An introduction to multi-armed bandits

Aditya Mahajan

March 13, 2017

1 Motivation

Suppose there is a drug, let's call it A, for treating a life threatening disease like cancer. From previous studies, it is known that the effectiveness of the drug is p_A (i.e., after being treated with drug A, a cancer patient survives with probability p_A and dies with probability $1-p_A$). A pharmaceutical company comes up with a new drug—drug B—with an unknown effectiveness p_B. Which drug should a doctor prescribe to a cancer patient? Should he play safe and prescribe A or should he take a leap of faith and prescribe B? If the doctor prescribes B and the patient survives, which drug should he prescribe to the next patient? What if the patient dies? What if the doctor has to treat a number of patients? This dilemma becomes trickier if the effectiveness of both drugs are unknown or if there are more than two drugs.

The above example highlights the *exploration versus exploitation* dilemma that arises in many walks of life: from something as simple as choosing a dish at a restaurant or as important as choosing which research problem to work on!

Such models are called *multi-armed bandits* (MAB), which is a play of words on the term "one armed bandit" used to describe a slot machine. It is meant to evoke the mental picture of the dilemma faced by gambler deciding how to maximize his fortune when he has the option to play multiple slot machines with unknown win probabilities.

In a general multi-armed bandit problem, a decision maker is faced with the task of choosing one of k alternatives at each time, where the alternatives yield *unknown* rewards. The objective is to choose a *good* strategy to pick alternatives. The solution depends on the precise definition of *unknown* and *good*.

Uncertainty may be quantified using either a frequentist or a Bayesian approach. Goodness may be quantified using a discounted total reward or a long term average reward[1]. However, it turns out that long run average is not a useful objective in this case because it does not penalize wrong choices as long as they are made a sub-linear number of times. One option is to look at cumulative reward (which will be unbounded)

[1] Given a reward sequence $\{R_t\}_{t\geq 0}$, the discounted reward with discount factor $\beta \in (0,1)$ is given by $\sum_{t=0}^{\infty} \beta^t R_t$ and the long term average is given by $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} R_i$.

and quantify the performance using the rate at which its cumulative reward reaches the maximum—a quantity that is called *regret*.

Historically, Bellman (1956) posed the discounted reward formulation of MAB and Robbins (1952) posed the regret formulation of MAB. We start with an example that explains the two formulations.

Example 1 (MAB with Bernuolli rewards) Consider a MAB with k alternatives (or arms), where the reward obtained from each arm is a Bernoulli process and is independent of the rewards obtained from other arms.

In the discounted reward formulation of Example 1, we assume a Beta(a_0^i, b_0^i) prior on the success probability of arm i. At time t, suppose arm i has been played n^i times and has yielded n_+^i successes and n_-^i failures. Then the posterior on p^i is a Beta(a_t^i, b_t^i) distribution[2] where $a_t^i = a_0^i + n_+^i$ and $b_t^i = b_0^i + n_-^i$. Therefore, $X_t^i \equiv (a_t^i, q_t^i)$ may be used as an *information state* (or a *sufficient statistic*) for bandit process i. This state evolves in a Markovian manner. In particular, if $X_t^i = (a_t^i, b_t^i)$ and arm i is played at time t, then

$$X_{t+1}^i = (a_{t+1}^i, b_{t+1}^i) = \begin{cases} (a_t^i + 1, b_t^i), & \text{w.p. } a_t^i/(a_t^i + b_t^i) \\ (a_t^i, b_t^i + 1), & \text{w.p. } b_t^i/(a_t^i + b_t^i) \end{cases}$$

and the average reward obtained by playing arm i is state (a_t^i, b_t^i) is the mean of Beta(a_t^i, b_t^i) distribution, that is:

$$R^i(a_t^i, b_t^i) = \frac{a_t^i}{a_t^i + b_t^i}.$$

* * *

There are two variation of the regret formulation. The first is the *stochastic setup* in which the success probability of arm i is assumed to have an unknown value p^i. Let $* = \arg\max_{i \in \{1,\ldots,k\}} p^i$ denote the best arm. Let U_t denote the arm played at time t and $X_t^{U_t}$ denote the reward obtained when arm U_t is played[3]. The expected *regret* (or the *weak* expected regret) after T plays is defined as

$$R_T = Tp^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t^{U_t}\right] = Tp^* - \sum_{i=1}^{k} \mathbb{E}[N_T^i] p^i = \sum_{i=1}^{k} \mathbb{E}[N_T^i] \Delta^i$$

where N_t^i denote the number of times arm i has been played up to time t and $\Delta^i = p^* - p^i$. To get a small regret one has to ensure that the non-optimal

[2] If a random variable X has a Beta(a, b) distribution, then its probability density $f_X(x)$ is proportional to $x^a(1-x)^b$. Beta distribution is a conjugate prior of Bernoulli, binomial, and geometric distributions.

[3] Note that the notation here is different from the Bayesian setup where X_t^i denotes the state (or the posterior distribution) and $R^i(X_t^i)$ denotes the reward. In the regret literature, the reward process is denoted by X_t^i.

