2022

Lecture 9: Stochastic Multi-Armed Bandits (Part I)

Week 9

Lecturer: Tianyu Wang

1 The Stochastic Multi-Armed Bandit Problem

Consider K unknown distributions supported on $[0, 1]$. Task: for $t = 1, 2, \cdots$, pick $I_t \in [K]$. Observe a random sample from the I_t-th distribution, and call it $Y_{I_t,t}$. We want to find a rule for picking distributions (a policy), such that, for any T (or for a given T), $\mathbb{E}\left[\sum_{i=1}^{T} Y_{I_t,t}\right]$ is maximized.

One of the biggest motivation is medical trials. Consider testing K different new medicines. We try them one by one and observe the treatment effect. We want to maximize the overall treatment effect.

In the literature, the K distributions are called arms. Selecting a distribution is called playing/pulling an arm. The sample $Y_{I_t,t}$ is called the reward/payoff at time t. The selection rule I_t is called policy.

2 Martingale processes and Azuma-Hoeffding inequality

Let's pause the discussion on bandit problems for a moment and put forward some concepts from probability theory.

Definition 2.1 (probability space)**.** A probability space is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ consisting of:

- the sample space Ω, which is an arbitrary non-empty set;
- the σ-algebra $\mathcal{F} \subseteq 2^{\Omega}$ (also called σ-field) – a set of subsets of Ω, called events, such that:
  - $\mathcal{F}$ contains the sample space: $\Omega \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under complements: if $A \in \mathcal{F}$, then $(\Omega \setminus A) \in \mathcal{F}$,
  - $\mathcal{F}$ is closed under countable unions: if $A_i \in \mathcal{F}$ for $i = 1, 2, \ldots$, then $(\bigcup_{i=1}^{\infty} A_i) \in \mathcal{F}$.
- the probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ – a function on $\mathcal{F}$ such that:
  - $\mathbb{P}$ is countably additive (also called σ-additive): if $\{A_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$ is a countable collection of pairwise disjoint sets, then $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$,
  - the measure of entire sample space is equal to one: $\mathbb{P}(\Omega) = 1$.

Definition 2.2 (filtered probability space)**.** A probability space is a tuple $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n=1}^{\infty}, \mathbb{P})$ such that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathcal{F}_n$ is a σ-algebra for all $n = 1, 2, \cdots$, and $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots \subseteq \mathcal{F}_\infty = \mathcal{F}$. The sequence $\{\mathcal{F}_t\}_{t=1}^{\infty}$ is called a filtered σ-algebra (or a filtration).

The notion of filtered probability space is useful in modelling stochastic processes. For example, let $X_1, X_2, X_3, \cdots$ be a sequence of $i.i.d.$ coin tosses. Let $\mathcal{F}_n$ be the space of all possible outcomes of the first n coin tosses. Then $\mathcal{F}_n$ is a σ-algebra.

Definition 2.3. Let $\{\mathcal{F}_t\}_{t=1}^{\infty}$ be a filtered σ-algebra. A sequence of random variables $\{X_t\}_{t=1}^{\infty}$ is adapted to $\{\mathcal{F}_t\}_{t=1}^{\infty}$ if $X_t \in \mathcal{F}_t$ for all $t = 1, 2, \cdots$. The condition $X_t \in \mathcal{F}_t$ is also called X_t is $\mathcal{F}_t$-measurable.

One can quickly see that a sequence of random variables $X_1, X_2, \cdots$ is adapted with the filtration generated by itself. Soon we will see that notion of filtration is extremely useful in describing the decision making processes.

Definition 2.4 (conditional expectation)**.** Consider

- $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space,
- $X \colon \Omega \to \mathbb{R}^n$ a random variable on that probability space with finite expectation,
- $\mathcal{H} \subseteq \mathcal{F}$ a sub-σ-algebra of $\mathcal{F}$.

A conditional expectation of X given $\mathcal{H}$, denoted as $\mathbb{E}[X|\mathcal{H}]$, is any $\mathcal{H}$-measurable function (meaning defined on a subset $\mathcal{H}$) that satisfies:

$$\int_H \mathrm{E}[X|\mathcal{H}] \, \mathrm{d}\mathbb{P} = \int_H X \, \mathrm{d}\mathbb{P}$$

for all $H \in \mathcal{H}$.

