Multi-armed Bandits

1 Introduction

Consider the problem of assigning a patient to one of a number of different clinical treatments, whose efficacies are not precisely known. How should we decide which treatment to use? The problem is impossible to solve if there is just a single patient, but in practice there is a large number of patients who arrive for treatment sequentially. A natural approach would be to assign patients at random initially, until we had built up a good estimate of the efficacies, and to subsequently assign all patients to the one with the best estimated efficacy. But how good an estimate is good enough, and what are the chances that we end up choosing a sub-optimal treatment? There is clearly a trade-off between exploring the different treatments for a long time in order to gain more certaintly about having identified the optimal treatment, and reaching a conclusion quickly in order not to subject a large number of patients to sub-optimal treatments.

This is a motivating example of the class of problems we will study in the first half of the course[1]. What they have in common is a choice between a finite number of actions, and a random reward or payoff associated with each action, whose distribution is unknown. However, the scenario is repeated indefinitely, sequentially over time, which makes it feasible to learn the reward distribution or some summary statistics. The goal is to "maximise" the total reward in the long run, in a sense to be made precise. In fact, there is more than one way to make precise the notion of long-run reward, and we shall be looking at a few different meanings of the term. We now introduce a formal model to describe this scenario, which is known as the multi-armed bandit problem.

[1] This particular example also has an ethical dimension, but that will not be explicitly included in the models we study.

Multi-armed bandit problem An agent is faced with a choice of K arms or actions. Each time the agent plays arm i, it receives a random real-valued reward or payoff drawn from a distribution ν_i, independent of the past. The distributions $\nu_1, \nu_2, \ldots, \nu_K$ are unknown to the agent. The agent seeks to maximise a measure of long-run reward.

More formally, we consider a probability space on which we define K mutually independent sequences of random variables, $X_i(t)$, $i = 1, 2, \ldots, K$, $t \in \mathbb{N}$. We think of $X_i(t)$ as denoting the reward that the agent would have obtained at time step t by choosing arm i. Thus, for each i, $X_i(t), t \in \mathbb{N}$ is an iid (independent and identically distributed) sequence of random variables, with distribution ν_i. We assume that these distributions have finite means, which we denote μ_i, $i = 1, 2, \ldots, K$. Define

$$\mu^* = \max_{i=1}^{K} \mu_i$$

to be the largest expected reward of any arm.

Let $I(t) \in \{1, \ldots, K\}$ denote the arm played by the agent at time step t. This arm will be chosen according to some strategy adopted by the agent, based on the information available to the agent at time t. We now make explicit our assumptions about this information. We assume that, at each time step, only the reward for the arm chosen in that time step is observed. Even though we have, for convenience, defined random variables corresponding to all arms, their realisations for arms that are not played are unobserved. Thus, at the beginning of time step t, the only information available to the agent is the values $X_{I(s)}(s)$, $s \leq t-1$. The agent must decide which arm to play based on this information. We allow the agent to adopt a randomised strategy. In other words, the agent has access to an external source of randomness independent of $X_i(t)$, $i = 1, 2, \ldots, K$, $t \in \mathbb{N}$, which she can use to make her decisions. It suffices to assume that she has access to a sequence of random variables $U(t), t \in \mathbb{N}$, which are iid and uniformly distributed on $[0, 1]$, and independent of the sequences $X_i(\cdot)$, $1 \leq i \leq K$. Now, we can formally describe as a strategy $I(\cdot)$ as a sequence of maps $I(t), t \in \mathbb{N}$, where, more precisely

$$I(t) = I\Big(t, \{I(s), s \leq t-1\}, \{X_{I(s)}(s), s \leq t-1\}, U(t)\Big) \in \{1, \ldots, K\}.$$

We now discuss some ways to specify what we mean by the long-run reward. The most natural choice is perhaps the long-run average reward of a given strategy. As this may not necessarily exist, we instead define

$$X_* = \liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[X_{I(t)}]. \tag{1}$$

We might then seek a strategy $I(t)$ that maximises X_*. Clearly, for any strategy in which $I(t)$ is based only on realisations of $X_i(s)$, $s \leq t-1$, we must have $\mathbb{E}[X_{I(t)}] \leq \mu^*$, for any t. Hence, it follows that $X_* \leq \mu^*$. If we could come up with a strategy for which $X_* = \mu^*$, then this strategy is necessarily optimal (for the long-run average reward criterion). It turns out that this is not hard to achieve; in fact, we shall study a more stringent criterion, which we now describe.