The other formulation is the *adversarial setup* in which an adversary sets the reward sequence $\{X_i\}_{i\geq 1}$. If this choice is independent of the decision maker's action, then the adversary is called an *oblivious* adversary; otherwise the adversary is called a *non-oblivious* adversary. The distinction is relevant only when the choice of actions are randomized. In the adversarial setup, regret is defined as

$$R_T = \max_{i\in\{1,\dots,k\}} \mathbb{E}\left[\sum_{t=1}^{T} X_t^i - \sum_{t=1}^{T} X_t^{U_t}\right]$$

The adversarial setup was motivated by the work of Baños (1968) on repeated games with unknown payoffs. A related concept is also considered in universal source compression (Ziv and Lempel, 1977, 1978).

* * *

We end this introduction with a comparison with other learning formulations: reinforcement learning and learning with expert advice. The objective of reinforcement learning to asymptotically identify an optimal policy, but traditionally the performance of the policy during the learning phase is not part of the learning objective. Thus, RL may be viewed as a purely exploration problem though in recent years there has been some work on exploration-exploitation trade-off in reinforcement learning. In learning with expert advice, as in MAB, one cares about the performance during the learning phase but unlike MAB the feedback after each action tells us how all experts would have performed. In MAB, the feedback that we get only tells us how the selected expert performed.

2 The discounted reward formulation

A MAB process consists of k independent Markov[4] process $\{X_t^i\}_{t\geq 0}$, $X_t^i \in \mathcal{X}^i$, $i \in \{1,\dots,k\}$. At each time, a decision maker chooses to play one of the k processes; the rest of the processes remain frozen. Let $U_t \in \{1,\dots,k\}$ denote the process played at time t. Then,

$$X_{t+1}^i = \begin{cases} \text{Updates in a Markov manner,} & \text{if } U_t = i \\ X_t^i, & \text{otherwise.} \end{cases}$$

There is a reward function R^i associated with each process. The process i that is played at time t yields a reward $R^i(X_t^i)$. Other processes do not yield a reward.

A *scheduling policy* $g = (g_1, g_2, \dots)$ is a decision rule that determines how to choose U_t as a function of the entire history $X_{1:t}^{1:k}$ of all processes and the history $U_{1:t-1}$ of all decisions, i.e.,

$$U_t = g_t(X_{1:t-1}^{1:k}, U_{1:t-1}).$$

[4]The Markovian assumption is made simply for the ease of exposition. The results easily extend to general processes under mild technical conditions; see Varaiya et al. (1985).

The objective is to choose a scheduling policy that maximizes

$$J(g) \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R_{U_t}(X_t^i)\right]$$

TODO: talk about mutli-dimensional dynamic program. Curse of dimensionality.

Bellman (1956) solved the so called one-and-a-half armed bandit problem (where one of the process yields a constant reward) for the Bernoulli case. Motivated by clinical trials, a version of this problem had been considered by Johnson and Karlin (1954). Bellman formulated the problem as a dynamic program, proved that a solution exists, and showed qualitative properties of the solution. However, very little progress was made beyond the one-and-a-half armed bandit setup until Gittins and Jones (1974); Gittins (1979) provided a complete characterization of the solution.

The idea behind Gittins solution is to assign an index $\nu^i \colon \mathcal{X}^i \to \mathbb{R}$ to each arm, which is given by

$$\nu^i(x^i) = \max_{\tau > 0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t R^i(X_t^i) \;\middle|\; X_0^i = x^i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \beta^t \;\middle|\; X_0^i = x^i\right]}, \tag{1}$$

where the maximization is over all stopping times τ. Then, there are two ways to describe the optimal solution. The first interpretation is that decisions are made in epochs of random duration. At the beginning of each epoch, play the arm with the highest index. The duration of the epoch equals the stopping time in (1). It can shown that if we play arm i in an epoch and start in state x^i, then the Gittins index of all states in that epoch are higher than $\nu^i(x^i)$. Therefore, the second interpretation of the optimal solution is to always play the arm with the highest index.

Gittins solution is remarkable because it reduces n-dimensional optimization problem (find the best scheduling strategy) to n one-dimensional optimal stopping problems (compute the Gittins index for each process).

TODO: Talk about forward induction interpretation

Gittins (1979) used an intricate interchange argument to prove the optimality of the index rule. Whittle (1980) came up with an alternative proof, where he showed that the Gittins index of a process at state x is related to the *retirement reward* $M(x)$ at which a decision maker is indifferent between playing the arm starting at state x or retiring and receiving a one-time reward $M(x)$. Subsequently, various other proofs for the optimality of Gittins index have been presented; perhaps the simplest being Weber (1992); Tsitsiklis (1994); Ishikida and Varaiya (1994). See Frostig and Weiss (2016) who present the different proofs in a unified notation.

