Multi-armed Bandits - Introduction

Henry Vu

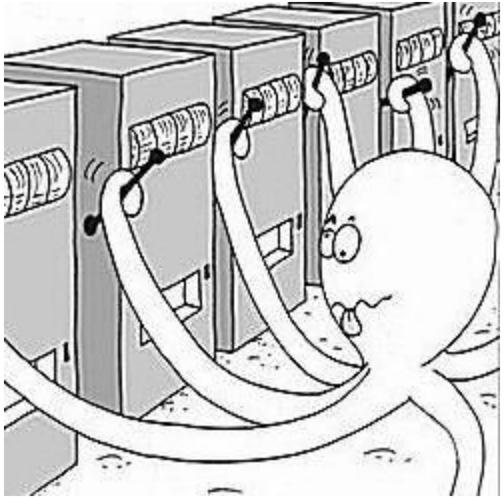
Jan 22, 2024

Outline

- Introduction
- Problem formulation
- Regret
- Stochastic bandit
- Some algorithms

Bandit Problem

- A sequential allocation problem defined by a set of actions. At each time step, a unit resource is allocated to an action and some payoff is obtained. The goal is to maximize the total payoff.
- Bandit is an instance of sequential decision making with limited information, and natural addresses the fundamental trade-off between **exploration** and **exploitation**



Exploration vs Exploitation

- Fundamental problem of Reinforcement Learning
  - exploit what has already been learned
  - explore to learn which actions give best reward
- Bandit could be considered a one-state RL problem, giving a simplified formulation of the **exploration-exploitation trade-off**.

Applications

- **Clinical Trials**: which drug to prescribe ⇒ health outcome
- **Online ad placement**: which ad to display ⇒ revenue from ads
- **Recommender system**: which movie to watch ⇒ revenue from users
- **Internet**: which TCP settings to use ⇒ connection quality
- **Games**: which version of the game to release ⇒ user engagement
- …

Problem Formulation

Given $A = \{1, 2, \ldots, K\}$ the set of action and (possibly) number of rounds $n \geq K$

for $t = 1, 2, \ldots, n$ **do**:

Algorithm pulls arm $I_t \in A$

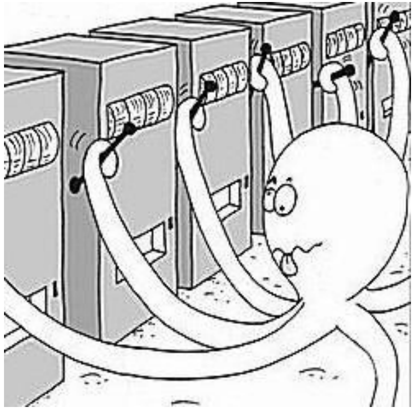
A reward vector $(X_{1, t}, X_{2, t}, \ldots, X_{n, t})$ is generated, usually scaled to $[0, 1]$

Algorithm observes reward $X_{It, t}$

end for

Goal: Minimizing the regret

Simple formulation, but no known tractable optimal solution



Regret

- **Regret**: The difference between cumulative reward after n time steps between the algorithm and an optimal strategy that consistently plays the best arm. Measures how much the algorithm "regrets" not knowing the best arm.

- The regret R_n after n plays $I_1, I_2, \ldots, I_n$ is defined by

$$R_n = \max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t} \,.$$

Regret (cont.)

- In general, reward can bet *stochastic* with regard to the draw of the rewards $X_{i,t}$. The expected regret of the algorithm over n rounds i given by

$$\mathbb{E}\, R_n = \mathbb{E}\left[\max_{i=1,\dots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]$$

- In practice, the expected regret is hard to work with. Therefore, we often minimize the pseudo-regret instead. The pseudo-regret is given by

$$\overline{R}_n = \max_{i=1,\dots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]$$

Regret (cont.)

