# Chapter 2
# Stochastic Multi-armed Bandit

**Abstract** In this chapter, we present the formulation, theoretical bound, and algorithms for the stochastic MAB problem. Several important variants of stochastic MAB and their algorithms are also discussed including multiplay MAB, MAB with switching costs, and pure exploration MAB.

## 2.1 Problem Formulation

In stochastic MAB problems, the reward processes do not depend on the players' actions. In this case, the reward processes can be viewed as being generated by environments. A $K$-armed bandit is defined by random variables $X_{i,n}$ for $1 \leq i \leq K$, and $n \geq 1$, where each $i$ is the index of a gambling machine (i.e., the "arm" of a bandit) and $n$ is the number of plays of the respective machine. Successive plays of machine $i$ yield rewards $X_{i,1}, X_{i,2}, \ldots$, which are IID according to an unknown distribution with unknown expectation $\mu$. In its most basic formulation, the rewards are also independent across different machines, namely, $X_{i,s}$ and $X_{j,t}$ are independent for each $1 \leq i < j \leq K$ and $s, t \geq 1$.

A policy, or an allocation strategy, $\pi$ is an algorithm that chooses the next machine to play based on the sequence of past plays and observed rewards. Let $T_i^{\pi}(n)$ be the number of times that machine $i$ has been played by $\pi$ during the first $n$ plays. Denote by $I_t^{\pi}$ the machine played at time $t$. To quantify the performance of a sequence of plays, the concept of *regret* is introduced. It measures the difference between the reward which could have been accumulated if the optimal action is selected, and the reward occurred by playing the actual actions according to the allocation strategy up to time $n$. Formally,

**Definition 2.1** (*Regret*) The cumulative *regret* of policy $\pi$ up to time $n$ is defined as the difference between the rewards achieved by following the optimal action instead of choosing the actual actions $I_1^{\pi}, \cdots, I_n^{\pi}$, that is,

$$R_n^{\pi} = \max_{i=1,\cdots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t^{\pi},t}. \tag{2.1}$$



R. Zheng and C. Hua, *Sequential Learning and Decision-Making in Wireless Resource Management*, Wireless Networks, DOI 10.1007/978-3-319-50502-2_2

Since both the rewards and player's actions are stochastic, the following two forms of average regrets have been defined:

**Definition 2.2** (*Expected regret*) The expected regret of playing $I_1^\pi, I_2^\pi, \ldots$ is

$$\mathbb{E}\left[R_n^\pi\right] = \mathbb{E}\left[\max_{i=1,\cdots,K} \sum_{t=1}^{n} X_{i,t} - \sum_{t=1}^{n} X_{I_t^\pi,t}\right], \tag{2.2}$$

and

**Definition 2.3** (*Pseudo-regret*) The pseudo-regret of playing $I_1^\pi, I_2^\pi, \ldots$ is

$$\overline{R_n^\pi} = \max_{i=1,\ldots,K} \mathbb{E}\left[\sum_{t=1}^{n} X_{i,t} - \sum_{j=1}^{K} X_{I_t^\pi,t}\right]. \tag{2.3}$$

In both definitions, the expectation $\mathbb{E}[\cdot]$ is taken with respect to the random draw of both rewards and the actions due to the policy. Clearly, pseudo-regret is a weaker form of regret as one competes against the action which is optimal only in expectation. The expected regret, in contrast, is with respect to the action which is optimal for the sequence of reward realizations. Therefore, we have $\overline{R_n^\pi} \leq \mathbb{E}\left[R_n^\pi\right]$. In subsequent discussion, unless noted otherwise, we consider pseudo-regrets only and thus drop the term "pseudo".

Let $\mu^* \stackrel{def}{=} \max_{1\leq i\leq K} \mu_i$, namely, the expected reward of the best arm, and $i^* \in \arg\max_{i=1,\ldots,K} \mu_i$. Equation (2.3) can be further simplified as,

$$\overline{R_n^\pi} = n\mu^* - \sum_{j=1}^{K} \mu_j \mathbb{E}\left[T_j^\pi(n)\right] \tag{2.4}$$

$$= n\mu^* - \mathbb{E}\left[\sum_{t=1}^{n} \mu_{I_t^\pi}\right]. \tag{2.5}$$

A couple of comments are in order to better understand the physical meaning of the regret defined above and the nature of MAB strategies.

*Remark 2.1* From (2.5), it is easy to see that $\overline{R_n^\pi}$ is positive and nondecreasing with $n$. Thus, in designing MAB policies, one can at best strive for reducing the growth rate of the regret. In general, sublinear regrets (with respective to time $n$) are desirable as linear regrets can be trivially achieved. To see why this is the case, let us consider two naive policies. The first policy, called the *super-conservative-exploitation (SCE) policy*, simply sticks to the first machine, say $I$, in sight and repeatedly plays it. The regret of the SCE policy can be easily computed as $n(\mu^* - \mu_I) = \Theta(n)$. The other extreme would be a policy, called *hyperexploration-with-commitment-phobia (HECP)*, that uniformly at random chooses a machine out of the $K$ machines at each play. Its regret can be computed as $n(\mu^* - \bar{\mu}) = \Theta(n)$, where $\bar{\mu} = \frac{1}{K}\sum_{j=1}^{K} \mu_j$ is the

average expected reward. Behaviors of the two extreme policies reveal the necessity of carefully balancing the trade-off between exploitation and exploration.

