4 Finite-Armed Stochastic Bandits

A ***K*-armed stochastic bandit** is a tuple of distributions $\nu = (P_1, P_2, \dots, P_K)$, where P_i is a distribution over the reals for each $i \in [K]$. The learner and the environment interact sequentially as summarized in Fig. 4.1. In each round t the learner chooses action $A_t \in \{1, 2, \dots, K\}$, which is fed to the environment. Then the environment samples reward $X_t \in \mathbb{R}$ from distribution P_{A_t} and reveals it to the learner. The interaction between the learner (or policy) and environment induces a measure on the sequence of outcomes $A_1, X_1, A_2, X_2, \dots, A_n, X_n$ where n is the horizon. Usually the horizon n is finite, but sometimes we allow the interaction to continue indefinitely ($n = \infty$). The interaction diagram above suggests that A_t and X_t should satisfy the following assumptions:

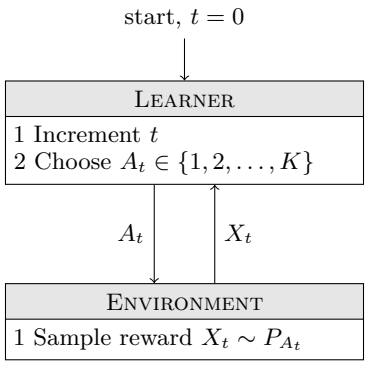


Figure 4.1 Interaction between learner and environment

(a) The conditional distribution of X_t given $A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t$ is P_{tA_t}, which captures the intuition that the environment samples X_t from P_{tA_t} in round t.

(b) $\mathbb{P}(A_t = a \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) = \pi_t(A_1, X_1, \dots, A_{t-1}, X_{t-1})$ where $\pi_1, \pi_2, \dots$ is a sequence of functions that characterize the learner with $\pi_t(a \mid a_1, x_1, \dots, a_{t-1}, x_{t-1})$ representing the probability that the learner chooses action a having observed $a_1, x_1, \dots, a_{t-1}, x_{t-1}$. The most important element of this assumption is the intuitive fact that the learner cannot use the future observations in current decisions.

A mathematician might ask on which probability space A_t and X_t are defined and the measure for which (a) and (b) are satisfied. We show how to do this in Section 4.4, but for now we move on.

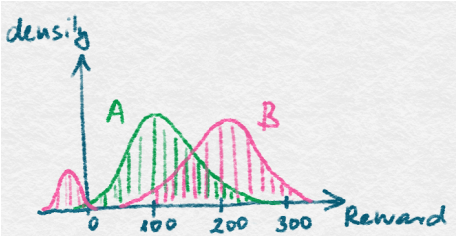
4.1 The learning objective

The learner's goal is to maximize the total reward $S_n = \sum_{t=1}^{n} X_t$, which is a random quantity that depends on the actions of the learner and the rewards sampled by the environment. This is not an optimization problem for three reasons:

1 The cumulative reward is a random quantity. Even if the reward distributions were known, then we require a measure of utility on distributions of S_n.
2 The learner does not know the distributions $(P_i)_i$ that determine the reward for each arm.
3 What is the value of n for which we are maximizing? Occasionally prior knowledge of the horizon is reasonable, but very often the learner does not know ahead of time how many rounds are to be played.

We address the first two points below. For issues relating to knowledge of the horizon it usually suffices to assume the horizon is known when designing a policy. Then if the horizon is unknown the objective is to design a new policy that does not depend on the horizon and is never much worse than what could be achieved for a known horizon. This is almost always quite easy and there exist generic approaches for making the conversion.

Expectation and risk
Suppose that S_n is the revenue of your company. The figure on the right shows the distribution of S_n for two different learners, call them A and B. Suppose you can choose between learners A and B. Which one would you choose? One choice is to go with the learner whose reward distribution has the larger expected value. This will be our default choice for stochastic bandits, but it bears remembering that there are other considerations, including the variance or tail behavior of the cumulative reward, which we will discuss occasionally. In particular, in the situation shown on the above figure, learner B achieves a higher expected total reward than A. However B has a reasonable probability of earning less than A, so a risk sensitive user may prefer learner A.