Let $I(t), t \in \mathbb{N}$ denote a sequence of chosen arms. The regret in the first n time steps, corresponding to this choice, is given by

$$R_n = \max_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}\Big(X_k(t) - X_{I(t)}(t)]\Big) = \sum_{t=1}^{n} \Big(\mu^* - \mathbb{E}[X_{I(t)}(t)]\Big). \tag{2}$$

Remark: Some authors call the above quantity the pseudo-regret, and define regret as

$$\tilde{R}_n = \max_{k=1}^{K} \sum_{t=1}^{n} \Big(X_k(t) - X_{I(t)}(t)]\Big),$$

which depends upon the sample path, and is hence a random variable rather than a number. Note that $\tilde{R}_n$ includes intrinsic randomness which cannot be learnt (as it is independent of the past, by definition) whereas R_n only involves expectations, which can be learnt from observations. As we are interested in quantifying how quickly it can be learnt, it makes sense to focus on R_n as the performance metric.

The study of bounds on achievable regret will take up the remainder of this chapter. But before that, we need to recall some mathematical preliminaries.

2 Probability Inequalities

What can we say about the probability of a random variable taking values in a certain set if we only know its moments, for instance, or its generating function? It turns out that they give us some bounds on the probability of the random variable taking values in certain specific sets. We now look at some examples.

Markov's inequality Let X be a non-negative random variable with finite mean $\mathbb{E}X$. Then, for all $c > 0$, we have

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}X}{c}.$$

The proof is straightforward. For an event A, let $\mathbf{1}(A)$ denote its indicator. In other words, $\mathbf{1}(A)$ is the random variable defined as $\mathbf{1}(A)(\omega) = 1$ if $\omega \in A$ and $\mathbf{1}(A)(\omega) = 0$ if $\omega \notin A$. Fix $c > 0$. We have

$$\mathbb{E}X \geq \mathbb{E}[X\mathbf{1}(X \geq c)] \geq \mathbb{E}[c\mathbf{1}(X \geq c)] = c\mathbb{P}(X \geq c).$$

Re-arranging this gives us Markov's inequality. (Why does X have to be non-negative?)

Chebyshev's inequality Let X be a random-variable, not necessarily non-negative, with finite mean $\mathbb{E}X$ and finite variance $\mathrm{Var}(X)$. Then, for all $c > 0$, we have

$$\mathbb{P}(|X - \mathbb{E}X| \geq c) \leq \frac{\mathrm{Var}(X)}{c^2}.$$

The proof is an easy consequence of Markov's inequality. Note that the event $|X - \mathbb{E}X| \geq c$ is the same as the event $(X - \mathbb{E}X)^2 \geq c^2$, and apply Markov's inequality to the non-negative random variable $Y = (X - \mathbb{E}X)^2$. Use the fact that $\mathbb{E}Y = \mathrm{Var}(X)$.