Various algorithms have been proposed to compute Gittins index efficiently. We refer the reader to Chakravorty and Mahajan (2014) for an overview. The off-line algorithms (i.e., algorithms that compute the Gittins index of all states) use properties of Markov chains to compute the index; on-line algorithms (i.e., algorithms that compute the Gittins

index of only the current state) use the retirement reward formulation of Whittle (1980) and compute the index either solving the dynamic program or solving a linear program. The computational complexity of the best off-line algorithms is $\frac{2}{3}m^3 + \mathcal{O}(m^2)$, where $m = |\mathcal{X}^i|$. There are also some closed form approximations to the Gittins index for bandits with Bernoulli and Gaussian reward processes (see Brezzi and Lai, 2002; Yao, 2006; Chakravorty and Mahajan, 2014)

* * *

The reason that the solution of the MAB problem is significantly simpler than a general stochastic control problem is that the MAB problem has following salient features:

(F1) only one process is played at each time. The evolution of the process that is played is controlled (i.e., we can choose which processor to operator but not how to operate it);

(F2) processes that are not played remain frozen;

(F3) frozen processes don't contribute to the reward

(F4) processes are independent.

Over the years, various extensions of MAB have been considered where these assumptions do not hold. These include MAB with multiple plays (where more than one arm may be played at each time); arm-acquiring bandits (where the number of arms may change with time); MAB with switching cost (where there is a cost associated with switching to a new arm); and restless bandits (where the armed that are not played continue to evolve and potentially convert to the reward). In general, the Gittins index rule is not optimal for these variations, though several sufficient conditions on the optimality of the index rule are available. We refer the reader to Mahajan and Teneketzis (2008) for a survey. In practice, even when its optimality cannot be established, t the index rule performs well (Glazebrook and Minty, 2009; Verloop, 2016) and is therefore often used as a heuristic.

* * *

Finally, we present two examples that might be relevant for the spectrum access challenge.

Example 2 (Pandora's problem (Weitzman, 1979)) There are k boxes. Box i, $i \in \{1, \dots, k\}$ contains a potential reward of X^i with a probability distribution function $F^i(X^i)$ independent of the other rewards. It costs c^i to open box i and learn its contents.

At each stage, Pandora has to decide whether or not to open a box. If she chooses to stop searching, Pandora collects at that time the maximum reward that she has uncovered so far. If she decides to continue, she must select the next box to be opened, pay the fee for opening it, and wait for the outcome. After looking at the outcome, she decides whether to continue or not.

The objective is to find a search strategy that maximizes the expected discounted reward.

Weitzman (1979) established an optimal strategy using an interchange argument. We present a MAB formulation of the problem (following Weber, 2016). Associate a bandit process with each box. The initial state of arm i is $\star$; the reward in state $\star$ is $-c^i$. When arm i is played for the first time, the next state X^i is chosen according to distribution F^i and it yields a reward $R^i(x^i) = x^i$. In all subsequent plays, the state remains X^i and yields reward x^i. The objective is to maximize the expected total reward.

For an opened box in state x^i, the Gittins index $\nu^i(x^i)$ is $R^i(x^i) = x^i$. For an unopened box, we can find the Gittins index by considering a one-and-a-half armed bandit. Suppose there are two arms: one that yields a constant reward of λ and the other that corresponds to box i. Then, the value of choosing arm i is

$$-c^i + \beta\left[\lambda \int_{-\infty}^{\lambda} dF^i(x^i) + \int_{\lambda}^{\infty} x^i dF^i(x^i)\right]$$

Now, the Gittins index of arm i is the value of λ for which we are indifferent between the two arms. Thus, $\nu^i(\star)$ is the smallest solution of

$$\frac{\lambda}{(1-\beta)} = -c^i + \frac{\beta}{1-\beta}\mathbb{E}[\max\{x^i, \lambda\}].$$

Weitzman (1979) calls this the *reservation price* of box i. The optimal policy can be described as follows: open boxes in the decreasing order of their reservation price $\nu^i(\star)$ until the time when the revealed prize is greater than the reservation price of all unopened boxes. Olszewski and Weber (2015) generalized this result to the case when the reward depends on all the opened boxed (and not just the box with the highest value).

Example 3 (Searching for a stationary object) Consider a stationary object hidden in one of k cells. The *a priori* probability that the target is in cell i is denoted by p_0^i. A sensor can search one location at each time. The search is imperfect. If the target is in cell i and the sensor searches in cell i, then the target is found with probability q^i. The

The objective is find a search strategy that minimizes the expected time to find the target.

Let p_t^i denote the posterior probability that the object is in cell i if the previous t searches have been unsuccessful. Suppose cell j is searched at time t and the object is not found. Then,

$$p_{t+1}^j = \frac{p_t^j(1-q^j)}{c} \quad \text{and} \quad p_{t+1}^i = \frac{p_t^i(1-q^j)}{c}, \quad \forall i \neq j$$

where $c = 1 - p_t^j q^j$. The reward at state p_t^i is $p_t^i q^i$ (the probability that the object is found).