Environment classes
Even if the horizon is known in advance and we commit to maximizing the expected value of S_n, there is still the problem that $\nu = (P_i)_i$ is unknown. A policy that maximizes the expectation of S_n for one ν can lead to very poor results on another. The learner usually has partial information about the distributions $(P_i)_i$. For example, that the rewards are binary, which means that P_i is Bernoulli

Name	Symbol	Definition
Bernoulli	$\mathcal{E}_{\mathcal{B}}^K$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^K\}$
Uniform	$\mathcal{E}_{\mathcal{U}}^K$	$\{(\mathcal{U}(a_i,b_i))_i : a,b \in \mathbb{R}^K \text{ with } a_i \le b_i \text{ for all } i\}$
Gaussian (known var.)	$\mathcal{E}_{\mathcal{N}}^K(\sigma^2)$	$\{(\mathcal{N}(\mu_i,\sigma^2))_i : \mu \in \mathbb{R}^K\}$
Gaussian (unknown var.)	$\mathcal{E}_{\mathcal{N}}^K$	$\{(\mathcal{N}(\mu_i,\sigma_i^2))_i : \mu \in \mathbb{R}^K \text{ and } \sigma^2 \in [0,\infty)^K\}$
Finite variance	$\mathcal{E}_{\mathbb{V}}^K(\sigma^2)$	$\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \le \sigma^2 \text{ for all } i\}$
Finite kurtosis	$\mathcal{E}_{\text{Kurt}}^K(\kappa)$	$\{(P_i)_i : \text{Kurt}_{X\sim P_i}[X] \le \kappa \text{ for all } i\}$
Bounded support	$\mathcal{E}_{[a,b]}^K$	$\{(P_i)_i : \text{Supp}(P_i) \subseteq [a,b]\}$
Subgaussian	$\mathcal{E}_{\text{SG}}^K(\sigma^2)$	$\{(P_i)_i : P_i \text{ is } \sigma\text{-subgaussian for all } i\}$

$\text{Supp}(P)$ is the support of distribution P. The kurtosis of a random variable X is a measure of its tail behavior and is defined by $\mathbb{E}[(X-\mathbb{E}[X])^4]/\mathbb{V}[X]^2$. Subgaussian distributions have similar properties to the Gaussian and will be defined in the next chapter.

Table 4.1 Typical environment classes for stochastic bandits

for each i. We represent this knowledge by defining a set of bandits $\mathcal{E}$ for which $(P_i)_i \in \mathcal{E}$ is guaranteed. Some typical choices are listed in Table 4.1. Of course, these are not the only choices, and the reader can no doubt find ways to construct more. For example, by allowing some arms to be Bernoulli and some Gaussian, or have rewards being exponentially distributed, or Gumbel distributed, or belonging to your favorite (non-)parametric family.

The Bernoulli, Gaussian and uniform distributions are often used as examples for illustrating some specific property of learning in stochastic bandit problems. The Bernoulli distribution is in fact a natural choice - think of applications like maximizing click-through rates in a web-based environment. A bandit problem is often called a 'distribution bandit' where 'distribution' is replaced by the underlying distribution from which the payoffs are sampled. Some examples are: Gaussian bandit, Bernoulli bandit or subgaussian bandit. Similarly we say 'bandits with X' where 'X' is a property of the underlying distribution from which the payoffs are sampled. For example, we can talk about bandits with finite variance, meaning the bandit environment where the a priori knowledge of the learner is that all payoff distributions are such that their underlying variance is finite.

Some of the environment classes, like Bernoulli bandits, are **parametric** while others, like subgaussian bandits, are **nonparametric**. The distinction is the number of degrees of freedom needed to describe an element of the environment. When the number of degrees of freedom is finite it is parametric and otherwise it is non-parametric. Of course, if a learner is designed for a specific environment class $\mathcal{E}$, then we might expect that it has good performance on all bandits $\nu \in \mathcal{E}$. What do we mean by 'good'? Keep reading! Some environment classes are subsets of other classes. For example, Bernoulli bandits are a special case of bandits with a finite variance, or bandits with bounded support. Something to keep in mind is that we expect that it will be harder to achieve a good performance in a larger

class. In a way, the theory of finite-armed stochastic bandits tries to quantify this expectation in a rigorous fashion.