The result extends conveniently to sums of iid (independent and identically distributed) random variables, since the variance of the sum is the sum of the variances. Let $X_1, X_2, \dots X_n$ be iid random variables, with finite mean μ and finite variance σ^2. Then, for all $c > 0$, we get

$$\mathbb{P}\Big(\Big| \sum_{i=1}^{n} X_i - n\mu \Big| \geq nc\Big) \leq \frac{\sigma^2}{nc^2}.$$

Chernoff bounds Let X be a random-variable, not necessarily non-negative, and suppose that its moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite for all θ. Then, for all $c \in \mathbb{R}$, we have

$$\mathbb{P}(X \geq c) \leq \inf_{\theta > 0} e^{-\theta c}\,\mathbb{E}[e^{\theta X}], \quad \mathbb{P}(X \leq c) \leq \inf_{\theta < 0} e^{-\theta c}\,\mathbb{E}[e^{\theta X}].$$

In fact, Chernoff only stated special cases of this inequality, and earlier versions were proved by Bernstein. But the general form follows easily from

Markov's inequality, and it has widely come to be known as Chernoff's bound, so we continue to use this terminology. The proof follows by noting that the event $X \geq c$ is identical to the event $e^{\theta X} \geq e^{\theta c}$ for all $\theta > 0$ (the inequality gets reversed for $\theta < 0$), applying Markov's inequality to the non-negative random variable $Y = e^{\theta X}$, and taking the best bound over all possible θ.

The result extends easily to sums of iid random variables. Let $X_1, X_2, \ldots, X_n$ be iid, and let $\phi(\theta) = \mathbb{E}[e^{\theta X_1}]$. Then, for all $c \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^n X_i \geq nc\Big) &\leq \inf_{\theta>0} e^{-n\theta c} \left(\mathbb{E}[e^{\theta X}]\right)^n, \\
\mathbb{P}\Big(\sum_{i=1}^n X_i \leq nc\Big) &\leq \inf_{\theta<0} e^{-n\theta c} \left(\mathbb{E}[e^{\theta X}]\right)^n.
\end{aligned}$$

Hoeffding's inequality This is a version of Chernoff's bounded for sums of iid bounded random variables (random variables taking values in a bounded interval), which doesn't require knowledge of their moment-generating functions (mgfs), but only of their mean.

Theorem 1 (Hoeffding) *Let $X_1, X_2, \ldots X_n$ be iid random variables, taking values in $[0, 1]$, and let $\mu = \mathbb{E}[X_1]$. Then,*

$$\mathbb{P}\Big(\sum_{i=1}^n X_i - n\mu > nt\Big) \leq e^{-2nt^2}, \quad \forall\, t > 0.$$

The proof uses the following lemma, also due to Hoeffding, which provides a uniform bound on the mgf of bounded random variables.

Lemma 1 *Let X be a random variable taking values in $[0, 1]$, with mean μ. Then,*

$$\mathbb{E}[e^{\theta(X-\mu)}] \leq e^{\theta^2/8}, \quad \forall\, \theta \in \mathbb{R}.$$

We now use the above lemma to prove Hoeffding's theorem. We won't prove the lemma as stated, as the proof is quite analytical and not very insightful. Instead, we shall present a proof of a weaker version of the lemma, with the constant 8 in the denominator replaced by 2; this proof uses an interesting technique called symmetrisation.

Proof of Hoeffding's theorem. First, observe from Chernoff's bound that

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i - n\mu > nt\Big) \leq e^{-\theta nt}\big(\mathbb{E}[\exp(\theta(X_1 - \mu))]\big)^n,$$

for all $\theta > 0$. Using Lemma 1 to bound the mgf, we get

$$\frac{1}{n}\log \mathbb{P}\Big(\sum_{i=1}^{n} X_i - n\mu > nt\Big) \leq -\theta t + \frac{\theta^2}{8}, \quad \forall\, \theta > 0.$$

We see that the RHS is minimised at $\theta = 4t$, which is positive if t is positive. The corresponding upper bound coincides with the bound in the statement of Theorem 1. □

Before discussing the proof of Lemma 1, we need to recall a fact about convex functions.

Defintion. A function $f : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if, for all $x, y \in \mathbb{R}$ and all $\alpha \in [0, 1]$, we have

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y).$$

The inequality is considered to be satisfied if both sides are $+\infty$.