Due to the normalization in the update of the posterior probability, the above model does not satisfy feature (F2) of the MAB problem. However, it is possible to view the

above problem as a multi-armed bandit problem as follow (Song and Teneketzis, 2004): Let X_t^i denote the number of times location i has been searched until time t. Then, if location i is searched at time t, the probability that the object is found is

$$R^i(X_t^i) = p^i q^i (1-q^i)^{X_t^i}.$$

The next state is $X_{t+1}^i = X_t^i + 1$. The above model satisfies features (F1)–(F4) of MAB. To compute the Gittins index, observe that the sequence of rewards

$$\pi_n^i = p^i q^i (1-q^i)^n, \quad n \geq 1$$

from arm i form a strictly decreasing sequence. Hence, the Gittins index is attained at $\tau = 1$. Thus,

$$\nu^i(x^i) = p^i q^i (1-q^n)^{x^i}$$

and the optimal strategy is to search the location with the highest index. Song and Teneketzis (2004) showed that the Gittins index rule remains optimal for multiple sensors as well.

An alternative proof of the optimality of the above procedure is given in DeGroot (1974, Sec 14.14). Let $\hat{p}_t^i$ denote the *unnormalized* posterior probability that the object is at location i. The unnormalized posterior evolves as follows: $\hat{p}_0^i = p_0^i$ and if location j is searched at time t then

$$p_{t+1}^j = p_t^j (1-q^j) \quad \text{and} \quad p_{t+1}^i = p_t^i (1-q^j), \quad \forall i \neq j.$$

The modified reward function is $R^i(\hat{p}^i) = \hat{p}^i q^i$. DeGroot (1974) showed that the optimal policy of the modified model is the same as the optimal policy of the original model. Since the reward sequence obtained from arm i is strictly decreasing, the Gittins index is always achieved at $\tau = 1$, and the Gittins index is given by

$$\nu^i(\hat{p}_t^i) = \hat{p}_t^i q^i.$$

Note that this index is same as the one obtained by considering the number of times a location has been searched as the state.

3 The stochastic version of the regret formulation

There are k arms. The rewards from arm i, $i \in \{1, \ldots, k\}$, are *i.i.d.* and distributed according to a *univariate* density $f(\cdot; \theta^i)$, where $f(\cdot;\cdot)$ is a known function (e.g., the exponential family) and $(\theta^1, \ldots, \theta^k)$ are unknown parameters belonging to some set Θ. Assume $\int_\infty^\infty |x| f(x;\theta) dx < \infty$ for all $\theta \in \Theta$. Let

$$\mu(\theta) := \int_{-\infty}^{\infty} f(x;\theta) dx$$

and define

$$\mu^* = \max\{\mu(\theta^1), \ldots, \mu(\theta^k)\} = \mu(\theta^*)$$

for some $* \in \{1, \dots, k\}$.

An adaptive allocation rule ϕ consists of a sequence of random variables $\{U_t\}_{t \geq 1}$ taking values in the set $\{1, \dots, k\}$ such that U_t is a (measurable) function of the history $(X_{1:t-1}, U_{1:t-1})$. The event $\{U_t = i\}$ denotes that we sample from arm i at time t. Let

$$N_T^i := \sum_{t=1}^{T} \mathbb{1}\{U_t = i\}$$

denote the number of times arm i has been played up to time T.

Given a parameter configuration $\boldsymbol{\theta} = (\theta^1, \dots, \theta^n)$, define the sample regret at time T as

$$R_T(\boldsymbol{\theta}) := T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} X_t\right] = \sum_{i \neq *} (\mu^* - \mu(\theta^i)) \mathbb{E}[N_T^i].$$

An allocation rule is called *uniformly good* if for every parameter configuration $\boldsymbol{\theta} = (\theta^1, \dots, \theta^k)$,

$$R_T(\boldsymbol{\theta}) = o(T^a) \quad \text{or equivalently} \quad \lim_{T \to \infty} \frac{R_T(\boldsymbol{\theta})}{T^a} = 0$$

for every $a > 0$. In the regret literature, attention is restricted to uniformly good allocation rules; others are considered uninteresting.

Recall that the Kullback-Leibler distance between two distributions is defined as

$$I(\theta; \lambda) = \int_{-\infty}^{\infty} \frac{\log f(x; \theta)}{\log f(x; \lambda)} f(x; \theta) dx.$$

TODO: Talk about why the problem is difficult. Greedy and ε-Greedy algorithms have linear regret.

Under some mild technical conditions on the distributions, Lai and Robbins (1985) showed that for any uniformly good allocation rule and any inferior arm $i \neq *$,

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_T^i]}{\log T} \geq \frac{1}{I(\theta^i; \theta^*)}.$$

This means that a uniformly good allocation rule must play all inferior arms at least a logarithmic number of times. Consequently,

$$\liminf_{T \to \infty} \frac{R_T(\boldsymbol{\theta})}{\log T} \geq \sum_{i \neq *} \frac{(\mu^* - \mu(\theta^i))}{I(\theta^i; \theta^*)}$$

See Garivier et al. (2016) for an elegant proof of this lower bound. Note that the model has a restrictive assumption: the rewards from each arm are specified by a density that depends on a *single* unknown parameter. Burnetas and Katehakis (1996) showed that a slightly weaker lower bound holds under no parametric assumptions.