*Remark 2.2* A relevant question is whether sublinear regrets are generally attainable. From (2.5), we see that to achieve sublinear regrets, $T_{i^*}^{\pi}(n)$, i.e., the number of times the optimal arm is played by time $n$ should grow superlinearly with $n$ (note that this is not true with HECP). In other words, the policy should choose the optimal arm more and more often. It is equivalent to say that the strategy converges to the optimal strategy.[1] This, intuitively, is not hard to do as long as we observe the reward from each arm *often enough* such that the best arm can be *eventually* identified and exploited. Clearly, the terms *often enough* and *eventually* need to be rigorously quantified. Nevertheless, the key insight is that there is a tight connection between stochastic MAB and parameter estimation.

## 2.2 Theoretical Lower Bound

From the discussion in Sect. 2.1, we aim to develop policies for stochastic MAB problems that achieve sublinear regrets. It is important to first establish the lower bound for the regrets of arbitrary policies.

In their seminar work [LR85], Lai and Robbins found, for specific families of reward distributions (Bernoulli indiced by a single real parameter), the lower bound for regret is logarithm in time. Specifically, for $p, q \in [0, 1]$, we denote by $D(p||q)$ the Kullback–Leibler (KL) divergence between a Bernoulli of parameter $p$ and a Bernoulli of parameter $q$, defined as

$$D(p||q) = p \ln \frac{p}{q} + (1 - p) \ln \frac{1 - p}{1 - q}.$$

Let $\Delta_i = \mu^* - \mu_i, i = 1, 2, \ldots, K$ be the suboptimality parameters. The following result was proven in [LR85].

**Theorem 2.1** (Distribution-dependent lower bound [LR85]) *Consider a strategy that satisfies* $\mathbb{E}[T_i(n)] = o(n^{\alpha})$ *for any set of Bernoulli reward distributions, any arm* $i$ *with* $\Delta_i > 0$, *and any* $\alpha > 0$. *Then, for any set of Bernoulli reward, we have*

$$\liminf_{n \to +\infty} \frac{\overline{R_n}}{\ln n} \geq \sum_{i:\Delta_i > 0} \frac{\Delta_i}{D(\mu_i||\mu^*)}.$$

[1] This type of strategies are also called *no-regret policies*. But this term is confusing and is thus omitted here.

Theorem 2.1 is quite a remarkable result. It states any policy that plays any inferior arm subpolynomial number of times over $n$ would incur at least logarithmic regrets asymptotically. Additionally, the constant multiplier of the logarithmic regret bound is determined by the suboptimality parameters (e.g., $\Delta_i$) and the KL divergences between the inferior arms and the optimal one.

Let $c = \sum_{j:\Delta_i>0} \frac{\Delta_i}{D(\mu_i||\mu^*)}$. From Pinsker's inequality and the fact that $\ln x \leq x - 1$, it is easy to show that

$$2(\mu_i - \mu^*)^2 \leq D(\mu||\mu^*) \leq \frac{(\mu_i - \mu^*)^2}{\mu^*(1-\mu^*)}.$$

Therefore,

$$\sum_{j:\Delta_i>0} \frac{\mu^*(1-\mu^*)}{\Delta_i} \leq \sum_{j:\Delta_i>0} \frac{\Delta_i}{D(\mu_i||\mu^*)} \leq \sum_{j:\Delta_i>0} \frac{1}{2\Delta_i} \tag{2.6}$$

Consider that the arms are arranged in the descending order of their expected rewards, namely, $\mu^* = \mu_1 \geq \mu_2 \geq \ldots \mu_K$, and, $0 = \Delta_1 \ll \Delta_2 \ll \ldots \ll \Delta_K$. From (2.6), we can see that $c$ is roughly bounded from below by $\frac{\mu^*(1-\mu^*)}{\Delta_2}$, where $\Delta_2$ is the difference between the expected rewards of the best and the second best arm. This should come at no surprise because if the expected rewards of the top two arms are similar, it is difficult to distinguish them through few observations. Thus, the lower bound reveals a key property of sequential learning, namely, the smaller the difference between the expected rewards of the top two actions, the slower it takes for a policy to converge.

One may question that the subpolynomial condition on the number of times inferior arms played in Theorem 2.1 is too restrictive. The answer is negative. As will be discussed in subsequent sections, there exist policies that achieve the same order in regrets as the lower bound (also called *order-optimal policies*). Clearly, these policies satisfy the subpolynomial condition as well. Furthermore, the condition in fact provides insights on the design of good policies—the number of times that inferior arms played should be kept subpolynomial.

## 2.3 Algorithms

One principle in designing sequential learning policies for the stochastic MAB problem is the so-called *optimism in face of uncertainty*. Consider the player has accumulated some data on the environment and must decide which arm to select next. First, a set of "plausible" environments (or hypothesis) that "agree" with the data are constructed. Next, the most "favorable" environment is identified in the set. Then, the decision that is the most optimal in this most favorable and plausible environment shall be made. We will make concrete the different aspects of the principle through the discussion of policies for stochastic MAB.

### 2.3.1 Upper Confidence Bound (UCB) Strategies

Our discussion of UCB strategies follows that of [BC12], which assumes that the distribution of rewards $X$ satisfies the following moment conditions. There exists a convex function $\psi$ on the reals such that, for all $\lambda > 0$,

$$\ln \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \psi(\lambda) \text{ and } \ln \mathbb{E}\left[e^{\lambda(\mathbb{E}[X]-X)}\right] \leq \psi(\lambda). \tag{2.7}$$

When the reward distributions have support in $[0, 1]$, one can take $\psi(\lambda) = \frac{\lambda^2}{8}$. In this case, (2.7) is also known as Hoeffding's lemma. Denote $\psi^*(\varepsilon)$ the Legendre–Fencher transform of $\psi$, namely,

$$\psi^*(\varepsilon) = \sup_{\lambda \in \mathbb{R}} (\lambda\varepsilon - \psi(\lambda)).$$

Let $\hat{\mu}_{j,s}$ be the sample mean of rewards obtained by pulling arm $j$ for $s$ times. In distribution, $\hat{\mu}_{j,s}$ is equal to $\frac{1}{s}\sum_{t=1}^{s} X_{j,t}$. The $(\alpha, \psi)$-UCB strategy, where $\alpha > 0$ is an input parameter, is an index-based policy. The index is the sum of two terms in (2.8). The first term is the sample mean of rewards obtained so far. The second term is related to the size of the one-sided confidence interval for the sample mean within which the true expected reward falls with high probability.

