DDA4230 Reinforcement learning Stochastic multi-armed bandits

Lecture 4

Lecturer: Baoxiang Wang *Scribe: Baoxiang Wang*

1 Goal of this lecture

To understand the formulation of multi-armed bandits and some preliminaries needed to study the problem.

Students should get familiar with learning preliminaries when needed, instead of learning a lump-sum of techniques in the beginning.

Suggested reading: Chapter 2 of *Reinforcement learning: An introduction*; Chapter 1, 2, 3, 4, and 5 of *Bandit Algorithms*;

2 Multi-armed bandits

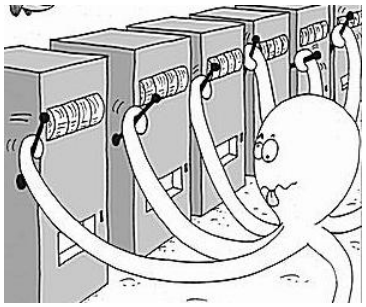

Figure 1: Multi-armed bandits.

The problem of multi armed bandits is a special case of the MDP we defined

- $\mathcal{S} = \{1\}$;
- $\mathcal{A} = [m] = \{1, 2, \ldots, m\}$;
- $\mathcal{T}(s, a) = 1$;
- $\mathcal{R}(s, a) = r(a)$ some unknown stochastic function $r(\cdot)$;
- $\rho_0 = 1$;
- $\gamma = 1$.

There are, however, some major differences in stochastic bandits compared with MDPs.

1. The MDP of multi-armed bandits has a finite horizon T.

2. The algorithms in bandits can leverage the knowledge that $\mathcal{S} = \{1\}$, $\mathcal{T}(s, a) = 1$, and $\mathcal{R}(s, a) = r(a)$. Though, $r(a)$ remains unknown.

3. Instead of obtaining an optimal expected return, the bandit algorithms aim to achieve asymptotically optimal expected return for $\lim_{T\to\infty}$.

Therefore, the optimal policy is a possibly stochastic policy that maps the time t to an action. This policy can be updated (i.e. reinforcement learning) through observing past rewards.

As the study focuses on the asymptotic performance, we mainly use the term regret instead of return to characterize the performance of an agent. The regret is defined as the difference between the maximum possible expected return and the expected return of the agent, as

$$\overline{R}_t = (t+1) \max_a \mathbb{E}[r(a)] - \mathbb{E}[\sum_{t'=0}^{t} r_{t'}].$$

In many bandit publications, the regret (instead of the return) is denoted by R_t.

It is common to denote $\mu_i = \mathbb{E}[r(i)]$ as the mean of the i-th arm's reward, and denote $\mu^* = \max_i \mu_i$ as the expected reward of an optimal arm. Also define $\Delta_i = \mu^* - \mu_i$, with which we rewrite $\overline{R}_t = \mathbb{E}[\sum_{t'=0}^{t} \sum_{i=1}^{m} \mathbb{1}\{a_{t'} = i\}\Delta_i]$. By letting $N_{t,i} = \mathbb{E}[\sum_{t'=0}^{t} \mathbb{1}\{a_{t'} = i\}]$, we alternatively write the regret into

$$\overline{R}_t = \sum_{i=1}^{m} N_{t,i}\Delta_i.$$

2.1 Type of feedback

Let's consider some (informal) examples of sequential decisions.

- Investment. Each morning, you choose one stock to invest into, and invest \$1. In the end of the day, you observe the change in value for each stock. Goal: maximize wealth.

- Dynamic pricing. A store is selling a digital good (e.g., an app or a song). When a new customer arrives, the store picks a price. Customer buys (or not) and leaves forever. Goal: maximize total profit.

- News site. When a new user arrives, the site picks a news header to show, observes whether the user clicks. Goal: maximize the number of clicks.

Immediately, we observe the difference in their feedback procedure and summarize it into the following table.

In fact, these examples correspond to the 3 types of feedback

- Full feedback. Reward is revealed for all arms (actions);

- Partial feedback. Reward is revealed for some but not necessarily for all arms;

- Bandit feedback. Reward is revealed only for the chosen arm.

Example	Action	Reward	Other feedback
Investment	a stock to invest into	change in value during the day	change in value for all other stocks
Dynamic pricing	a price p	p if sale; 0 otherwise	sale $\Rightarrow$ sale at any smaller price; no sale $\Rightarrow$ no sale at any larger price
News site	an article to display	1 if clicked, 0 otherwise	none

Table 1: Type of feedback in our examples.