All the environments mentioned so far are **unstructured**, by which we mean that knowledge about the distribution of one arm does not restrict the range of possibilities for other arms. This means the only way to learn about the distribution for an arm is to play it. When we refer to finite-armed stochastic bandits with no further qualifications the reader should take it as assumed that we mean unstructured finite-armed stochastic bandits. Later we will see that much changes in **structured** bandit problems when this property does not hold.

4.2 The regret

In Chapter 1 we informally defined the regret as being the deficit suffered by the learner relative to the optimal policy. Let ν be a K-armed stochastic bandit and define $\mu_i(\nu) = \int_{-\infty}^{\infty} x dP_i(x)$, which is the mean of P_i. Then let $\mu^*(\nu) = \max_{i \in [K]} \mu_i(\nu)$ be the largest mean of all the arms. Of course $\mu_i(\nu)$ could be undefined or infinite, so for the remainder of the book we assume that $\mu_i(\nu)$ exists and is finite for all stochastic bandit instances. For stochastic bandits we define the regret of policy π in bandit ν by

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right], \tag{4.1}$$

where the expectation is taken with respect to the measure on outcomes induced by the interaction of π and ν. Minimizing the regret is equivalent to maximizing the expectation of S_n, but the normalization inherent in the definition of the regret is useful when stating results, which would otherwise need to be stated relative to the optimal action.

If the context is clear we will often drop the dependence on ν and π in various quantities. For example, by writing $R_n = n\mu^* - \mathbb{E}[\sum_{t=1}^{n} X_t]$. Similarly, when we think readers can work out ranges of symbols in a unique way, we abbreviate sums, or maxima. For example: $\mu^* = \max_i \mu_i$

The regret is always nonnegative and for every bandit ν there exists a policy π for which the regret vanishes.

LEMMA 4.1 *Let ν be a stochastic bandit environment. Then,*

(a) $R_n(\pi, \nu) \geq 0$ *for all policies* π.
(b) The policy π *choosing* $A_t \in \operatorname{argmax}_i \mu_i$ *for all* t *satisfies* $R_n(\pi, \nu) = 0$.
(c) If $R_n(\pi, \nu) = 0$ *for some policy* π *then for all* t, $A_t \in [K]$ *is optimal with probability one:* $\mathbb{P}(\mu_{A_t} = \mu^*) = 1$.

We leave the proof for the reader (Exercise 4.7). Part (b) of Lemma 4.1 shows that for every bandit ν there exists a policy for which the regret is zero (the best possible outcome). According to Part (c), achieving zero is possible if and only if the learner knows which bandit it is facing (or at least, what is the optimal arm). In general, however, the learner only knows that $\nu \in \mathcal{E}$ for some environment class $\mathcal{E}$. So what can we hope for? A relatively weak objective is to find a policy π with sublinear regret on all $\nu \in \mathcal{E}$. Formally, this objective is to find a policy π such that

$$\text{for all } \nu \in \mathcal{E}\,, \qquad \lim_{n\to\infty} \frac{R_n(\pi,\nu)}{n} = 0\,.$$

If the above holds, then at least the learner is choosing the optimal action almost all of the time as the horizon tends to infinity. One might hope for much more, however. For example that for some specific choice of $C > 0$ and $p < 1$ that

$$\text{for all } \nu \in \mathcal{E}\,, \qquad R_n(\pi,\nu) \le Cn^p\,. \tag{4.2}$$

Yet another alternative is to find a function $C : \mathcal{E} \to [0,\infty)$ and $f : \mathbb{N} \to [0,\infty)$ such that

$$\text{for all } n \in \mathbb{N},\ \nu \in \mathcal{E}\,, \qquad R_n(\pi,\nu) \le C(\nu)f(n)\,. \tag{4.3}$$

This factorization of the regret into a function of the instance and a function of the horizon is not uncommon in learning theory and appears in particular in supervised learning (for example, Györfi et al. 2002).

