Lecture 3 - Stochastic Multi-Armed Bandits

Guiliang Liu

The Chinese University of Hong Kong, Shenzhen

DDA4230: Reinforcement Learning
Course Page: [Click]

DDA 4230 Resources

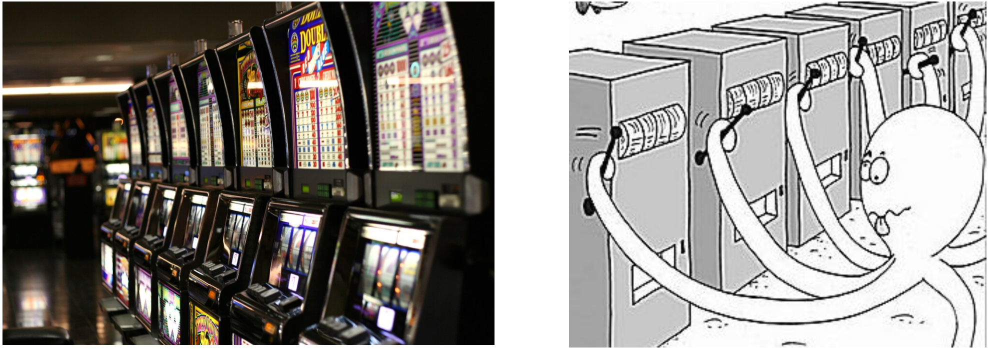
Please join our Slack group.

https://join.slack.com/t/slack-us51977/shared_invite/zt-22g8b40v8-0qSs9o0G3~8hXHwWydlCpw

Please check our course page.

https://guiliang.github.io/courses/cuhk-dda-4230/dda_4230.html

Multi-Armed Bandits



Multi-Armed Bandits

The problem of multi-armed bandits (MAB) is a special case of the MDP (focusing on exploration), we defined

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

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

Multi-Armed Bandits

The key properties of a MAB problem are:

- The reward functions $r(a)$ are not known a priori and can only be inferred using historical observations.
- The multi-armed bandit problem is a simple MDP with a dummy state while we investigate it with model-based methods, recall $\mathcal{S} = \{1\}$, $\mathcal{T}(s, a) = 1$, and $\mathcal{R}(s, a) = r(a)$.
- The MAB has a finite horizon T. the optimal policy $\pi(\cdot, t)$ maps the historical data and the time t to an action. Though, bandit algorithms aim to achieve asymptotically optimal expected return for $\lim_{T \to \infty}$.

Multi-Armed Bandits

The key properties of a MAB problem are:

- The optimal policy could be a stochastic policy that maps the historical data and the time t to an action.
- We can view the difference of $\pi(\cdot, t)$ and $\pi(\cdot, t+1)$ as if this policy is updated through historical data at time t.

Multi-Armed Bandits

The performance of an agent is characterized by the term regret: the difference between the maximum possible expected return and the expected return of the agent, as:

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

Remark:

1. $(t+1)\max_a \mathbb{E}[r(a)]$ is a constant.
2. Maximizing R_t (cumulative rewards) is equivalent to minimize $\overline{R}_t$.

Multi-Armed Bandits

- The mean of the reward of the i-th arm (action): $\mu_i = \mathbb{E}[r(i)]$.
- The expected reward of an optimal arm: $\mu^* = \max_i \mu_i$.
- The optimality action gap: $\Delta_i = \mu^* - \mu_i$ (unity loss due to sub-optimality).
- The natural filtration: $N_{i,t} = \sum_{t'=0}^{t} \mathbb{1}\{a_{t'} = i\}$.

Based on the aforementioned definitions, we alternatively write the regret into:

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

Some Examples of Bandits

- **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: to maximize wealth.

Example	Action	Reward	Full feedback
Investment	a stock to invest into	change in value during the day	change in value for all other stocks

Some Examples of Bandits

- **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: to maximize total profit.

Example	Action	Reward	Partial feedback
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

Some Examples of Bandits

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

Example	Action	Reward	Bandit feedback
News site	an article to display	1 if clicked, 0 otherwise	none

Type of Feedback

These examples correspond to the 3 types of feedback

- Full feedback. The reward is revealed for all arms;
- Partial feedback. The reward is revealed for some but not necessarily for all arms;
- Bandit feedback. The reward is revealed only for the chosen arm.

