Multi-armed Bandits
[Sutton and Barto, Reinforcement Learning, 2nd Edition, Chapter 2]

CS 486 / 686

University of Waterloo

Lecture 22: July 17, 2017

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

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?

Examples

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

Simple yet difficult problem

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

Simple heuristics

- Greedy strategy: select the arm with the highest average so far
  - May get stuck 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

Regret

- Let $R(a)$ be the unknown average reward of a
- Let $r^* = \max_a R(a)$ and $a^* = argmax_a R(a)$

- Denote by $loss(a)$ the expected regret of a

$$loss(a) = r^* - R(a)$$

- Denote by $Loss_n$ the expected cumulative regret for n time steps

$$Loss_n = \sum_{t=1}^{n} loss(a_t)$$

Theoretical Guarantees

- When ϵ is constant, then
  - For large enough t: $\Pr(a_t \neq a^*) \approx \epsilon$
  - Expected cumulative regret: $Loss_n = O(n)$
    - 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: $Loss_n = O(\log n)$
    - Logarithmic regret

Empirical mean

- Problem: how far is the empirical mean $\tilde{R}(a)$ from the true mean $R(a)$?

- If we knew that $\left|R(a) - \tilde{R}(a)\right| \leq bound$
  - 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 $bound$.

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. $\lim_{n\to\infty} UB_n(a) = R(a)$
- Optimistic algorithm
  - At each step, select $argmax_a \; UB_n(a)$

Convergence

- Theorem: An optimistic strategy that always selects $\mathrm{argmax_a} UB_n(a)$ 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.

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., $\Pr\left(R(a) \leq f(a)\right) \geq 1 - \delta$
  - 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

Upper Confidence Bound (UCB)

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

UCB(h)

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

Repeat until $n = h$

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

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

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

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