We will spend a lot of time in the following chapters finding policies satisfying Eq. (4.2) and Eq. (4.3) for different choices of $\mathcal{E}$. The form of Eq. (4.3) is quite general, so much time is also spent discovering what are the possibilities for f and C, both of which should be 'as small as possible'. All of the policies are inspired by the simple observation that in order to make the regret small, the algorithm must discover the action/arm with the largest mean. Usually this means the algorithm should play each arm some number of times to form an estimate of the mean of that arm, and subsequently play the arm with the largest estimated mean. The question essentially boils down to discovering exactly how often the learner must play each arm in order to have reasonable statistical certainty that it has found the optimal arm.

There is another candidate objective called the **Bayesian regret**. If Q is a prior probability measure on $\mathcal{E}$ (which must be equipped with a σ-algebra $\mathcal{F}$), then the Bayesian regret is the average of the regret with respect to the prior Q.

$$\mathrm{BR}_n(\pi, Q) = \int_{\mathcal{E}} R_n(\pi,\nu)dQ(\nu)\,, \tag{4.4}$$

which is only defined by assuming (or proving) that the regret is a measurable function with respect to $\mathcal{F}$. An advantage of the Bayesian approach is that having settled on a prior and horizon, the problem of finding a policy that minimizes the Bayesian regret is just an optimization problem. Most of this book is devoted to

analyzing the frequentist regret, but Bayesian methods are covered in Chapters 34 and 35.

4.3 Decomposing the regret

We now present a lemma that forms the basis of almost every proof for stochastic bandits. Let $\nu = (P_i)_{i=1}^K$ be a stochastic bandit and define $\Delta_i(\nu) = \mu^*(\nu) - \mu_i(\nu)$, which is called the **suboptimality gap** or **action gap** or **immediate regret** of action i. Further, let

$$T_i(t) = \sum_{s=1}^{t} \mathbb{I}\{A_s = i\}$$

be the number of times action i was chosen by the learner after the end of round t. In general, $T_k(n)$ is random, which may seem surprising if we think about a deterministic policy that chooses the same action for any fixed history. So why is $T_k(n)$ random in this case? The reason is because for all rounds t except for the first, the action A_t depends on the rewards observed in rounds $1, 2, \ldots, t-1$, which are random, hence A_t will also inherit their randomness. We are now ready to state the second and last lemma of the chapter. In the statement of the lemma we use our convention that the dependence of the various quantities involved on the policy π and the environment ν is suppressed.

LEMMA 4.2 (Regret Decomposition Lemma) *For any policy π and K-armed stochastic bandit environment ν and horizon $n \in \mathbb{N}$, the regret R_n of policy π in ν satisfies*

$$R_n = \sum_{i=1}^{K} \Delta_i \mathbb{E}\left[T_i(n)\right] . \tag{4.5}$$

The lemma decomposes the regret in terms of the loss due to using each of the arms. It is useful because it tells us that to keep the regret small, the learner should try to minimize the weighted sum of expected action-counts, where the weights are the respective action gaps.

Lemma 4.2 tells us that a learner should aim to use an arm with a larger action gap proportionally fewer times.

Proof of Lemma 4.2 Since R_n is based on summing over rounds, and the right-hand side of the lemma statement is based on summing over actions, to convert one sum into the other one we introduce indicators. In particular, note that for any fixed t we have $\sum_k \mathbb{I}\{A_t = k\} = 1$. Hence, $S_n = \sum_t X_t = \sum_t \sum_k X_t \mathbb{I}\{A_t = k\}$

and thus

$$R_n = n\mu^* - \mathbb{E}[S_n] = \sum_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = k\}]. \tag{4.6}$$