**Algorithm 2.1**: $(\alpha, \psi)$-UCB

**Init:** *Play each machine once.*
**Init:** $T_i(K) = 1$, $j = 1, 2, \ldots K$.
**for** *t=K+1, K+2, ..., n* **do**

$$I_t = \arg\max_{i=1,\ldots,K} \left[\hat{\mu}_{j,T_i(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right)\right]. \tag{2.8}$$

Play arm $I_t$ and observe $X_{I_t,t}$.

From Markov's inequality and (2.7), one obtains that

$$\mathbb{P}\left(\mu_j - \hat{\mu}_{j,s} > \varepsilon\right) \leq e^{-s\psi^*(\varepsilon)}, \forall j \tag{2.9}$$

In other words, with probability at least $1 - \delta$,

$$\hat{\mu}_{j,s} + (\psi^*)^{-1}\left(\frac{1}{s}\ln\frac{1}{\delta}\right) > \mu_j. \tag{2.10}$$

Set $\delta = t^{-\alpha}$ and $s = T_j(t-1)$ in (2.10). With probability at least $1 - t^{-\alpha}$, the following holds:

$$\hat{\mu}_{j,T_j(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_j(t-1)}\right) > \mu_j. \tag{2.11}$$

Furthermore, (2.10) holds with probability at least $1 - t^{-\alpha}$, which diminishes as $t$ grows for $\alpha > 0$.

Revisiting the optimism in face of uncertainty principle, we see that the $(\alpha, \psi)$-UCB policy constructs based on past observations, hypotheses of the expected reward of each action ("plausible environments"), and picks the action with the highest plausible rewards ("the most optimal in the most plausible environments").

It has been proven in [BC12] that the $(\alpha, \psi)$-UCB policy achieves logarithmic regrets. Since the proof technique is representative of UCB-like policies, we provide a proof sketch here.

**Theorem 2.2** (Pseudo-regret of $(\alpha, \psi)$-UCB [BC12]) *Assume that the reward distributions satisfy* (2.7)*. Then* $(\alpha, \psi)$*-UCB with* $\alpha > 2$ *satisfies,*

$$\overline{R_n} \leq \sum_{i:\Delta_i>0} \left( \frac{\alpha \Delta_i}{\psi^*(\Delta_i/2)} \ln n + \frac{\alpha}{\alpha - 2} \right). \tag{2.12}$$

*Proof* The main idea is to show that suboptimal arms are played at most logarithmic number of times among $n$ plays. If $I_t = i$, one of the following three inequalities must be true for any suboptimal arm $i$,

$$\hat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_{i^*}(t-1)}\right) \leq \mu^* \tag{2.13}$$

$$\hat{\mu}_{i,T_i(t-1)} > \mu_i + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right) \tag{2.14}$$

$$T_i(t-1) < \frac{\alpha \ln t}{\psi^*(\Delta_i/2)}. \tag{2.15}$$

To see why this is true, assume all inequalities are false. We have

$$\begin{aligned} \hat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_{i^*}(t-1)}\right) &> \mu^* \\ &= \mu_i + \Delta_i \\ &\overset{\text{negation of (2.15)}}{\geq} \mu_i + 2(\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right) \\ &\overset{\text{negation of (2.14)}}{\geq} \hat{\mu}_{i,T_i(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right) \end{aligned}$$

The last inequality implies $I_t \neq i$ and thus a contradiction.

Let $l = \left\lceil \frac{\alpha \ln n}{\psi^*(\Delta_i/2)} \right\rceil$. It is easy to show that

$$
\begin{aligned}
T_i(n) &= \sum_{t=1}^{n} \mathbb{I}_{\{I_t=i\}} \\
&\le l + \sum_{t=1}^{n} \mathbb{I}_{\{I_t=i, T_i(t-1)\ge l\}} \\
&\le l + \sum_{t=l+1}^{n} \mathbb{I}_{\{(2.13)\}} + \mathbb{I}_{\{(2.14)\}}
\end{aligned}
$$

Taking expectation on both sides, we have

$$
\begin{aligned}
\mathbb{E}[T_i(n)] \quad \le \quad & l + \sum_{t=l+1}^{n} \mathbb{P}\left(\hat{\mu}_{i^*,T_{i^*}(t-1)} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_{i^*}(t-1)}\right) \le \mu^*\right) \\
& + \sum_{t=l+1}^{n} \mathbb{P}\left(\hat{\mu}_{i,T_i(t-1)} > \mu_i + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right)\right) \\
\overset{\text{by union bound}}{\le} \quad & l + \sum_{t=l+1}^{n} \sum_{s=1}^{t} \mathbb{P}\left(\hat{\mu}_{i^*,s} + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{s}\right) \le \mu^*\right) \\
& + \sum_{t=l+1}^{n} \sum_{s=l}^{t} \mathbb{P}\left(\hat{\mu}_{i,T_i(t-1)} > \mu_i + (\psi^*)^{-1}\left(\frac{\alpha \ln t}{T_i(t-1)}\right)\right) \\
\overset{\text{by (2.11)}}{\le} \quad & l + 2\sum_{t=l+1}^{n} \sum_{s=1}^{t} t^{-\alpha} \\
\le \quad & l + 2\sum_{t=l+1}^{\infty} t^{-\alpha+1} \\
\le \quad & \frac{\alpha \ln n}{\psi^*(\Delta_i/2)} + \frac{\alpha}{\Delta_i(\alpha-2)}.
\end{aligned}
$$

