 2.958 \cdot \ln 1000 \approx 2.958 \cdot 6.908 \approx 20.43.$$

Q2

The figure below shows reward probabilities (or empirical rewards) for two agents—a random agent and an optimal agent—across multiple arms in a Multi-Armed Bandit problem.

- Random agent rewards: $[0.33, 0.33, 0.33, 1.00, 0.16, 0.21, 0.65]$
- Optimal agent rewards: missing

A) Why might the random agent's rewards for the first three arms all be 0.33? What does this suggest about the agent's strategy?
B) The fourth arm has reward 1.00. If the optimal agent always chooses this arm, what would its long-term average reward converge to?

A) Random agent strategy
The first three arms all yield rewards of 0.33 for the random agent. This suggests that the agent is *choosing arms uniformly at random*, and the observed rewards are roughly equal due to averaging over multiple trials. This is consistent with a non-strategic exploration policy with no preference for high-reward arms.

B) Optimal agent long-term reward
The fourth arm has the highest reward (1.00). If the optimal agent always chooses this arm, its long-term average reward will converge to:

$$\mathbb{E}[r] = 1.00.$$

This represents the maximum possible expected reward per pull in this setting.

Q3

Explain what the learner should do if:
"The environment does not reveal the rewards of the arms not pulled by the learner."

When the environment only reveals rewards for the arms actually pulled, the learner faces a *partial feedback* or *bandit feedback* problem.
Recommended strategy:

- The learner should **explore** different arms to gather information about their expected rewards.
- Simultaneously, the learner should **exploit** arms that appear to have high rewards based on past observations.
- This requires an *exploration-exploitation trade-off*, typically handled by algorithms such as:
  - ϵ-greedy
  - Upper Confidence Bound (UCB)
  - Thompson Sampling

In other words, the learner must carefully balance trying new arms to learn their rewards while favoring arms that have historically performed well.