Informally speaking, the type of feedback will decide the tool needed to study the problem. In full feedback problems like stock investing, our action has no impact on the information collected and thus needs only to focus on leveraging the collected information (i.e. exploitation). The problem is then studied by optimization algorithms and online optimization algorithms.

The problem with bandit feedback corresponds to the MDP we specified in this lecture notes. In this case, the agent needs to both exploit the historical information to choose high reward arms (exploitation) but also deploy action to collect more information (exploration). The exploration-exploitation tradeoff one of the most important problems in RL and bandits is a simple model to focus on this.

Interestingly, investment in high frequency can be a different story. If the market's response depends on the action taken, e.g. by placing a large order the agent can maneuver the price and the order book, the feedback becomes partial or even bandit. This makes high-frequency investment an application of RL.

2.2 Type of rewards

In our formulation of bandits, we assume the reward function to depend only on a, i.e. $\mathcal{R}(s, a) = r(a)$. This then prescribes the reward signal to be generated from $r_t \sim r(a_t)$ and each reward to be independent. The setting is the i.i.d. reward setting in bandits, known as stochastic bandits.

- i.i.d. rewards. The reward for each arm is drawn independently from a fixed distribution that depends on the arm but not on the round index t;
- Adversarial rewards. Rewards are chosen by an adversary;
- Strategic rewards. Rewards are chosen by an adversary with known constraints, such as reward of each arm can change by at most B from one round to another, reward of each arm can change by at most B from the original reward, or reward of each arm can change for at most B times;
- Stochastic rewards. Reward of each arm follows some stochastic process or random walk.

2.3 More applications of bandits

Table 2 illustrates more applications of bandits for reference.

Application	Action	Reward
medical trials	drug to give	health outcomes
internet ads	which ad to display	bid value if clicked, 0 otherwise
content optimization	e.g.: font color or page layout	clickthrough rate
sales optimization	which products to sell at which prices	revenue
recommendation systems	suggest a movie, restaurants, etc.	recommendation success rate
computer systems	which server(s) to route the job to	job completion time
crowdsourcing systems	which tasks to give to which workers; which price to offer?	quality of completed work; number of completed tasks
wireless networking	which frequency to use?	transmissions success rate
network routing	which path to transmit data	minimize package loss
robot control	a "strategy" for a given state and task	number of tasks successfully completed
game playing	an action for a given game state	game win rate
Bayesian optimization	the point to evaluate the function	optimality
Boolean satisfiability problem	the variable to toggle	correctly output the satisfiability
hyperparameter tuning	the set of parameter to continue training with	model performance

Table 2: Applications of bandits.

3 Some preliminaries

Studying RL and bandits requires the audience to refresh their background on probability and random variables, including discrete random variables and continuous random variables.

3.1 Concentration inequalities

Let $X_1, \ldots, X_n$ be independent random variables. These variables are not necessarily identically distributed. Let $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ denote the average. Then, the strong law of

large number indicates that when n approaches infinity,

$$\mathbb{P}(\overline{X} = E[\overline{X}]) = 1.$$

A concentration inequality is characterization of the convergence described by the strong law of large number, by bounding both the error term and the probability term in n.

$$\mathbb{P}(|\overline{X} - E[\overline{X}]| \leq \varepsilon(n)) \geq 1 - \delta(n),$$

where $\varepsilon(n)$ and $\delta(n)$ converge to 0 when n approaches infinity.

Hoeffding's inequality If each X_i is bounded and assume $0 \leq X_i \leq 1$ without loss of generality

$$\mathbb{P}(|\overline{X} - E[\overline{X}]| \leq \sqrt{\frac{\alpha \log T}{n}}) \geq 1 - 2T^{-2\alpha}.$$

A more general form of Hoeffding's inequality

$$\mathbb{P}(|\overline{X} - E[\overline{X}]| \leq t) \geq 1 - e^{-2nt^2}.$$

By restricting $X_i \in \{0, 1\}$, this inequality reduces to the Chernoff bound.

Acknowledgement

This lecture notes partially use material from *Reinforcement learning: An introduction* and *COMS E6998: Bandits and Reinforcement Learning* from Columbia. Figure 1 is from the blogpost *Efficient experimentation and the multi-armed bandit* by Ian Osband. Table 1 and part of Table 2 are from *CMSC 858G: Bandits, Experts, and Games*, University of Maryland, by Alex Slivkins. Proofread by Shaokui Wei.