Finally, since $\overline{R_n} = \sum_{j:\Delta_i>0} \Delta_j T_i(n)$, we have the statement in the theorem. □

When the rewards are characterized by [0, 1]-valued random variables, let $\psi(\lambda) = \frac{\lambda^2}{8}$ and thus $\psi^*(\varepsilon) = 2\varepsilon^2$. Substituting the corresponding term in (2.12), we obtain the following regret bound for the UCB policy:

$$
\overline{R_n} \le \sum_{j:\Delta_i>0} \left(\frac{2\alpha}{\Delta_i} \ln n + \frac{\alpha}{\alpha-2}\right). \tag{2.16}
$$

This gives rise to the $\alpha$-UCB policy, which at time $t$ plays the arm

$$
I_t = \arg\max_{j=1,\ldots,K} \hat{\mu}_{j,T_i(t-1)} + \sqrt{\frac{\alpha \ln t}{2T_i(t-1)}}.
$$

It is identical to the UCB1 policy first proposed by Auer et al. with $\alpha = 4$ in [ABF02].

**Algorithm 2.2**: UCB1-normal Policy

**for** $t = 1, \ldots, n$ **do**
  Play any machine that has been played less than $\lceil 8 \log t \rceil$ times.
  Otherwise, play machine $i$ that maximizes

$$\hat{\mu}_i + \sqrt{16 V_{i,T_i(t-1)} \cdot \frac{\ln(t-1)}{T_i(t-1)-1}}. \tag{2.17}$$

  $T_i(t) = T_i(t-1) + 1$;
  Update $\hat{\mu}_i$ and $V_{i,T_i(t)}$ upon observing the reward.

Another interesting special case is when the rewards follow Gaussian distributions with standard deviations bounded above by $\sigma_{max}$. In this case, using the moment generation function, one can show that $\psi(\lambda) = \frac{1}{2}\sigma_{max}^2\lambda^2$ and thus $\psi^*(\varepsilon) = \frac{1}{2}\frac{\varepsilon^2}{\sigma_{max}^2}$. This allows us to derive a $\alpha$-normal-UCB policy with a logarithmic regret bound. However, the inclusion of the maximum standard deviation term $\sigma_{max}$ requires prior knowledge of the reward processes, which may not always be available. This difficulty can be attributed to the fact that the $(\alpha, \psi)$-UCB policy only takes into account sampling means. Consideration of sampling variances gives rise to a class of UCB policies with tighter regret bounds [AMS09]. In what follows, we only discuss one such policy called UCB-normal proposed by Auer [ABF02].

Consider the rewards $X$ of each action follow normal distribution with unknown mean and variance. The UCB1-normal policy (Algorithm 2.2) utilizes the estimation of sample variances as $V_{j,s} = \frac{1}{s}\sum_{t=1}^{s}(X_{j,t} - \hat{\mu}_{j,s})^2 = \frac{1}{s}\sum_{t=1}^{s} X_{j,t}^2 - \hat{\mu}_{j,s}^2$. Theorem 2.3 states the regret bound of the UCB1-normal policy. The correctness of the results is based on certain bounds on the tails of the $\chi^2$ and the Student's t-distribution that can only be verified numerically.

**Theorem 2.3** (Regret bound for UCB1-normal [ABF02]) *For all $K > 1$, if policy UCB1-normal is run on $K$ machines having normal reward distributions $P_1$, $P_2, \ldots, P_K$, then its expected regret after any number $n$ of plays is at most*

$$256\left(\sum_{i:\Delta_i>0}\frac{\sigma_i^2}{\Delta_i}\right)\log n + \left(1 + \frac{\pi^2}{2} + 8\log n\right)\sum_{j=1}^{K}\Delta_j,$$

*where $\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2$ are the variances of the normal distributions $P_1, P_2, \ldots, P_K$.*

### 2.3.2 ε-Greedy Policy

A simple and well-known heuristic for the bandit problem is the so-called $\varepsilon$-greedy rule. The idea is to play with probability $1-\varepsilon$ the machine with the highest average reward, and with probability $\varepsilon$ a randomly chosen machine. $\varepsilon$ is also called the *exploration* probability. Clearly, with a fixed $\varepsilon$, the average regret incurred will grow linearly. A simple fix is to let $\varepsilon$ go to zero at a certain rate. In [ABF02], Auer show that a rate of $1/t$, where $t$ is the index of the current play, allows for a logarithmic bound on the regret. The resulting policy, $\varepsilon_n$-greedy is described in Algorithm 2.3.

**Algorithm 2.3**: $\varepsilon_t$-Greedy Policy

**Input**: $c > 0$ and $0 < d < 1$

**Init:** *Define the sequence* $\varepsilon_t \in (0, 1]$, $t = 1, 2, \ldots,$ *by*

$$\varepsilon_t \stackrel{def}{=} \min\left\{1, \frac{cK}{d^2 t}\right\}.$$

**for** $t = 1, 2, \ldots, n$ **do**
- $I_t = \arg\max_{j=1,\ldots,K} \hat{\mu}_j$.
- Draw $u$ uniformly from $[0, 1]$.
- **if** $u > \varepsilon_t$ **then**
  - Play arm $I_t$.
- **else**
  - Play a random arm.