Consider the following example. If X and Y are discrete random variables, the conditional expectation of X given Y is

$$\begin{aligned} \mathbb{E}[X|Y \in A] &= \sum_x x \mathbb{P}(X = x \mid Y \in A) \\ &= \sum_x x \frac{\mathbb{P}(X = x \text{ and } Y \in A)}{\mathbb{P}(Y \in A)}. \end{aligned}$$

If $\mathcal{H}$ is the σ-algebra generated by Y, then $\mathbb{E}[X|\mathcal{H}]$ is a function that equals $\mathbb{E}[X|Y \in A]$ for every $A \in \mathcal{H}$ unless A occurs with zero probability.

Definition 2.5 (martingale)**.** A process (a sequence of random variables) $\{M_n\}_{n=1}^{\infty}$ is a martingale with respect to a filtration $\{\mathcal{F}_n\}_{n=1}^{\infty}$ if

- $\{M_n\}_{n=1}^{\infty}$ is adapted to $\{\mathcal{F}_n\}_{n=1}^{\infty}$,

- M_n is absolutely integrable: $\mathbb{E}|M_n| < \infty$ for all n.
- $\mathbb{E}[M_n|\mathcal{F}_{n-1}] = M_{n-1}$ for all $n = 2, 3, 4, \cdots$.

Theorem 2.6 (Azuma-Hoeffding). *Suppose $\{M_k\}_{k=0}^{\infty}$ is a martingale (adapted to $\{\mathcal{F}_k\}_{k=0}^{\infty}$ for some filtration $\{\mathcal{F}_k\}_{k=0}^{\infty}$) and*

$$|M_k - M_{k-1}| \leq c_k,$$

almost surely (meaning with probability 1), for $c_k \in \mathcal{F}_{k-1}$. Then for all positive integers N and all positive reals ϵ,

$$\mathbb{P}(|M_N - M_0| \geq \epsilon) \leq 2\exp\left(\frac{-\epsilon^2}{2\sum_{k=1}^{N} c_k^2}\right).$$

One can use Hoeffding's inequality is a special case.

Theorem 2.7 (Hoeffding's inequality). *Let $X_1, \cdots, X_n$ 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 for all $\epsilon > 0$*

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

Let X_i and S_n be defined as in Theorem 2.7. Let $M_0 = 0$ and let $M_k = S_k - \mathbb{E}[S_k]$. Then the sequence $M_0, M_1, M_2, \cdots$ is a martingale sequence (adapated to the filtration generated by $X_1, X_2, \cdots$). Applying Azuma-Hoeffding inequality recovers the Hoeffding's inequality.

By Hoeffding's inequality, we can estimate the means of the distributions via observed samples.

2.1 Back to bandit problems

Let μ_i be the mean of the i-th distribution. Let $\widehat{\mu}_{t,i}$ be the estimator of the mean of the i-th distribution at time t, which is defined as

$$\widehat{\mu}_{t,i} = \frac{\sum_{s=1}^{t} Y_{I_s,s}\mathbb{I}_{[I_s=i]}}{\sum_{s=1}^{t}\mathbb{I}_{[I_s=i]}}.$$

We will use the Azuma-Hoeffding inequality to derive concentration bounds for $\widehat{\mu}_{t,i}$.

Fix any i, consider the random variables $\{Y_{I_s,s}\mathbb{I}_{[I_s=i]}\}_{s=1}^{\infty}$, and the filtration $\{\mathcal{F}_s\}_{s=1}^{\infty}$ where $\mathcal{F}_s = \sigma\left(\{Y_{r,1}, Y_{r,2}, \cdots, Y_{r,K}\}_{r=1}^{s}, \{I_r\}_{r=1}^{s}\right)$ (the σ-algebra generated by all possible outcomes of $\{\{Y_{r,1}, Y_{r,2}, \cdots, Y_{r,K}\}_{r=1}^{s}$ and $\{I_r\}_{r=1}^{s}$). Although at any s, only one of $Y_{s,1}, Y_{s,2}, \cdots, Y_{s,K}$ is observed, the σ-algebras consider all possible outcomes of all distributions.