- Pseudo-regret is a weaker notion of regret and is upper-bounded by expected regret, i.e. $\widetilde{R}_n \leq \mathbb{E}R_n$

- Proof:

$$\begin{aligned}
\widetilde{R}_n &= \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right] \\
&\leq \mathbb{E}\left[\max_{i=1,\ldots,K}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right]\right] \\
&= \mathbb{E}\left[\max_{i=1,\ldots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t,t}\right] \\
&= \mathbb{E}R_n
\end{aligned}$$

Stochastic Multi-armed Bandits

- We assume that there is no dependency among arms. The rewards of arm i are i.i.d according to a fixed probability distribution v_1, v_2, …, v_K on [0, 1]. These distributions are <u>unknown</u> to the algorithm

- Let:

$$\mu^* = \max_{i=1,\ldots,K} \mu_i \qquad \text{and} \qquad i^* \in \operatorname*{argmax}_{i=1,\ldots,K} \mu_i \ .$$

- In the stochastic setting, pseudo-regret can be written as

$$\widetilde{R}_n = n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t}\right]$$

Assumptions

- Mean reward for each arm are **unknown**, it is necessary to make some assumptions:
  - Independent arms
  - Stationary reward distribution
  - Bounded rewards
  - I.I.D, adversarial rewards
  - Reward distribution is Bernoulli, Gaussian, etc.

Regret: Another Perspective

Let $\Delta_i = \mu^* - \mu_i$, and let $T_i(s)$ denote the number of times the algorithm chose arm i on the first s rounds.

- Regret is also a function of $T_i(s)$ and Δ_i.

- Proof:

$$\overline{R}_n = \left(\sum_{i=1}^{K} \mathbb{E}\, T_i(n)\right)\mu^* - \mathbb{E}\sum_{i=1}^{K} T_i(n)\mu_i = \sum_{i=1}^{K} \Delta_i\, \mathbb{E}\, T_i(n)$$

Some heuristics

- **Naive**:

  Greedily plays the arm with the highest empirical mean ⇒ may get stuck due to lack of exploration, regret is linear n.

  Play all arms an equal number of times ⇒ pure exploration, regret is linear in n

- **e-greedy**:

  Exploitation: greedily plays the arm with the highest empirical mean (observed rewards) so far with probability 1-ϵ,

  Exploration: plays a random arm (including empirically best arm) with probability ϵ.

ϵ-Greedy

- With an infinite number of time steps, the probability of picking a suboptimal arm is ~ $\epsilon(k-1)/k \Rightarrow \mathbb{E}[T_i(n)] = n\epsilon(k-1)/k \Rightarrow$ Regret is linear in n.

- Choose ϵ_t ~ 1/t at each time step t
  Regret ~ O(log(n)) ⇒ Logarithmic regret.

Finite-time Analysis of the Multiarmed Bandit Problem*

PETER AUER pauer@igi.tu-graz.ac.at
University of Technology Graz, A-8010 Graz, Austria

NICOLÒ CESA-BIANCHI cesa-bianchi@dti.unimi.it
DTI, University of Milan, via Bramante 65, I-26013 Crema, Italy

PAUL FISCHER fischer@ls2.informatik.uni-dortmund.de
Lehrstuhl Informatik II, Universität Dortmund, D-44221 Dortmund, Germany

Hoeffding's Inequality

Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Consider the sum of these random variables,

$$S_n = X_1 + \cdots + X_n.$$

Then Hoeffding's theorem states that, for all $t > 0$,[3]

$$\mathrm{P}(S_n - \mathrm{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

$$\mathrm{P}(|S_n - \mathrm{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Here $\mathrm{E}[S_n]$ is the expected value of S_n.

- Here, $a_i = 0$ and $b_i = 1$ for all i. S_n and $\mathbb{E}[S_n]$ can be scaled by 1/n to represent the empirical mean and actual mean.
- n (refers to the n from the image above) is the number of times each arm is pulled.

Concentration Inequalities

- Provide mathematical bounds on the probability of a r.v X deviating from some value (usually $\mathbb{E}[X]$)
  - Azuma-Hoeffding's inequality
  - Chebyshev's inequality
  - Markov's inequality
  - …

Next Week

- Optimism in the face of uncertainty
- Upper Confidence Bound (UCB) algorithm

Thank you!