**Theorem 2.4** (Regret bound for $\varepsilon_n$-greedy [ABF02]) *For all $K > 1$ and for all reward distribution $P_1, \ldots, P_K$ with support in $[0, 1]$, if policy $\varepsilon_n$-greedy is run with input parameter,*

$$0 < d \leq \min_{j:\Delta_j > 0} \Delta_j,$$

*then the probability that after any number $n \geq cK/d$ of plays $\varepsilon_n$-greedy chooses a suboptimal machine $j$ is at most*

$$\frac{c}{d^2 n} + 2\left(\frac{c}{d^2} \ln \frac{(n-1)d^2 e^{\frac{1}{2}}}{cK}\right)\left(\frac{cK}{(t-1)d^2 e^{\frac{1}{2}}}\right)^{c/5d^2} + \frac{4e}{d^2}\left(\frac{cK}{(n-1)d^2 e^{\frac{1}{2}}}\right)^{c/2}. \tag{2.18}$$

*Remark 2.3* To achieve logarithmic regrets in the $\varepsilon$-greedy algorithm, one needs to know the lower bound $d$ on the difference between the reward expectations of the best and the second best machines. One remedy is to replace $d^2$ with a slowly decreasing term, for instance, $\varepsilon_t \stackrel{def}{=} \min\left\{1, \frac{\ln \ln t}{t}\right\}$. This would result in a regret that grows in the order of $\ln \ln(t) \cdot \ln(t)$, which is (slightly) worse than the bound in (2.18).

**Algorithm 2.4**: Thompson Sampling for general stochastic bandits

**Init:** *set* $S_j = 0, F_j = 0$ *for each arm* $j = 1, 2, \ldots, K$.
**for** $t = 1, 2, \ldots,$ **do**
  For each arm $j = 1, ..., K$, sample $\theta_j(t)$ from the $Beta(S_j + 1, F_j + 1)$ distribution.
  Play arm $j(t) := \arg\max_j \theta_j(t)$ and observe reward $\tilde{r_t}$.
  Perform a Bernoulli trial with success probability $\tilde{r_t}$ and observe output $r_t$.
  **if** $r_t = 1$ **then**
    $S_{j(t)} = S_{j(t)} + 1$
  **else**
    $F_{j(t)} = F_{j(t)} + 1$.

### 2.3.3 Thompson Sampling Policy

Thompson sampling (TS) [Tho33] is a natural strategy for MAB problems and has gained a lot of interests recently due to its flexibility to incorporate prior knowledge on the arms. The basic idea is to assume a simple prior distribution on the parameters of the reward distribution of every arm, and at each time step, play an arm according to its posterior probability of being the best arm.

Agrawal and Goyal [AG12] established the first logarithmic regret bound for stochastic MAB using TS. The reward distributions are assumed to have support on [0, 1]. The basic procedure is described in Algorithm 2.4. In Algorithm 2.4, the probability of a successful Bernoulli when arm $i$ is played evolves toward its mean reward. Thus, $\theta_i(t)$ sampled from the Beta distribution would approach the mean reward as arm $i$ is played more times.

The regret bound for $K$ arms using Thompson sampling is given by the following theorem:

**Theorem 2.5** (Regret bounds for TS [AG12]) *For the K-armed stochastic bandit problem with support in* [0, 1]*, Algorithm 2.4 has an expected regret bound*

$$O\left(\left(\sum_{j:\Delta_j>0}\frac{1}{\Delta_j^2}\right)^2 \ln n\right)$$

*in time n. An alternative bound can be obtained as*

$$O\left(\frac{\Delta_{max}}{\Delta_{min}^3}\left(\sum_{j:\Delta_j>0}\frac{1}{\Delta_j^2}\right)\ln n\right),$$

*where* $\Delta_{max} = \max_{j:\Delta_j>0}\Delta_j$ *and* $\Delta_{max} = \min_{j:\Delta_j>0}\Delta_j$.

## 2.4 Variants of Stochastic Multi-armed Bandit

Many variants of the stochastic MAB have been investigated in literature. In this section, we discuss a few variants that are the most relevant in addressing resource management problems in wireless networks.

### *2.4.1 Multiplay MAB*

In multiplay MAB problems, a player can pull multiple arms in each round. This can be used to model problems, where a decision maker needs to take multiple actions simultaneously. For instance, consider an advertiser who creates an ad campaign to promote her products through a search engine. To participate in search ad auctions, she needs to choose multiple keywords for her campaign. Each keyword can be treated as an arm. As another example, $m$ cooperative users try to access $K$ wireless channels concurrently. The goal is to maximize their sum throughput. In this example, each wireless channel is an arm. If more than two users access the same channel, the reward is zero. Thus, the $m$ channels selected shall be distinct.

One can trivially extend the policies for single-play MAB such as $(\alpha, \phi)$-UCB policy or the $\varepsilon$-greedy policy, by viewing each of the $\binom{K}{m}$ possible combinations as an arm. However, doing so ignores the dependence among the rewards from the arms and results in slower learning rate. Recall that the lower bound of the pseudo-regret for Bernoulli bandit in Theorem 2.1 is proportional to $\sum_{j:\Delta_j>0} \frac{\Delta_j}{D(\mu_j||\mu^*}$. When the number of arms is large, the lower bound tends to be higher. Though the regret still grows logarithmically, the constant multiplier is bigger.