It is easy to check if a smooth function is convex; if f is twice continuously differentiable, then f is convex if and only if $f''(x) \geq 0$ for all $x \in \mathbb{R}$.

Jensen's inequality. Let f be a convex function and let X be a random variable. Then, $\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$. (The inequality is considered to be satisfied if both sides as $+\infty$.)

Proof of (weaker version of) Lemma 1. Let X' be an independent copy of X, so that it has the same mean μ. It is easy to verify that $f(x) = e^{\theta x}$ is a convex function for any $\theta \in \mathbb{R}$. Hence, by Jensen's inequality,

$$\mathbb{E}[e^{-\theta X'}] \geq \exp(-\theta \mathbb{E}[X']) = e^{-\theta\mu}.$$

Consequently, we obtain that

$$\mathbb{E}[e^{\theta(X-\mu)}] \leq \mathbb{E}[\mathbb{E}[e^{\theta(X-X')}|X]] = \mathbb{E}[e^{\theta(X-X')}], \tag{3}$$

where the equality holds because of the independence of X and X'.

Since X and X' take values in $[0, 1]$, $X - X'$ takes values in $[-1, 1]$. Moreover, as X and X' have the same distribution, $X - X'$ has mean zero and a distribution that is symmetric around zero. Let S be a discrete random variable that takes values ± 1, with equal probability $1/2$, and is independent of X and X'. We have by the symmetry of $X - X'$, and its independence from S, that $X - X'$ has the same distribution as $S(X - X')$. Hence, for all $\theta \in \mathbb{R}$,

$$\mathbb{E}[e^{\theta(X-X')}] = \mathbb{E}[e^{\theta S(X-X')}] = \mathbb{E}[\mathbb{E}[e^{(\theta(X-X'))S}|X, X'] \leq \frac{1}{2}(e^{\theta} + e^{-\theta}). \quad (4)$$

To see the last inequality, not that, conditional on X and X', $\theta(X - X') = \eta$ for some $\eta \in [-\theta, \theta]$, as $X - X' \in [-1, 1]$. Now use the fact that $x \mapsto e^x + e^{-x}$ is decreasing on $(-\infty, 0)$ and increasing on $(0, \infty)$ and symmetric around 0. Finally, we observe using Taylor series that

$$\frac{1}{2}(e^{\theta} + e^{-\theta}) = \sum_{n=0}^{\infty} \frac{\theta^{2n}}{(2n)!} \leq \sum_{n=0}^{\infty} \frac{(\theta^2/2)^n}{n!} = \exp(\theta^2/2). \quad (5)$$

Combining (3), (4) and (5), we obtain that

$$\mathbb{E}[e^{\theta(X-\mu)}] \leq e^{\theta^2/2},$$

which is the version of Lemma 1 that we set out to prove. □

3 A heuristic based on hypothesis testing

In this section, we consider a heuristic for the simplest version of the problem, with an emphasis on gaining intuition rather than finding optimal algorithms. Consider a bandit with two arms. The rewards $X_1(n), n \in \mathbb{N}$ from the first arm are iid Bernoulli random variables with parameter μ_1, while those from the second arm, $X_2(n), n \in \mathbb{N}$ are iid Bern(μ_2). Assume without loss of generality that $\mu_1 > \mu_2$, that the parameters μ_1 and μ_2 are known to the player, but it is not known which arm has which parameter. Finally, suppose that the player is given a fixed time horizon T, and seeks to minimise the regret up to time T.

An obvious idea is the following: play each arm a fixed number of times, N, and subsequently play the arm that had the larger empirical mean reward. The question is how to choose N to minimise the regret of this strategy. We shall address this question using Chernoff bounds for binomial random variables.