Lai and Robbins (1985) also proposed index based strategies that were asymptotically optimal, i.e., for any configuration $\boldsymbol{\theta}$

$$R_T(\boldsymbol{\theta}) \approx \left\{ \sum_{i \neq *} \frac{(\mu^* - \mu(\theta^i))}{I(\theta^i; \theta^*)} \right\} \log T \quad \text{as } T \to \infty.$$

The strategies depend on an two sequence of function. The first is *upper confidence bound* functions $\{g_{m,n}\}_{n\in\mathbb{N},m\in\{1,\dots,n\}}$ that satisfy

1. For any $m \in \mathbb{N}$, the sequence $\{g_{m,n}\}_{n\geq m}$ is non-decreasing.

2. For every $\theta \in \Theta$, and any $r < \mu(\theta)$,

$$\mathbb{P}_\theta\Big(r \leq g_{m,n}(Y_{1:m}) \text{for al } m \leq n\Big) = 1 - o(n^{-1}).$$

3. For any $\theta, \lambda \in \Theta$ such that $\mu(\lambda) > \mu(\theta)$,

$$\lim_{\varepsilon\downarrow 0}\left[\limsup_{n\to\infty}\sum_{m=1}^{n}\mathbb{P}_\theta\Big(g_{m,n}(Y_{1:m}) \geq \mu(\lambda) - \varepsilon\Big)/\log n\right] \leq 1/I(\theta;\lambda).$$

The second sequence is estimators $\{h_n\}_{n\geq 1}$ of the mean such that

1. For any $m, n \in \mathbb{N}$ such that $n \geq m$,

$$h_m \leq g_{m,n}.$$

2. For every $\theta \in \Theta$, $\varepsilon > 0$, and $\delta \in (0, 1)$,

$$\mathbb{P}_\theta\Big(\max_{\delta n\leq m\leq n}|h_m(Y_{1:m}) - \mu(\theta)| \geq \varepsilon\Big) = o(n^{-1}).$$

Note that the sample mean $(Y_1 + \cdots + Y_m)/m$ satisfies the last condition. Lai and Robbins (1985) constructed mean estimators and upper confidence bound functions that satisfy the above properties for normal (with known variance), Bernoulli, exponential, and Poisson distributions.

Now, at any time T, define MEAN_T^i and UCB_T^i as the mean and upper confidence bound of arm i, i.e.,

$$\text{MEAN}_T^i = h_{N_T^i}(X_{1:N_T^i}^i) \quad \text{and} \quad \text{UCB}_T^i = g_{N_T^i,T}(X_{1:N_T^i}^i).$$

Finally consider the following allocation strategy. Sample all arms once. Pick any $\delta \in (0, 1/k)$ and at any time $T > k$, let

$$\text{MEAN}_T^* = \max_{i\in\{1,\dots,T\}}\{\text{MEAN}_T^i : N_t^i \geq \delta T\}$$

and let i_T^* denote the corresponding arg max. We call i_T^* as the leader. Furthermore, let

$$\text{UCB}_T^* = \max_{i\in\{1,\dots,T\}}\{\text{UCB}_T^i\}$$

and let $j{*}_T$ denote the corresponding arg max. We call j_T^* as the alternate candidate.

If $\text{MEAN}_T^* \geq \text{UCB}_T^*$ then sample the leader, else sample the alternative candidate. Lai and Robbins (1985) showed that the under this sampling policy, for every configuration $\boldsymbol{\theta} = (\theta^1, \ldots, \theta^k)$ and every $i \neq *$,

$$\mathbb{E}_\theta[N_T^i] \leq \left(\frac{1}{I(\theta^i; \theta^*)} + o(1) \right) \log T.$$

Thus, the strategy proposed by Lai and Robbins (1985) samples the sub-optimal arms at the optimal rate.

* * *

Following Lai and Robbins (1985) various authors have obtained asymptotically efficient allocation rules for several variations. Anantharam et al. (1987a) extended the results to MAB with multiple plays (i.e., instead of playing one bandit at each time, m bandits can be played). Let $\sigma(m)$ be the m-worst arm and $K^{(m)}$ be the collection of all arms which are worse than $\sigma(m)$. Then, under mild technical conditions, for any arm $i \in K^{(m)}$

$$\liminf_{T \to \infty} \frac{\mathbb{E}[N_T^i]}{\log T} \geq \frac{1}{I(\theta^i; \theta^{\sigma(m)})}$$

and

$$\liminf_{T \to \infty} \frac{R_T(\boldsymbol{\theta})}{\log T} \geq \sum_{i \in K^{(m)}} \frac{(\mu(\sigma(m)) - \mu(\theta^i))}{I(\theta^i; \theta^*)}.$$

They also showed that an obvious generalization of Lai and Robbins' scheme is asymptotically optimal.