In [AVW87], Anantharam et. al formulated the multiplay MAB problem, where $K$ arms follow one-parameter reward distributions with the unknown parameter set $(\theta_1, \theta_2, \ldots, \theta_K)$. Successive plays of each arm produce IID rewards. At each stage, one is required to play a fixed number, $m$, of the arms ($1 \leq m \leq K$). Upon playing $m$ arms, one can observe the reward from each of the arms. An asymptotically efficient allocation rule has been proposed. However, unless the reward distributions are Bernoulli, Poisson, or normal, it is computationally prohibitive to update the statistics required for allocation upon new observation under the proposed allocation rule.

More recently, a Thompson sampling based approach has been proposed for the multiplay MAB problem in [KHN15], where each arm is associated with Bernoulli distribution with unknown means. Similar to the Thompson sampling policy for single-play MAB discussed in Sect. 2.3.3, the success probability $\theta_j$ of arm $j$ is assumed to follow a Beta distribution. At time $t$, having observed $S_j(t)$ successes (reward = 1) and $F_j(t)$ failures (reward = 0), the posterior distribution of $\theta_j$ is updated as $Beta(S_j(t) + 1, F_j(t) + 1)$. The main difference in the multiplay case is to select top $m$ arms ranked by the posterior sample $\theta_j(t)$ from $Beta(S_j(t) + 1, F_j(t) + 1)$.

Komiyama et al. proved that such a simple policy can indeed achieve an any-time logarithmic regret bound.

In [GKJ12], a combinatorial optimization problem with unknown variables and linear rewards has been formulated. Consider a discrete time system with $K$ unknown random processes $X_j(t)$, $1 \leq j \leq K$. $X_j(t)$ is IID with finite support in $[0, 1]$ and mean $\mu_j$. At each decision period $t$, a $K$-dimensional action vector $\mathbf{a}(t)$ is selected from a finite set $\mathscr{F}$. Upon selecting an action $\mathbf{a}(t)$, the value of $X_j(t)$ is observed for all $j$'s such that $a_j(t) \neq 0$. The combinatorial optimization problem is thus formulated as

$$\begin{array}{ll} \max & \sum_t \sum_{j=1}^{K} a_j(t) X_j(t) \\ s.t. & \mathbf{a}(t) \in \mathscr{F}. \end{array} \tag{2.19}$$

When the action space $\mathscr{F}$ is restricted to binary vectors in $K$-dimension with at most $m$ nonzero elements, it is easy to see that the problem is equivalent to the multiplay stochastic MAB problem. Gai et al. proposed a UCB-like policy and proved that it incurs logarithmic regrets over time. At time $t$, the policy plays an action $\mathbf{a}$ that solves the following maximization problem:

$$\mathbf{a} = \arg\max_{\mathbf{a} \in \mathscr{F}} \sum_{j \in \mathscr{A}_{\mathbf{a}}} a_j \left( \hat{\mu}_j + \sqrt{\frac{(m+1) \ln t}{T_j(t-1)}} \right), \tag{2.20}$$

where $\mathscr{A}_{\mathbf{a}} = \{j | a_j \neq 0, j = 1, 2, \ldots, K\}$, $\hat{\mu}_j$ is the sample mean of observed rewards of arm $j$, and $m = \min\{\max_{\mathbf{a} \in \mathscr{F}} |\mathscr{A}_{\mathbf{a}}|, K\}$ (i.e., the maximum number of arms that can be played in each round). In the special case of the multiplay stochastic MAB, we can see that the above policy plays the top $m$ arms with the highest indices.

### 2.4.2 MAB with Switching Costs

To this end, our discussion of MAB problems does not consider the costs of switching between arms. In practice, switching costs are important concerns [Jun04]. For example, a worker who switches jobs must pay nonnegligible costs. In job scheduling, switching jobs from one machine to another incurs a variety of setup costs. In wireless networks, changing the operational channels of a radio transmitter incurs nonnegligible delay during which no data can be transmitted. The policies we discuss so far are designed to optimize regrets by playing suboptimal arms less and less often. However, they do not explicitly control the frequency of switching between arms.

Consider $K$ arms whose reward processes are IID with support on $[0, 1]$ and unknown means $\mu_1, \mu_2, \ldots, \mu_K$. Let

$$S_n(j) = \sum_{t=2}^{n} \mathbb{I}_{\{I_{t-1}=j, I_t \neq j\}} \tag{2.21}$$

denote the number of times to switch from arm $j$ to another arm. The switching regret of policy $\pi$ is then defined as

$$SW_n^\pi = C \sum_{j=1}^{K} \mathbb{E}\left[S_n(j)\right], \tag{2.22}$$

where $C > 0$ is the fixed switching cost. Therefore, the total regret incurred by a policy $\pi$ consists two parts, the regret from sampling suboptimal arms, called sampling regret $\overline{R_n^\pi}$ in this context, and the switching regret $SW_n^\pi$.

Banks and Sundaram [BS94], and later Sundaram [Sun05] have shown the suboptimality of index-based policies when switching costs are nonnegligible.[2] In response, researchers have taken three different lines of approaches: characterization of the optimal policy, exact derivation of the optimal policy in restricted settings, and development of order-optimal policies. An excellent survey on this topic can be found in [Jun04]. More recently, Guha and Munagala have investigated the problem, where the switching costs are determined by distances in a metric space [GM09].