The expected reward in round t conditioned on A_t is μ_{A_t}, which means that

$$\begin{aligned}
\mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = k\} \mid A_t] &= \mathbb{I}\{A_t = k\}\,\mathbb{E}[\mu^* - X_t \mid A_t] \\
&= \mathbb{I}\{A_t = k\}\,(\mu^* - \mu_{A_t}) \\
&= \mathbb{I}\{A_t = k\}\,(\mu^* - \mu_k) \\
&= \mathbb{I}\{A_t = k\}\,\Delta_k\,.
\end{aligned}$$

The result is completed by plugging this into Eq. (4.6) and using the definition of $T_k(n)$. □

4.4 The canonical bandit model (†)

In most cases the underlying probability space that supports the random rewards and actions is never mentioned. Occasionally, however, it becomes convenient to choose a specific probability space, which we call the **canonical bandit model**.

Finite horizon

Let $n \in \mathbb{N}$ be the horizon. A policy and bandit interact to produce the outcome, which is the tuple of random variables $H_n = (A_1, X_1, \dots, A_n, X_n)$. The first step towards constructing a probability space that carries these random variables is to choose the measurable space. For each $t \in [n]$ let $\Omega_t = ([K] \times \mathbb{R})^t \subset \mathbb{R}^{2t}$ and $\mathcal{F}_t = \mathfrak{B}(\mathbb{R}^{2t})_{|\Omega_t}$ be the restriction of the Borel σ-algebra to Ω_t (see Exercise 2.4). The random variables $A_1, X_1, \dots, A_n, X_n$ that make up the outcome are defined by their coordinate projections:

$$A_t(a_1, x_1, \dots, a_n, x_n) = a_t \qquad \text{and} \qquad X_t(a_1, x_1, \dots, a_n, x_n) = x_t\,.$$

The probability measure on $(\Omega_n, \mathcal{F}_n)$ depends on both the environment and the policy. Our informal definition of a policy is not quite sufficient now.

DEFINITION 4.1 A policy π is a sequence $\pi_1, \dots, \pi_n$ where π_t is a Markov kernel from $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to $([K], \rho)$ where ρ is the counting measure. Since the latter space is discrete we adopt the notational convention that for $i \in [K]$,

$$\pi_t(i \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) = \pi_t(\{i\} \mid a_1, x_1, \dots, a_{t-1}, x_{t-1})\,.$$

Let $\nu = (P_i)_{i=1}^K$ be a stochastic bandit where each P_i is a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. We want to define a measure on $(\Omega_n, \mathcal{F}_n)$ that respects our understanding of the sequential nature of the interaction between the learner and a stationary stochastic bandit. Since we only care about the law of the random variables (X_t)

and (A_t) the easiest way to enforce this is to directly list our expectations, which are:

(a) The conditional distribution of action A_t given $A_1, X_1, \ldots, A_{t-1}, X_{t-1}$ is $\pi_t(\cdot \mid A_1, X_1, \ldots, A_{t-1}, X_{t-1})$ almost surely.
(b) The conditional distribution of reward X_t given $A_1, X_1, \ldots, A_t$ is P_{A_t} almost surely.

The sufficiency of these assumptions is asserted by the following proposition, which we ask you to prove in Exercise 4.1.

PROPOSITION 4.1 *Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are measures on an arbitrary measurable space $(\Omega, \mathcal{F})$ and $A_1, X_1, \ldots, A_n, X_n$ are random variables on Ω. If both $\mathbb{P}$ and $\mathbb{Q}$ satisfy (a) and (b), then the law of the outcome under $\mathbb{P}$ is the same as under $\mathbb{Q}$:*

$$\mathbb{P}_{A_1, X_1, \ldots, A_n, X_n} = \mathbb{Q}_{A_1, X_1, \ldots, A_n, X_n} .$$

Next we construct a measure on $(\Omega_n, \mathcal{F}_n)$ that satisfies (a) and (b). To emphasize that what follows is intuitively not complicated, imagine that $X_t \in \{0, 1\}$ is Bernoulli, which means the set of possible outcomes is finite and we can define the measure in terms of a distribution. Let $p_i(0) = P_i(\{0\})$ and $p_i(1) = 1 - p_i(0)$ and define

$$p_{\nu\pi}(a_1, x_1, \ldots, a_n, x_n) = \prod_{t=1}^{n} \pi(a_t \mid a_1, x_1, \ldots, a_{t-1}, x_{t-1}) p_{a_t}(x_t) .$$