Fix any $i \in \{1, 2, \cdots, K\}$. Consider the martingale $M_0 = 0$, and

$$M_t = \sum_{s=1}^{t} Y_{I_s,s}\mathbb{I}_{[I_s=i]} - \mu_i \sum_{s=1}^{t}\mathbb{I}_{[I_s=i]}.$$

We know that

$$|M_t - M_{t-1}| = |Y_{I_t,t}\mathbb{I}_{[I_t=i]} - \mu_i \mathbb{I}_{[I_t=i]}| \leq \mathbb{I}_{[I_t=i]},$$

where the inequality uses that $Y_{I_t,t}$ is supported on $[0,1]$ for all t.

If the rule I_s is $\mathcal{F}_{s-1}$-measurable (meaning the decision at s is made with observations up to $s-1$), we can apply the Azuma-Hoeffding inequality to $\{M_t\}_{t=1}^{\infty}$, and get, for any t and ϵ,

$$\mathbb{P}\left(\left|\sum_{s=1}^{t} Y_{I_s,s}\mathbb{I}_{[I_s=i]} - \mu_i \sum_{s=1}^{t} \mathbb{I}_{[I_s=i]}\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2\sum_{s=1}^{t}\mathbb{I}_{[I_s=i]}}\right).$$

Define $n_{t,i} = \sum_{s=1}^{t} \mathbb{I}_{[I_s=i]}$, we can write the above as

$$\mathbb{P}\left(\left|\sum_{s=1}^{t} Y_{I_s,s}\mathbb{I}_{[I_s=i]} - \mu_i n_{t,i}\right| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2n_{t,i}}\right), \quad \forall \epsilon > 0, t \in \mathbb{N}_+,$$

or equivalently

$$\mathbb{P}\left(|\widehat{\mu}_{t,i} - \mu_i| \geq \frac{\epsilon}{n_{t,i}}\right) \leq 2\exp\left(\frac{-\epsilon^2}{2n_{t,i}}\right), \quad \forall \epsilon > 0, t \in \mathbb{N}_+.$$

Letting $\delta = 2\exp\left(-\frac{n_{t,i}\epsilon^2}{2}\right)$ gives

$$\mathbb{P}\left(|\widehat{\mu}_{t,i} - \mu_i| \geq \sqrt{\frac{2\log(2/\delta)}{n_{t,i}}}\right) \leq \delta, \quad \forall \delta > 0, t \in \mathbb{N}_+. \tag{1}$$

We have built the concentration inequality for the decision making process in (1).

3 Regret, a Naive Attempt, and the Exploration-Exploitation Tradeoff

Recall that in stochastic MAB problems, we need to maximize the total expected reward $\mathbb{E}\left[\sum_{t=1}^{T} Y_{I_t,t}\right]$ up to time T. Equivalently, we can minimize the regret up to time T

$$Reg(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} Y_{I_t,t}\right] = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right],$$

where μ_i is the expectation of arm i (distribution i) and μ_* is the expectation of the optimal arm (distribution).

With the inequality (1), finding the optimal arm is simple, but minimizing regret is not so simple. To find the optimal arm, we can play the arms one by one, cyclically. At time T, the policy plays each arm $\frac{T}{K}$ times, and the for all i,

$$\mathbb{P}\left(|\widehat{\mu}_{T,i} - \mu_i| \geq \sqrt{\frac{2K\log(2/\delta)}{T}}\right) \leq \delta, \quad \forall \delta > 0, T \in \mathbb{N}_+.$$

When T is large, we can for sure find the optimal arm. But what about the regret? At any T, the regret of this policy is

$$Reg(T) = T\mu_* - \mathbb{E}\left[\sum_{t=1}^{T} \mu_{I_t}\right] = T\mu_* - \sum_{i=1}^{K} \frac{T\mu_i}{K} = T\left(\mu_* - \sum_{i=1}^{K} \frac{\mu_i}{K}\right),$$

which grows linearly in T. In terms of regret, this kind of policy is as terrible as never playing the optimal arm!

To ensure a sub-linear regret rate, we need to (1) explore all arms, so that the uncertainty about all arms are decreasing, and (2) exploit the good arms according to past observations, so that with high probability we are frequently playing the best arm. This dilemma is called the exploration-exploitation tradeoff, and it's a key dilemma in almost all online decision making processes.

Acknowledgement

The stochastic MAB problem is due to Lai & Robbins.