EE675 - IIT Kanpur, 2022-23-II

Lecture 2: n-Armed Bandits

11.01.23

Lecturer: Prof. Subrahmanyu Swamy Peruru Scribe: Ananya Mehrotra and Shamayeeta Dass

1 Recap

1.1 Paradigms of Machine Learning and Basic Terminology

There are three paradigms of Machine Learning:

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

We primarily focus on Reinforcement Learning (RL) in this course. Some common terms used in Reinforcement Learning are described below:

- **State** : Representation of the environment of the task. Commonly denoted as S_t
- **Action** : Refers to the reaction to State. Commonly denoted as A_t
- **Environment** : The combination of the new State and Reward to the previous action
- **Reward** : Feedback to the User in correspondence with the previous action and the new state. Commonly denoted as R_t
- **Return** : Cumulative Reward. It is equivalent to $\sum_{t=1}^{T} R_t$
- **Policy** : Methodology which evaluates history and predicts best possible action. A policy can be segregated into the following two categories:
  - **Deterministic** : The output is a single action which is predicted to be the best possible choice
  - **Stochastic** : The output is a set of probabilities of the actions. As per the policy, the higher the probability of a particular action, the better it is.

The main objective of an RL algorithm is to maximize the cumulative reward over several rounds.

2 Basics of n-Armed Bandits

n-Armed Bandits is a special scenario of Reinforcement Learning where there is no time sequence dependence. In simpler terms, n-Armed Bandits are Single State Scenarios. A single action leads to a reward with no intermediate states, after which the environment resets, and the process repeats.
Some Common Notations :

- **Arms** : Refer to actions taken by the user. The i^{th} arm is denoted by a_i, and the total number of arms is taken to be K

- **Expectation of a Random Variable 'X'** : Denoted by $E[X]$.

$$E[X] = \int x f_X(x) dx$$

  where $f_X(x)$ refers to the Probability Density Function of a continuous RV 'X'

- **True means of arms** : Denoted by μ_i for i^{th} arm. We know,

$$E[R_t | A_t = a_i] = \mu_i$$

  The best true mean is given by μ^* such that,

$$\mu^* = \max_i \mu_i$$

- **Estimated Sample Averages**: Denoted by $\bar{\mu_i}$ for i^{th} arm.

There can be two main objectives in a Multi-Armed Bandit problem:

- **Best Arm Identification** : This methodology focuses more on exploration. Over T rounds, P(Identified Arm is the Optimal Arm) is maximised to help identify the optimal arm

- **Regret Minimzation** : This methodology minimzes the Expected Regret.
  The Expected Regret is

$$\mu^* T - \sum_{i=1}^{\boldsymbol{T}} \mu_i(a_t)$$

  and, the Actual Regret is

$$\mu^* T - \sum_{t=1}^{\boldsymbol{T}} R_t$$

  where $\mu(a_t) = E[R_t | a_t]$ and $\mu_i = E[R_t | a_t = a_i]$

2.1 Algorithms

There are two main algorithms for a Multi-Armed Bandit problem.

2.1.1 Explore Then Commit Algorithm

In this methodology, we explore all arms uniformly and select the arm with the highest sample average reward.

Algorithm:

1. Explore each arm N times.
2. Pick arm $\hat{a}$ with the highest sample average mean.
3. Play arm $\hat{a}$ in all remaining rounds.

If we take a large number of samples, the sample mean will be equal to the expected mean. In our setting, the average reward $\tilde{\mu}(a)$ for any action a should be close to the expected reward $\mu(a)$. We make use of the following inequality to set a bound.

Hoeffding's Inequality: *It states that for N sample means and* $\forall\ \epsilon$*,*

$$P[|\bar{\mu}(a) - \mu(a)| \geq \epsilon] < 2e^{-2\epsilon^2 N}$$

2.1.2 Regret Minimisation in ETC Algorithm:

In Hoeffding's Inequality, we assume that $R_t \in [0, 1]$. As we want the probability to be small, we take $\varepsilon \sim \sqrt{\frac{2\log(T)}{N}}$. This leads to the upper bound in the above expression to be of the order: $\mathcal{O}(\frac{1}{T^4})$.

Let us take an example of K = 2 arms. This indicates one of the arms will be the optimal arm, and the other one will be a sub-optimal arm.

Let a^* denote the best arm. If we choose the sub-optimal arm, then it must have been because the sub-optimal arm had a higher sample mean, $\bar{\mu}(a) > \bar{\mu}(a^*)$. We also know by using Hoeffding's Inequality:

$$\mu(a) + \epsilon \geq \bar{\mu}(a) > \bar{\mu}(a^*) \geq \mu(a^*) - \epsilon$$

Thus,

$$\mu(a^*) - \mu(a) \leq 2\epsilon$$

In order to analyse regret: $R(T)$, we need to divide it into two parts -

- **Exploration Regret** :If we sample every arm N times, we will have a term of order N for sampling the sub-optimal arm in every round.
- **Exploitation Regret** : In the remaining $T - 2N$ rounds, we will have a regret term of 2ϵ in each round for picking the sub-optimal arm.

Hence,

$$R(T) \leq N + 2\epsilon(T - 2N) < N + 2\epsilon T$$
$$R(T) < N + 2T\sqrt{2\log T/N}$$

For $N = T^{2/3}(\log T)^{1/3}$, the upper bound in Hoeffding's Inequality is minimum. Thus, we get

$$R(T) \leq O(T^{2/3}(\log T)^{1/3})$$

Let A be the event that Hoeffding's Inequality holds for every arm and A' be its complementary event. Then, the Expected Regret can be expressed as

$$E[R(T)] = E[R(T)|A]P(A) + E[R(T)|A']P(A')$$

If Hoeffding's inequality doesn't hold, then a regret term of order $1/T^4$ is considered for each round. As $P(A) \leq 1$

$$E[R(T)] \leq R(T) + T.\mathcal{O}(1/T^4)$$
$$E[R(T)] \leq \mathcal{O}(T^{2/3}(\log T)^{1/3}$$

2.1.3 Epsilon-greedy Algorithm

Algorithm

1. Toss a coin that lands in the head with a probability of ϵ.
2. If coin lands in head pick any arm at random, else pick the arm with best sample average.

If the Exploration Probability $\epsilon_t \approx t^{-1/3}$, then

$$E[R(t)] \leq t^{2/3}\mathcal{O}((K\log t)^{1/3})$$

References

[1] A. Slivkins. *Introduction to Multi-Armed Bandits*. Foundations and Trends in Machine Learning, 2022.

[2] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2020.

[1] [2]