The reader can check that $p_{\nu\pi}$ is a distribution on $([K] \times \{0, 1\})^n$ and that the associated measure satisfies (a) and (b) above. Making this argument rigorous when (P_i) are not discrete requires the use of Radon-Nikodym derivatives. Let λ be a σ-finite measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ for which P_i is absolutely continuous with respect to λ for all i. Next let $p_i = dP_i/d\lambda$ be the Radon-Nikodym derivative of P_i with respect to λ, which is a function $p_i : \mathbb{R} \to \mathbb{R}$ such that $\int_B p_i d\lambda = P_i(B)$ for all $B \in \mathfrak{B}(\mathbb{R})$. The density $p_{\nu\pi} : \Omega \to \mathbb{R}$ can now be defined with respect to the product measure $(\rho \times \lambda)^n$ by

$$p_{\nu\pi}(a_1, x_1, \ldots, a_n, x_n) = \prod_{t=1}^{n} \pi(a_t \mid a_1, x_1, \ldots, a_{t-1}, x_{t-1}) p_{a_t}(x_t) . \tag{4.7}$$

The reader can again check (more abstractly) that (a) and (b) are satisfied by the measure $\mathbb{P}_{\nu\pi}$ defined by

$$\mathbb{P}_{\nu\pi}(B) = \int_B p_{\nu\pi}(\omega)(\rho \times \lambda)^n(d\omega) \qquad \text{for all } B \in \mathcal{F}_n .$$

It is important to emphasize that this choice of $(\Omega_n, \mathcal{F}_n, \mathbb{P}_{\nu\pi})$ is not unique. Instead, all that this shows is that a suitable probability space does exist. Furthermore, if some quantity of interest depends on the law of H_n, by Proposition 4.1, there is no loss in generality in choosing $(\Omega_n, \mathcal{F}_n, \mathbb{P}_{\nu\pi})$ as the probability space.

A choice of λ such that $P_i \ll \lambda$ for all i always exists since $\lambda = \sum_{i=1}^{K} P_i$ satisfies this condition. For direct calculations another choice is usually more convenient. For example, the counting measure when (P_i) are discrete and the Lebesgue measure for continuous (P_i).

There is another way define the probability space, which can be useful. Define a collection of independent random variables $(X_{si})_{s\in[n],i\in[K]}$ such that the law of X_{ti} is P_i. By Theorem 2.2 these random variables may be defined on $(\Omega, \mathcal{F})$ where $\Omega = \mathbb{R}^{nK}$ and $\mathcal{F} = \mathfrak{B}(\mathbb{R}^{nK})$. Then let $X_t = X_{tA_t}$ and A_t is a probability kernel from $(\Omega, \mathcal{F}_t)$ to $([K], \rho)$ where $\mathcal{F}_t = \sigma(A_1, X_1, \ldots, A_t, X_t)$. Yet another way is to define $(X_{si})_{s,i}$ as above, but let $X_t = X_{T_{A_t}(t),A_t}$. This corresponds to sampling a stack of rewards for each arm at the beginning of the game. Each time the learner chooses an action they receive the reward on top of the stack. All of these models are convenient from time to time. The important thing is that it does not matter which model we choose because the quantity of ultimate interest (usually the regret) only depends on the law of $A_1, X_1, \ldots, A_n, X_n$ and this is the same for all choices.

Infinite horizon

We never need the canonical bandit model for the case that $n = \infty$. It is comforting to know, however, that there does exist a probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\nu\pi})$ and infinite sequences of random variables $X_1, X_2, \ldots$ and $A_1, A_2, \ldots$ satisfying (a) and (b). The result follows directly from the theorem of Ionescu Tulcea (Theorem 3.3).