Lemma 2 *Let X have a binomial distribution, Bin(n, α), with parameters $n \in \mathbb{N}$, $\alpha \in (0,1)$. Then, for any $\beta > \alpha$, we have that*

$$\mathbb{P}(X \geq \beta n) \leq e^{-nK(\beta;\alpha)},$$

where

$$K(\beta;\alpha) = \begin{cases} \beta \log \frac{\beta}{\alpha} + (1-\beta) \log \frac{1-\beta}{1-\alpha}, & \text{if } \beta \in [0,1], \\ +\infty, \text{ otherwise,} \end{cases}$$

with $x \log x$ defined to be 0 if $x = 0$.

Similarly, if $\beta' < \alpha$, then

$$\mathbb{P}(X \leq \beta' n) \leq e^{-nK(\beta;\alpha)}.$$

Here, $K(\beta;\alpha)$ is called the relative entropy, or Kullback-Leibler divergence, of a Bern(β) distribution with respect to a Bern(α) distribution. The proof of the lemma is left as a homework problem.

We will use this result to analyse the proposed strategy. Notice that $S_1(N) := \sum_{t=1}^{N} X_1(t)$ has a Bin(N, μ_1) distribution, while $S_2(N) := \sum_{t=1}^{N} X_2(t)$ has a Bin(N, μ_2) distribution. Moreover, these two random variables are independent, as the sequences $X_1(\cdot)$ and $X_2(\cdot)$ are independent of each other. Let $\beta \in (\mu_2, \mu_1)$. Then, we have by Lemma 2 that

$$\mathbb{P}(S_1(N) < \beta N, S_2(N) > \beta N) \leq e^{-N(K(\beta;\mu_1)+K(\beta;\mu_2))} = e^{-NJ(\mu_1,\mu_2)},$$

where

$$J(\mu_1, \mu_2) = \inf_{\beta \in [\mu_2, \mu_1]} K(\beta; \mu_1) + K(\beta; \mu_2).$$

The value of β which solves the minimisation problem above describes the most likely way for the event $S_1(N) < S_2(N)$ to occur. While it is not obvious, it turns out that it also captures the exponential decay rate of the probability of this event. More precisely,

$$\mathbb{P}(S_1(N) < S_2(N)) \leq e^{-NJ(\mu_1,\mu_2)},$$

and

$$\lim_{N \to \infty} \frac{1}{N} \log \mathbb{P}(S_1(N) < S_2(N)) = -J(\mu_1, \mu_2).$$

We will not prove this here. But taking this for granted, we will describe how to choose N to minimise regret.

If the correct decision is taken, regret is incurred only during the exploration phase, and is equal to $N(\mu_1 - \mu_2)$, since the wrong arm is played N times, incurring a regret of $\mu_1 - \mu_2$ each time. If the incorrect decision is taken, then the regret incurred is $(T - N)(\mu_1 - \mu_2)$ is the wrong arm is played N times during the exploration phase, and $T - 2N$ times during the exploitation phase. Hence, the overall regret up to time T is given by

$$\begin{aligned} \mathcal{R}(T) &= (\mu_1 - \mu_2)(T - 2N)\mathbb{P}(S_1(N) < S_2(N)) + (\mu_1 - \mu_2)N \\ &\approx (\mu_1 - \mu_2)\big(N + Te^{-NJ(\mu_1,\mu_2)}\big). \end{aligned}$$

It is easy to see that the last expression is minimised for N close to the solution of $TJ(\mu_1, \mu_2)e^{-NJ(\mu_1,\mu_2)} = 1$, i.e., when $N = \log T/J(\mu_1, \mu_2)+O(1)$. Moreover, the corresponding regret is

$$\mathcal{R}(T) = \frac{(\mu_1 - \mu_2)}{J(\mu_1, \mu_2)} \log T + O(1).$$

It can be shown that, if μ_1 and μ_2 are very close to each other, then $J(\mu_1, \mu_2) \approx (\mu_1 - \mu_2)^2$, and the above expression becomes $\mathcal{R}(T) = \log T/(\mu_1 - \mu_2) + O(1)$. We shall see later that this expression captures the correct scaling of the best achievable regret, for large T and small $\mu_1 - \mu_2$.