Order-optimal policies aim to control the growth of cumulative switching costs such that it is slower than that of the sampling regret from the reward processes. Such policies typically utilize "block" sampling, namely, an action once selected, is in play for a period of time, called an *epoch*. The epoch length is initially short when the uncertainty in the reward processes is high (thus more exploration), and increases as more knowledge is gained so as to amortize the switching costs. Next, we present one such policy based on UCB that was originally proposed by Auer in [ABF02]. Let $\tau(r) = \lceil (1+\alpha)^r \rceil$, where $\alpha > 0$. Clearly, $\tau(r+1) - \tau(r) \approx \alpha(1+\alpha)^r$ grows exponentially with $r$. We further denote

$$a_{t,r} = \sqrt{\frac{(1+\alpha)\ln(et/\tau(r))}{2\tau(r)}}. \tag{2.23}$$

Policy UCB2 in Algorithm 2.5 is an index-based policy and plays arms over epochs that increase exponentially in length over rounds. It has been proven in [ABF02] that policy UCB2 attains logarithmic regrets in time. Furthermore, it is easy to show that when the switching cost is constant between any two arms, the cumulative switching cost incurred is $O(\ln \ln n)$.

### 2.4.3 Pure Exploration MAB

The last variant of stochastic MAB problems we discuss is the so-called *pure exploration* MAB first studied by Bubeck et al. in [BMS11]. Unlike conventional stochastic MAB problems, where exploitation needs to be performed at the same time as

[2] MAB with switching costs can be cast as a restless bandit problem discussed in Chap. 3.

**Algorithm 2.5**: UCB2 Policy

**Input**: $0 < \alpha < 1$.
**Init:** *Set* $r_j = 0$, *for* $j = 1, \ldots, K$. *Play each machine once.*
**for** $t = 1, 2, \ldots,$ **do**
  $j = \arg\max_{j=1,\ldots,K}\{\hat{\mu}_j + a_{t,r_j}\}$, where $\hat{\mu}_j$ is the average reward obtained from machine $j$.
  Play machine $j$ exactly $\tau(r_j + 1) - \tau(r_j)$ times.
  Set $r_j \leftarrow r_j + 1$

exploitation, in pure exploration MAB, the two operate at different stages. A player first selects and observes the rewards of arms for a given number of times $n$ (not necessarily known in advance). She is then asked to provide a recommendation, either deterministic or in the form of a probability distribution over the arms. The player's strategy is evaluated by the difference between the average payoff of the best arm and the average payoff obtained by her recommendation.

Pure exploration MAB is suitable for applications with a preliminary exploration phase in which costs are not measured in terms of rewards but rather in terms of resources consumed. Typically, a limited budget is associated with the resources. For instance, in wireless networks, probing available channels incurs extra delay and energy costs. In formulating channel probing as a pure exploration MAB problem, we aim to design a strategy consisting of a sequence of channels to probe given the delay or energy cost constraints. At the end of the procedure, the "best" channel for data communication is selected and no further probing is needed.

Consider $K$ arms with mean rewards $\mu_1, \mu_2, \ldots, \mu_K$. Denote $\boldsymbol{P}\{1, ..., K\}$ the set of all probability distributions over the indexes of the arms. At each round $t$, a player decides which arm to pull, denoted by $I_t$, according to a distribution $\varphi_t \in \boldsymbol{P}\{1, \ldots, K\}$ based on past rewards. After pulling the arm, the player gets to see the reward $X_{I_t, T_{I_t}}(t)$, where $T_{I_t}(t)$ is the number of times that arm $I_t$ has been pulled by round $t$. The primary task is to output at the end of each round $t$, a recommendation $\psi_t \in \boldsymbol{P}\{1, ..., K\}$ to be used to form a randomized play in a one-shot instance if the exploration phase is signaled to be over. The sequence $\psi_t$ is referred to as a recommendation strategy. The *simple regret* $r_t$ of a recommendation $\psi_t = (\psi_{j,t})_{j=1,\ldots,K}$ is defined as the expected regret on a one-shot instance of the game, if a random action is taken according to $\psi_t$. Formally,

$$r_t = r(\psi_t) = \mu^* - \mu_{\psi_t},$$

where $\mu^* = \max_{j=1,\ldots,K} \mu$ and $\mu_{\psi_t} = \sum_{j=1,\ldots,K} \psi_{j,t}\mu_j$.

From the description of the problem, we see that a player needs to devise two strategies, namely, the allocation strategy $\varphi_t$ and the recommendation strategy $\psi_t$. Interestingly, the cumulative expected regret of the allocation strategy is intrinsically related to the simple regret of the recommendation strategy. For Bernoulli reward processes, Bubeck et al. established the following results [BMS11]:

**Theorem 2.6** (Simple regret vs. pseudo-regret) *For all allocation strategies* $(\varphi_t)$ *and all functions* $f : 1, 2, \ldots \to \mathbb{R}$ *such that for all (Bernoulli) distributions* $P_1$, $P_2, \ldots, P_K$ *on the rewards, there exists a constant* $C \geq 0$ *with* $\mathbb{E}[R_n] \leq Cf(n)$. *Then, for all recommendation strategies* $(\psi_t)$ *based on the allocation strategies* $(\varphi_t)$, *there exists a constant* $D \geq 0$, *such that*

$$\mathbb{E}[r_n] \geq \frac{\min_{j:\Delta_j>0} \Delta_j}{2} e^{-Df(n)}.$$

Theorem 2.6 implies that the smaller the cumulative regret, the larger the simple regret. In [BMS11], several combinations of allocation and recommendation strategies have been analyzed. An immediate consequence of Theorem 2.6 is a lower bound on the simple regret from the fact that the cumulative regrets are always bounded by $n$:

**Corollary 1** *For allocation strategies* $(\varphi_t)$, *all recommendation strategies* $(\psi_t)$, *and all sets of* $K \geq 3$ *(distinct, Bernoulli) distributions on the rewards, there exist two constants* $\beta > 0$ *and* $\gamma \geq 0$ *such that, up to the choice of a good ordering of the considered distributions,*

$$\mathbb{E}[r_n] \geq \beta e^{-\gamma n}. \tag{2.24}$$

The lower bound in (2.24) can be achieved using a trivial uniform allocation policy that samples each arm with equal probability and a recommendation indicating the empirical best arm. However, for moderate values of $n$, strategies not pulling each arm a linear number of the times in the exploration phase can have interesting simple regrets. One such strategy was discussed in [BMS11], called $UCB(p)$, which is the same as $\alpha$-UCB policy in Sect. 2.3.1 by setting $\alpha = 2p$. Empirically, it was shown that when the number of arms is large, for moderate $n$, UCB($p$) combined with a recommendation indicating the arm with the highest sample mean or the most played arm incurs smaller simple regrets.