In a MAB problem, the agent needs to both:

- Exploit the historical information to choose high-reward arms (exploitation)
- Deploy actions to collect more information (exploration).

The exploration-exploitation tradeoff is most important in RL!

Type of Rewards

In our MAB, the reward function depends only on a, i.e. $\mathcal{R}(s, a) = r(a)$.

- Rewards that are i.i.d. 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 (Maximize $\overline{R}_t$).
- 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.
- Stochastic rewards. Reward of each arm follows some stochastic process or random walk.

Concentration Inequalities

The setting:

- Let $X_1, \ldots, X_n$ be independent random variables and assume that $\mathbb{E}[X_i]$ exists.
- 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,

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

A concentration inequality bounds both the error term and the probability term in the number n of samples:

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

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

Concentration Inequalities

Lemma (Chebyshev's inequality)

Let $X_1, \ldots, X_n$ *be i.i.d and assume that the variance* $\mathbb{V}[X_i] = \sigma^2$ *exists, then*

$$\mathbb{P}(|\overline{X} - \mathbb{E}[\overline{X}]| \leq z) \geq 1 - \frac{\sigma^2}{nz^2} .$$

Note that $\frac{\sigma^2}{nz^2}$ is $O(\frac{1}{n})$, not very ideal for RL.

Proof: See *Chebyshev's inequality* on Wikipedia.

Concentration Inequalities

Lemma (Hoeffding's inequality)

If $0 \leq X_i \leq c$ *for each* X_i*, then for*

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \leq z) \geq 1 - \exp(-\frac{2nz^2}{c^2}).$$

Note that $\exp(-\frac{2nz^2}{c^2})$ is $O(\frac{1}{e^n})$, better for RL.

Proof: See *Hoeffding's lemma* on Wikipedia.

Concentration Inequalities

Lemma (The Chernoff-Hoeffding inequality)

For $\alpha > 0$ *and* $t > 1$, *if* $X_i \sim \mathcal{N}(0, 1)$ *for each* X_i, *then for*

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

Concentration Inequalities

For random variables that are not necessarily identically distributed and not necessarily independent, similar results hold when the conditional expectations are constant.

Lemma (The Azuma-Hoeffding inequality)

For random variables $X_1, \ldots, X_n \in [0,1]$ *with constant conditional expectations* $\mu_i = \mathbb{E}[X_i \mid X_{i-1}, \ldots, X_1]$ *for* $i = 1, \ldots, n$, *then*

$$\mathbb{P}(|\overline{X} - \frac{1}{n}(\mu_1 + \cdots + \mu_n)| \leq \sqrt{\frac{\alpha \log t}{n}}) \geq 1 - 2t^{-2\alpha}.$$

Concentration Inequalities

Lemma (Bernstein's inequalities)

For independent Rademacher random variables $X_1, \ldots, X_n \in \{-1, 1\}$,

$$\mathbb{P}(|\overline{X}| \leq z) \geq 1 - 2\exp(-\frac{nz^2}{2(1+\frac{z}{3})}).$$

An alternative form of Bernstein's inequalities states that for Bernoulli random variables where the total variance $\sum_{i=1}^{n} \mathbb{V}[x_i \mid x_{i-1}, \ldots, x_1] = \sigma^2$, then

$$\mathbb{P}(\overline{X} - \mathbb{E}[\overline{X}] \leq z) \geq 1 - \exp(-\frac{n^2 z^2}{2\sigma^2 + nz}).$$

Tail bounds

Lemma (Gaussian tail bound)

If $X \sim \mathcal{N}(0,1)$, then for $x > 0$,

$$\frac{1}{\sqrt{2\pi}}(\frac{1}{x} - \frac{1}{x^3})\exp(-\frac{x^2}{2}) \leq \mathbb{P}(X \geq x) \leq \frac{1}{\sqrt{2\pi}x}\exp(-\frac{x^2}{2}) .$$

Tail bounds

Lemma (Gaussian tail bound)

For a σ^2*-sub-Gaussian random variable* X, *for* $z \geq 0$,

$$\mathbb{P}(X - \mathbb{E}[X] \leq z) \geq 1 - \exp(-\frac{z^2}{2\sigma^2}).$$

Question and Answering (Q&A)