4.5 Notes

1 It is not obvious why the expected value is a good summary of the reward distribution. Decision makers who base their decisions on expected values are called risk-neutral. In the example shown on the figure above, a risk-averse decision maker may actually prefer the distribution labeled as A because occasionally distribution B may incur a very small (even negative) reward. Risk-seeking decision makers, if they exist at all, would prefer distributions with occasional large rewards to distributions that give mediocre rewards only. There is a formal theory of what makes a decision maker rational (a decision maker in a nutshell is rational if he/she does not contradict himself/herself). Rational decision makers compare stochastic alternatives based on the alternatives' expected utilities, according to the Von-Neumann-Morgenstern utility theorem. Humans are known to be not doing this, i.e., they are irrational. No surprise here.

2 Recall that $A \times B$ stands for the Cartesian product of sets A and B. Formally, any of the unstructured environments shown in Table 4.1 are of the form $\mathcal{E}^K = (\mathcal{P})^K$ where $(\mathcal{P})^K = \mathcal{P} \times \mathcal{P} \times \cdots \times \mathcal{P}$ and $\mathcal{P}$ is some set of distributions

over the reals. The upper index in $\mathcal{E}^K$ hints on this and also reminds us that there are K arms. Because of the product form, unstructured environments can also be called 'product environments', or 'rectangle environments'.

3 Note that every unstructured environment $\mathcal{E}^K$ is symmetric in the following sense: for any $(P_i)_i \in \mathcal{E}^K$ and any bijection $\pi : [K] \to [K]$, $(P_{\pi(i)})_i \in \mathcal{E}^K$ also holds. Since a bijection over $[K]$ is also known as a **permutation**, we also say that $\mathcal{E}^K$ is invariant to permutations. While an unstructured environment is necessarily symmetric, a symmetric environment can be structured. Consider for example $K = 2$ and consider the symmetric environment $\mathcal{E} = \{(\mathcal{B}(0), \mathcal{B}(1)), (\mathcal{B}(1), \mathcal{B}(0))\}$. Clearly, $\mathcal{E}$ is symmetric. It is not unstructured, however. If a nonzero reward is observed for the first arm, then the mean of the second arm must be zero.

4 While the canonical model introduced in Section 4.4 is enough for finite-armed bandits, in later chapters require similar constructions for more complicated settings. For example, the action space may be infinite or the learner may receive side information that evolves according to a sequence of Markov kernels. In all cases one could construct a canonical model using the same techniques, a task that we leave for connoisseurs of measure theory to tackle for themselves.

5 The study of utility and risk has a long history, going right back to (at least) the beginning of probability [Bernoulli, 1954, translated from original Latin, 1738]. The research can broadly be categorized into two branches. The first deals with *describing* how people actually make choices (**descriptive theories**) while the second is devoted to characterizing how a rational decision maker *should* make decisions (**prescriptive theories**). A notable example of the former type is 'prospect theory' [Kahneman and Tversky, 1979], which models how people handle probabilities (especially small ones) and earned Daniel Khaneman a Nobel prize (after the death of his long-time collaborator, Amos Tversky). Further descriptive theories concerned with alternative aspects of human decision-making include bounded rationality, choice strategies, recognition-primed decision making, and image theory [Adelman, 2013].

6 The most famous example of a prescriptive theory is the von Neumann-Morgenstern expected utility theorem, which states that under (reasonable) axioms of rational behavior under uncertainty, a rational decision maker must choose amongst alternatives by computing the expected utility of the outcomes [Neumann and Morgenstern, 1944]. Thus, rational decision makers, under the chosen axioms, differ only in terms of how they assign utility to outcomes (that is, rewards). Finance is another field where attitudes toward uncertainty and risk are important. Markowitz [1952] argues against expected return as a reasonable metric that investors would use. His argument is based on the (simple) observation that portfolios maximizing expected returns will tend to have a single stock only (unless there are multiple stocks with equal expected returns, a rather unlikely outcome). He argues that such a complete lack of diversification is unreasonable. He then proposes that investors should minimize the variance of the portfolio's return subject to a constraint on the

portfolio's expected return, leading to the so-called **mean-variance optimal portfolio choice theory**. Under this criteria, portfolios will indeed tend to be diversified (and in a meaningful way: correlations between returns are taken into account). This theory eventually won him a Nobel-prize in economics (shared with 2 others). Closely related to the mean-variance criterion are the 'Value-at-Risk' (VaR) and the 'Conditional Value-at-Risk', the latter of which has been introduced and promoted by Rockafellar and Uryasev [2000] due to its superior optimization properties. The distinction between the prescriptive and descriptive theories is important: Human decision makers are in many ways violating rules of rationality in their attitudes towards risk.