In [AB10], Audibert et al. proposed an elegant algorithm called Successive Reject (SR) for the pure exploration MAB when the number of rounds $n$ of the exploration phase is known. The idea is quite simple. First, the algorithm divides the $n$ rounds into $K - 1$ phases. At the end of each phase, the algorithm dismisses the arm with the lowest sample mean. In next phase, it pulls equally often each arm which has not been dismissed yet. At the end of the $n$ rounds, the last surviving arm is recommended. The length of each phase is chosen carefully such that the simple regret diminishes exponentially with $n$.

Other variations of the pure exploration problems have been investigated in literature, including the problem of finding the top-$m$ arms [KS10, Kal+12], the multi-bandit problem of finding the best arms simultaneously from several disjoint sets of arms [Gab+11], and the more general combinatorial pure exploration (CPE) problem [Che+14].

## 2.5 Summary

In this chapter, we have presented the formulation of the stochastic MAB problem and several of its important variants. Representative strategies and their regret bounds have been discussed. Most stochastic MAB problems admit simple index-based policies that can achieve logarithmic regrets in time. It is important to note that the optimality of these index-based policies relies on the assumption of IID reward processes and is expressed in terms of the growth rate in time (as opposed to the number of the arms). When the reward process of the arms are dependent, the logarithmic bounds still apply but it is important to exploit the correlation structure among the arms to make learning more efficient.

## References

[AB10] Jean-Yves Audibert and Sébastien Bubeck. "Best arm identification in multi-armed bandits". In: *COLT-23th Conference on Learning Theory*. 2010, 13–p.

[ABF02] P. Auer, N. C. Bianchi, and P. Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Mach. Learn*. 47.2-3 (May 2002), pp. 235–256. ISSN: 0885-6125.

[AG12] Shipra Agrawal and Navin Goyal. "Analysis of Thompson Sampling for the Multi-armed Bandit Problem." In: 2012.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. "Exploration-exploitation Tradeoff Using Variance Estimates in Multi-armed Bandits". In: *Theor. Comput. Sci*. 410.19 (Apr. 2009), pp. 1876–1902. ISSN: 0304-3975.

[AVW87] V. Anantharam, P. Varaiya, and J. Walrand. "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part I: I.I.D. rewards". In: *Automatic Control, IEEE Transactions on* 32.11 (Nov. 1987), pp. 968–976. ISSN: 0018-9286.

[BC12] Sébastien Bubeck and Nicol'o Cesa-Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems". In: *Foundations and Trends in Machine Learning* 5.1 (2012), pp. 1–122.

[BMS11] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. "Pure exploration in finitely-armed and continuous-armed bandits". In: *Theor. Comput. Sci*. 412.19 (2011), pp. 1832–1852.

[BS94] Jeffrey S Banks and Rangarajan K Sundaram. "Switching costs and the Gittins index". In: *Econometrica* 62.3 (1994), pp. 687–694.

[Che+14] Shouyuan Chen et al. "Combinatorial pure exploration of multi-armed bandits". In: *Advances in Neural Information Processing Systems*. 2014, pp. 379–387.

[Gab+11] Victor Gabillon et al. "Multi-bandit best arm identification". In: *Advances in Neural Information Processing Systems*. 2011, pp. 2222–2230.

[GKJ12] Y. Gai, B. Krishnamachari, and R. Jain. "Combinatorial Network Optimization with Unknown Variables: Multi-Armed Bandits with Linear Rewards and Individual Observations". In: *IEEE/ACM Transactions on Networking* 20.5 (Oct. 2012), pp. 1466–1478.

[GM09] Sudipto Guha and Kamesh Munagala. "Multi-armed bandits with metric switching costs". In: *International Colloquium on Automata, Languages, and Programming*. Springer. 2009, pp. 496–507.

[Jun04] Tackseung Jun. "A survey on the bandit problem with switching costs". In: *De Economist* 152.4 (2004), pp. 513–541.

[Kal+12] Shivaram Kalyanakrishnan et al. "PAC subset selection in stochastic multi-armed bandits". In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*. 2012, pp. 655–662.

[KHN15] Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. “Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays”. In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*. 2015, pp. 1152–1161.

[KS10] Shivaram Kalyanakrishnan and Peter Stone. “Efficient selection of multiple bandit arms: Theory and practice”. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, pp. 511–518.

[LR85] T L Lai and H. Robbins. “Asymptotically efficient adaptive allocation rules”. In: *Advances in Applied Mathematics* 6.1 (1985), pp. 4–22.

[Sun05] Rangarajan K Sundaram. “Generalized bandit problems”. In: *Social choice and strategic decisions*. Springer, 2005, pp. 131–162.

[Tho33] William R Thompson. “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples”. In: *Biometrika* 25.3/4 (1933), pp. 285–294.

http://www.springer.com/978-3-319-50501-5

Sequential Learning and Decision-Making in Wireless Resource Management

Zheng, R.; Hua, C.

2016, XIII, 118 p. 22 illus., Hardcover
ISBN: 978-3-319-50501-5