Anantharam et al. (1987b) then considered MAB with Markovian rewards. Each arm is given by a *univariate* parametric family of stochastic transition matrices $P(\theta) = P(x, y; \theta)$. It is assumed that for every $\theta \in \Theta$, $P(\theta)$ is irreducible and aperiodic. Let $\pi(\cdot; \theta)$ denote the stationary distribution of $P(\theta)$ and define

$$\mu(\theta) = \sum_{x \in \mathcal{X}} x \pi(x; \theta).$$

In this case, the Kullback-Leibler distance between two arm $\theta, \lambda \in \Theta$ is defined as

$$I(\theta; \lambda) = \sum_{x \in \mathcal{X}} \pi(x) \sum_{y \in \mathcal{X}} P(x, y; \theta) \log \frac{P(x, y; \theta)}{P(x, y; \lambda)}.$$

With the above modifications, Anantharam et al. (1987b) showed that the lower bounds on regret is similar to the i.i.d. case and a minor modification to the scheme of Anantharam et al. (1987a) is asymptotically optimal.

Agrawal et al. (1988) considered the MAB problem with switching cost; i.e., a fixed cost has to paid each time we switch arms. In this case, the regret is defined as the usual regret plus the expected switching cost. The lower bound of Lai and Robbins

(1985) is still a lower bound on the regret with switching costs. Agrawal et al. (1988) constructed an ingenious scheme that asymptotically achieves the same regret! Their achievable scheme operated in blocks. Time is first divided into "frames" and each frame is further subdivided into "blocks" of equal size. The block length increases linearly with each frame and the number of blocks increase roughly exponentially with each frame. The basic idea of Agrawal et al. (1988) was to sample from the same arm for an entire block. Thus, the decision to select an arm is made only at the beginning of each block. This decision is based on the upper confidence bound functions, similar to Lai and Robbins (1985). The choice of number of blocks and block length ensures that the expected number of samples from each inferior population is $\mathcal{O}(\log T)$ and the expected number of switches is $o(\log T)$.

Further generalizations to controlled i.i.d. processes and controlled Markov chains were considered by Agrawal et al. (1989a,b), for certainty equivalence with forcing by Agrawal and Teneketzis (1989) and to Markovian rewards where the state keeps on evolving (restless bandits) by Liu et al. (2013).

The upper confidence bounds proposed by Lai and Robbins (1985) were quite complicated and typically relied on the entire sequence of observations from the corresponding arm. Inspired by the scheme in Agrawal and Teneketzis (1989), Agrawal (1995) proposed sample mean-based upper confidence bounds (for one parametric exponential family including Gaussian, exponential, Bernoulli, Poisson, and Laplacian) and showed that an index policy using these upper confidence bounds is asymptotically optimal.

* * *

In a remarkable result, Auer et al. (2002) showed that a slight variation of the sample path based upper confidence bound functions of Agrawal (1995) achieves a logarithmic regret uniformly over time rather than only asymptotically. The upper confidence bound function defined in Auer et al. (2002) was

$$\text{UCB}_T^i = \text{MEAN}_T^i + \sqrt{\frac{2\log T}{N_T^i}}$$

where MEAN_T^i is the sample mean of arm i at time T. The index policy is to play the arm with the highest UCB index. This policy is called UCB 1 in the literature.

Auer et al. (2002) showed that when reward distribution have bounded support on $[0,1]$, then under the UCB 1 strategy

$$\mathbb{E}[N_T^i] \leq \frac{8}{(\mu^* - \mu^i)^2}\log T + \text{small constant}$$

and the regret is

$$R_T = \left[8\sum_{i\neq *}\frac{\log T}{(\mu^* - \mu^i)}\right] + \left(1 + \frac{\pi^2}{3}\right)\left(\sum_{i=1}^{k}(\mu^* - \mu^i\right)$$

The leading constant $8/(\mu^* - \mu^i)^2$ is worse than the $1/I(\theta^i; \theta^*)$ constant in Lai and Robbins (1985). Thus, UCB 1 is not asymptotically optimal. Auer et al. (2002) presented a more complicated strategy, which they called UCB 2 that can bring the constant arbitrarily close to $2/(\mu^* - \mu^i)^2$. Bubeck et al. (2012) generalize UCB 1 to distributions that satisfy a moment condition and show that this generalization has a similar regret bound.

Inspired by the lower bounds of Burnetas and Katehakis (1996), **?** proposed a Kullback-Leibler distance based UCB algorithm for reward distributions with bounded support on $[0, 1]$. The KL-UCB index is given by

$$\text{KL-UCB}_T^i = \max\{p \in [0, 1] : N_T^i I(\text{MEAN}_T^i, p) \leq \log T + c \log\log T\}$$

where c is a tunable parameter.

Garivier and Cappé (2011) showed that (for $c = 3$) the number of times KL-UCB algorithm chooses a suboptimal arm i is upper bounded by

$$\mathbb{E}[N_T^i] \leq \frac{\log T}{I(\mu(\theta); \mu(\theta^*))}(1 + \varepsilon) + C \log\log T + \mathcal{O}(1)$$

which matches the lower bound of Lai and Robbins (1985); Burnetas and Katehakis (1996). Therefore, the finite-time regret of KL-UCB is

$$\frac{R_T}{\log T} \leq (1 + \varepsilon) \sum_{i \neq *} \frac{(\mu^* - \mu(\theta^i)}{I(\mu(\theta^i); \mu^*)} + o(1),$$

which is also asymptotically optimal.

