

Bandits

CS 486 / 686
University of Waterloo
Lecture 23: July 21, 2015

Exploration/Exploitation Tradeoff

- Fundamental problem of RL due to the active nature of the learning process
- Consider one-state RL problems known as **bandits**

Stochastic Bandits

- Formal definition:
 - Single state: $S = \{s\}$
 - A: set of actions (also known as **arms**)
 - Space of rewards (typically assumed to be $[0, 1]$)
- **No transition function to be learned** since there is a single state
- We simply need to **learn the stochastic reward function**

3

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Origin

- The term bandit comes from gambling where **slot machines can be thought as one-armed bandits**.
- Problem: which slot machine should we play at each turn when their payoffs are not necessarily the same and initially unknown?



4

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Examples

- Design of experiments (Clinical Trials)
- Online ad selection
- Games
- Networks (packet routing)

5

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Online Ad Optimization

The screenshot shows a web browser window with the URL <http://www.theglobeandmail.com/>. The browser's address bar and tabs are visible at the top. The main content area displays the Globe and Mail website. At the top of the page, there is a large advertisement for IBM with the headline "Can your business anticipate shifts in the marketplace?". Below this, there is a search bar and navigation links. The main content area features several news articles, including "Six Ontarians charged in alleged \$200-million investment fraud" and "122 'potential objects' spotted in ocean offer fresh jet lead". On the right side, there is a yellow advertisement for Porter's plans with the headline "ASK your Toronto City Councillor TO VOTE YES on April 1 for Porter's plans".

6

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Online Ad Optimization

- Problem: **which ad should be presented?**
- Answer: present ad with highest payoff

$$\text{payoff} = \text{clickThroughRate} \times \text{payment}$$

- Click through rate: probability that user clicks on ad
- Payment: \$\$ paid by advertiser
 - Amount determined by an auction

7

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Simplified Problem

- Assume payment is 1 unit for all ads
- Need to estimate click through rate
- Formulate as a bandit problem:
 - Arms: the set of possible ads
 - Rewards: 0 (no click) or 1 (click)
- In what order should ads be presented to maximize revenue?
 - **How should we balance exploitation and exploration?**

8

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Simple yet difficult problem

- Simple: description of the problem is short
- Difficult: **no known tractable optimal solution**

9

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Simple heuristics

- **Greedy strategy**: select the arm with the highest average so far
 - May get stuck in local optimum due to lack of exploration
- **ϵ -greedy**: select an arm at random with probability ϵ and otherwise do a greedy selection
 - Convergence rate depends on choice of ϵ

10

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Regret

- Let $R(a)$ be the unknown average reward of a
- Let $r^* = \max_a R(a)$ and $a^* = \operatorname{argmax}_a R(a)$
- Denote by $\operatorname{loss}(a)$ the **expected regret** of a
$$\operatorname{loss}(a) = r^* - R(a)$$
- Denote by Loss_n the **expected cumulative regret** for n time steps

$$\operatorname{Loss}_n = \sum_{t=1}^n \operatorname{loss}(a_t)$$

11

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Theoretical Guarantees

- When ϵ is constant, then
 - For large enough t : $\Pr(a_t \neq a^*) \approx \epsilon$
 - Expected cumulative regret: $\operatorname{Loss}_n = O(n)$
 - **Linear regret**
- When $\epsilon_t \propto 1/t$
 - For large enough t : $\Pr(a_t \neq a^*) \approx \epsilon_t = O\left(\frac{1}{t}\right)$
 - Expected cumulative regret: $\operatorname{Loss}_n = O(\log n)$
 - **Logarithmic regret**

12

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Empirical mean

- Problem: how far is the empirical mean $\tilde{R}(a)$ from the true mean $R(a)$?
- If we knew that $|R(a) - \tilde{R}(a)| \leq bound$
 - Then we would know that $R(a) < \tilde{R}(a) + bound$
 - And we could select the arm with best $\tilde{R}(a) + bound$
- Overtime, additional data will allow us to refine $\tilde{R}(a)$ and compute a tighter *bound*.

13

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Positivism in the Face of Uncertainty

- Suppose that we have an oracle that returns an **upper bound** $UB_n(a)$ on $R(a)$ for each arm based on n trials of arm a .
- Suppose the upper bound returned by this oracle converges to $R(a)$ in the limit:
 - i.e. $\lim_{n \rightarrow \infty} UB_n(a) = R(a)$
- **Optimistic algorithm**
 - At each step, **select** $\operatorname{argmax}_a UB_n(a)$

14

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Convergence

- **Theorem:** An optimistic strategy that always selects $\operatorname{argmax}_a UB_n(a)$ will converge to a^*
- Proof by contradiction:
 - Suppose that we converge to suboptimal arm a after infinitely many trials.
 - Then $R(a) = UB_\infty(a) \geq UB_\infty(a') = R(a') \forall a'$
 - But $R(a) \geq R(a') \forall a'$ contradicts our assumption that a is suboptimal.

15

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Probabilistic Upper Bound

- Problem: We can't compute an upper bound with certainty since we are sampling
- However we can obtain measures f that are upper bounds most of the time
 - i.e., $\Pr(R(a) \leq f(a)) \geq 1 - \delta$
 - Example: Hoeffding's inequality

$$\Pr\left(R(a) \leq \tilde{R}(a) + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n_a}}\right) \geq 1 - \delta$$

where n_a is the number of trials for arm a

16

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Upper Confidence Bound (UCB)

- Set $\delta_n = 1/n^4$ in Hoeffding's bound
- Choose a with highest Hoeffding bound

UCB(h)

$V \leftarrow 0, n \leftarrow 0, n_a \leftarrow 0 \quad \forall a$

Repeat until $n = h$

Execute $\operatorname{argmax}_a \tilde{R}(a) + \sqrt{\frac{2 \log n}{n_a}}$

Receive r

$V \leftarrow V + r$

$\tilde{R}(a) \leftarrow \frac{n_a \tilde{R}(a) + r}{n_a + 1}$

$n \leftarrow n + 1, n_a \leftarrow n_a + 1$

Return V

17

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

UCB Convergence

- Theorem: Although Hoeffding's bound is probabilistic, **UCB converges**
- Proof: As n increases, the term $\sqrt{\frac{2 \log n}{n_a}}$ increases, ensuring that all arms are tried infinitely often
- Expected cumulative regret: **$Loss_n = O(\log n)$**
– **Logarithmic regret**

18

CS486/686 Lecture Slides (c) 2015 Pascal Poupart

Summary

- Stochastic bandits
 - Exploration/exploitation tradeoff
- ϵ -greedy and UCB
 - Theory: logarithmic expected cumulative regret
- In practice:
 - UCB often performs better than ϵ -greedy
 - Many variants of UCB improve performance