DS 598
Introduction to RL

Xuezhou Zhang

Chapter 7: Exploration

(Modern Challenges)

Existing algorithms are very inefficient..

AlphaZero:

44,000,000 games

Human Pro Player:

~ 50,000 games

Existing algorithms are very inefficient..

1000 times less efficient than human!!

What is exploration?

- 4200 restaurants in Boston.
- Find your favorite one.
- What do you do?

How to quantify exploration efficiency?

- **Sample Complexity**: how many episodes do you need to find an ϵ-optimal policy?
- π is ϵ-optimal if $J(\pi^\star) - J(\pi) \leq \epsilon$.

- **Regret**: $\sum_{t=1}^{T} J(\pi^\star) - J(\pi_t)$.
- e.g. how many bad meals do you have to suffer.

- They are interchangeable to some extent (whiteboard).

Multi-armed Bandit – a.k.a. the Boston Restaurant problem

- K restaurants (arms): $a_1, \dots, a_K$
- Unknown reward distribution:
  - $r_k \sim \nu_k \in \Delta_{[0,1]}$ with mean $\mu_k = \mathbb{E}[r_k]$.
- Optimal arm: $k^\star = \mathrm{argmax}_k \mu_k$

- Interactive Learning Process:
- For $t = 1, \dots T$
  - Learner pulls arm $I_t \in \{1, \dots, K\}$.
  - Learner observes i.i.d. reward $r_t \sim \nu_{I_t}$.

Pure exploration

What are some naïve strategies?

Attempt 1: Uniform Exploration

- Try each restaurant n times. Estimate their reward. Pick the best one.
- **Uncertainty Estimation**: Hoeffding's Inequality

Given a distribution $\mu \in \Delta([0,1])$, and N i.i.d samples $\{r_i\}_{i=1}^{N} \sim \mu$, w/ probability at least $1-\delta$, we have:

$$\left| \sum_{i=1}^{N} r_i/N - \mu \right| \leq O\left(\sqrt{\frac{\ln(1/\delta)}{N}} \right)$$

- Total sample complexity to find an ϵ-optimal policy: $O(K/\epsilon^2)$
- Already pretty good!

Attempt 1: Uniform Exploration

Can we improve?

Some restaurants are obviously bad, no need to keep trying them!

Attempt 2: Arm Elimination

- Give up those arms that are clearly suboptimal.
- Gap: $\Delta_k = \mu^\star - \mu_k$

- Q: How many times will each arm be tried?
- A: Roughly $O\left(\frac{1}{\Delta_k^2}\right)$.

- Total sample complexity: $\sum_{\{\mathrm{k}|\ \Delta_k \geq \epsilon\}} \frac{1}{\Delta_k^2}$.

Regret Minimization

Minimize the regret of eating bad food.

Attempt 1: Greedy Algorithm

- Try each restaurant once.
- Going to the best restaurant I’ve been to previously.

- Problem: a good restaurant may give bad experience by chance.

- Missing the best restaurant forever!

- $O(T)$ regret!

Attempt 2: Explore and then Commit

- Do uniform exploration for N rounds per arm.
- Commit to the empirically best arm.

- What's the regret?
- Exploration stage: $O(NK)$
- Exploitation stage: $O\left(T\sqrt{\frac{1}{N}}\right)$

Attempt 2: Explore and then Commit

- Total regret: $O\left(NK + T\sqrt{\frac{1}{N}}\right)$
- Cauchy Schwarz Inequality:

$$NK + T\sqrt{\frac{1}{N}} \leq K^{1/3}T^{2/3}$$

- This is achieved by $N^* = \left(\frac{T}{K}\right)^{2/3}$

Regret Minimization

Can we do better than $O\left(K^{1/3}T^{2/3}\right)$?

Yes! Next time!