Tekin and Liu (2012) consideed MAB with Markovian rewards where the arms that are not played keep on evolving (the restless bandit setup) and proposed a variation of the UCB algorithm (called UCB-M) that achieves logarithmic regret.

* * *

All of the above models and results are non-Bayesian. In recent years, the regret formulation has also been investigated in the Bayesian setup.

Kaufmann (2016) considered the MAB problem where the rewards come from a single parameter exponential family with a conjugate prior. Let π_t^i denote the posterior distribution on the parameters of arm i at time t. π_t^i will also belongs to the exponential family. The Bayes-UCB index is defined as

$$\text{Bayes-UCB}_T^i = Q(1 - \alpha_t; \pi_T^i), \quad \text{where } \alpha_t = \frac{1}{t(\log T)^c}$$

where c is a tunable parameter and $Q(\alpha, \pi)$ is the quantile function associated with distribution π such that $\mathbb{P}_\pi(X \leq Q(\alpha, \pi)) = \alpha$. Kaufmann (2016) showed that Bayes-UCB index policy is asymptotically optimal.

Various authors had noted that the sampling algorithm proposed in **?** (which is now called Thompson sampling) performs well in practice. Thompson sampling proceeds as follows: let π_T^i denote the posterior distribution on the reward process for arm i. At time t, take a sample $Y_T^i \sim \pi_T^i$ from every arm $i \in \{1, \ldots, k\}$ and pick the arm with the highest sampled valued!

Agrawal and Goyal (2012) obtained the first logarithmic upper bound on the performance of Thompson sampling (for Bernoulli rewards) and showed that

$$R_T \leq C \sum_{i \neq *} \frac{\log T}{(\mu^* - \mu(\theta^i))^2} + o(\log T).$$

Kaufmann et al. (2012) obtained a tighter bound (again for Bernoulli rewards) and showed that

$$R_T \leq (1+\varepsilon) \sum_{i \neq *} \frac{\mu^* - \mu(\theta^i)}{I(\mu(\theta^i); \mu^*)} \log T + o(\log T),$$

which is asymptotically optimal.

4 The adversarial version of the regret formulation

I don't know the literature on adversarial bandits; the short overview below is taken from Bubeck et al. (2012).

As before, it is assumed that there are k arms, $\{X_t^i\}_{t \geq 1}$. It is convenient to assume that these are loss processes rather than reward processes, i.e., the objective is to minimize the loss $\sum_{t=1}^T X_t^{U_t}$. In the simplest setting, it is assumed that at each stage, the adversary assigns to each arm a loss $X_t^i \in [0,1]$ simultaneously with the player's choice of arm $U_t \in \{1, \ldots, k\}$.

If the arms are chosen by a deterministic algorithm, then the adversary knows the choice of U_t and can set the losses such that the regret is 1 at all times. The idea to get around this is to choose the arms using a randomized algorithm and look at average regret.

The basic algorithm that does this is called EXP 3. The algorithm is parametrized by a non-increasing sequence $\{\eta_t\}_{t \geq 1}$ of real numbers. It maintains a probability distribution $\pi_t = (\pi_t^1, \ldots, \pi_t^k)$ over the arms and estimated cumulative losses $(L_t^1, \ldots, L_t^k)$ of each arm. It starts with a uniform distribution π_1 over all arms and $L_0^i = 0$ and then at each round

1. Draw an arm U_t from the probability distribution π_t.
2. For arm U_t, set $\ell_t^{U_t} = X_t^{U_t} / \pi_t^{U_t}$; for all other arms set $\ell_t^i = 0$. Update $L_t^i = L_{t-1}^i + \ell_t^i$.
3. Compute the new probability distribution π_{t+1} over the arms as

$$\pi_{t+1}^i = \frac{\exp(-\eta_t L_t^i)}{\sum_{i=1}^K \exp(-\eta_t L_t^i)}$$

If the forecaster knows the number of rounds and chooses $\eta_t = \eta = \sqrt{2 \log k/Tk}$, then

$$\mathbb{E}[R_T] \leq \sqrt{2Tk \log K}.$$

If the forecaster does not know the number of rounds and chooses $\eta_t = \sqrt{\log k/tk}$ then

$$\mathbb{E}[R_T] \leq 2\sqrt{Tk \log K}.$$

The lower bound on the regret must be $\mathcal{O}(\sqrt{Tk})$. The intuition is the following. Suppose all arms are Bernoulli. At least one arm must be pulled less than T/k times and for this time one cannot differentiate between a Bernoulli with parameter $1/2$ and a Bernoulli whose parameter is $1/2 \pm \sqrt{k/T}$. Thus, if one arm is Bernoulli $1/2 - \sqrt{K/n}$ and all other arms are Bernoulli $1/2$, then the forecaster will incur a regret of order $T/\sqrt{k/T} = \sqrt{kT}$.

References

Agrawal, R. (1995). Sample mean based index policies with o(log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

Agrawal, R., Hedge, M. V., and Teneketzis, D. (1988). Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost. *IEEE Trans. Autom. Control*, 33(10):899–906.

Agrawal, R. and Teneketzis, D. (1989). Certainty equivalence control with forcing: revisited. *Systems & Control Letters*, 13(5):405 – 412.

Agrawal, R., Teneketzis, D., and Anantharam, V. (1989a). Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: finite parameter space. *IEEE Transactions on Automatic Control*, 34(3):258–267.

Agrawal, R., Teneketzis, D., and Anantharam, V. (1989b). Asymptotically efficient adaptive allocation schemes for controlled markov chains: finite parameter space. *IEEE Trans. Autom. Control*, 34(12):1249–1259.

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pages 39–1.

Anantharam, V., Varaiya, P., and Walrand, J. (1987a). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards. *IEEE Trans. Autom. Control*, 32(11):968–976.

Anantharam, V., Varaiya, P., and Walrand, J. (1987b). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards. *IEEE Trans. Autom. Control*, 32(11):977–982.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256.

Baños, A. (1968). On pseudo-games. *Ann. Math. Statist.*, 39(6):1932–1945.

Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics*, 16(3/4):221–229.

Brezzi, M. and Lai, T. L. (2002). Optimal learning and experimentation in bandit problems. *Journal of Economic Dynamics and Control*, 27(1):87–108.

Bubeck, S., Cesa-Bianchi, N., et al. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Burnetas, A. N. and Katehakis, M. N. (1996). Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122 – 142.

Chakravorty, J. and Mahajan, A. (2014). Multi-armed bandits, Gittins index, and its calculation. In *Methods and applications of statistics in clinical trials. Vol. 2*, Wiley Ser. Methods Appl. Statist., pages 416–435. Wiley, Hoboken, NJ.

DeGroot, M. (1974). Reaching a consensus. *Journal of the American Statistical Association*, (345):118–121.

Frostig, E. and Weiss, G. (2016). Four proofs of Gittins' multiarmed bandit theorem. *Annals of Operations Research*, 241(1):127–165.

Garivier, A. and Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, pages 359–376.

Garivier, A., Ménard, P., and Stoltz, G. (2016). Explore first, exploit next: The true shape of regret in bandit problems. *arXiv:1602.07182*.

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177.

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the discounted multiarmed bandit problem. In *Progress in Statistics*, volume 9, pages 241–266. North-Holland, Amsterdam, Netherlands.

Glazebrook, K. D. and Minty, R. (2009). A generalized Gittins index for a class of multiarmed bandits with general resource requirements. *Mathematics of Operations Research*, 34(1):26–44.

Ishikida, T. and Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications*, 83(1):113–154.

Johnson, S. and Karlin, S. (1954). A bayes model in sequential design. Technical Report P-328, Rand Corporation.

Kaufmann, E. (2016). On bayesian index policies for sequential resource allocation. *arXiv:1601.01190*.

Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pages 199–213. Springer.

Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4 – 22.

Liu, H., Liu, K., and Zhao, Q. (2013). Learning in a changing world: Restless multiarmed bandit with unknown dynamics. 59(3):1902–1916.

Mahajan, A. and Teneketzis, D. (2008). Multi-armed bandits. In *Foundations and Applications of Sensor Management*, pages 121–151. Springer-Verlag.

Olszewski, W. and Weber, R. (2015). A more general pandora rule? *Journal of Economic Theory*, 160:429 – 437.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58(5):527–535.

Song, N.-O. and Teneketzis, D. (2004). Discrete search with multiple sensors. *Mathematical Methods of Operations Research*, 60(1):1–13.

Tekin, C. and Liu, M. (2012). Online learning of rested and restless bandits. *IEEE Transactions on Information Theory*, 58(8):5588–5611.

Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. *Ann. Appl. Probab.*, 4(1):194–199.

Varaiya, P., Walrand, J., and Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: The discounted case. *IEEE Trans. Autom. Control*, 30(5):426–439.

Verloop, I. M. (2016). Asymptotically optimal priority policies for indexable and nonindexable restless bandits. *Ann. Appl. Probab.*, 26(4):1947–1995.

Weber, R. (1992). On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.*, 2(4):1024–1033.

Weber, R. (2016). Mutli-armed bandits and the Gittins index theorem. http://www.statslab.cam.ac.uk/~rrw1/oc/ocgittins.pdf.

Weitzman, M. L. (1979). Optimal search for the best alternative. *Econometrica: Journal of the Econometric Society*, pages 641–654.

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149.

Yao, Y.-C. (2006). *Some results on the Gittins index for a normal reward process*, volume Volume 52 of *Lecture Notes–Monograph Series*, pages 284–294. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Trans. Inf. Theory*, 23(3):337–343.

Ziv, J. and Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory*, 24(5):530–536.