7 We defined the regret as an expectation, which makes it unusable in conjunction with measures of risk because the randomness has been eliminated by the expectation. When using a risk measure in a bandit setting we can either base this on the **random regret** or **pseudo-regret** defined by:

$$\hat{R}_n = n\mu^* - \sum_{t=1}^{n} X_t \,. \qquad \text{(random regret)}$$

$$\bar{R}_n = n\mu^* - \sum_{t=1}^{n} \mu_{A_t} \,. \qquad \text{(pseudo-regret)}$$

While $\hat{R}_n$ is influenced by the noise $X_t - \mu_{A_t}$ in the rewards, the pseudo-regret filters this out, which arguably makes it a better basis for measuring the 'skill' of a bandit policy. As these random regret measures tend to be highly skewed, using variance to assess risk suffers not only from the problem of penalizing upside risk, but also from failing to capture the skew of the distribution.

8 What happens if the distributions of the arms are changing with time? Such bandits are unimaginatively called **nonstationary** bandits. With no assumptions there is not much to be done. Because of this it is usual to assume the distributions change slowly. We'll eventually see that techniques for stationary bandits can be adapted quite easily to this setup (see Chapter 31).

4.6 Bibliographical remarks

There is now a huge literature on stochastic bandits, much of which we will discuss in detail in the chapters that follow. The earliest reference to the problem that we know of is by Thompson [1933], who proposed an algorithm that forms the basis of many of the currently practical approaches in use today. Thompson was a pathologist who published broadly and apparently did not pursue bandits much further. Sadly his approach was not widely circulated and the algorithm (now called Thompson sampling) did not become popular until very recently. Two decades after Thompson, the bandit problem was formally restated in a short but influential paper by Robbins [1952], an American statistician now most famous for his work on empirical Bayes. Robbin's introduced the notion of regret

and minimax regret in his 1952 paper. The regret decomposition (Lemma 4.2) has been used in practically every work on stochastic bandits and its origin is hard to pinpoint. All we can say for sure is that it does *not* appear in the paper by Robbins [1952], but does appear in the work of Lai and Robbins [1985]. Denardo et al. [2007] considers risk in a (complicated) Bayesian setting. Sani et al. [2012] consider a mean-variance approach to risk, while Maillard [2013] considers so-called coherence risk measures (CVaR, is one example of such a risk measure), and with an approach where the regret itself is redefined. Value-at-Risk is considered in the context of a specific bandit policy family by Audibert et al. [2007, 2009].

4.7 Exercises

4.1 Prove Proposition 4.1.

4.2 Prove that the measure defined in terms of the density in Eq. (4.7) satisfies the conditions (a) and (b) in Section 4.4.

Use the properties of the Radon-Nikodym derivative in combination with Fubini's theorem.

4.3 Implement a Bernoulli bandit environment in Python using the code snippet below (or adapt to your favorite language).

```
class BernoulliBandit:
  # accepts a list of K >= 2 floats, each lying in [0,1]
  def __init__(self, means):
    pass

  # Function should return the number of arms
  def K(self):
    pass

  # Accepts a parameter 0 <= a <= K-1 and returns the
  # realisation of random variable X with P(X = 1) being
  # the mean of the (a+1)th arm.
  def pull(self, a):
    pass

  # Returns the regret incurred so far.
  def regret(self):
    pass
```

4.4 Implement the following simple algorithm called 'Follow-the-Leader', which chooses each action once and subsequently chooses the action with the largest average observed so far. Ties should be broken randomly.

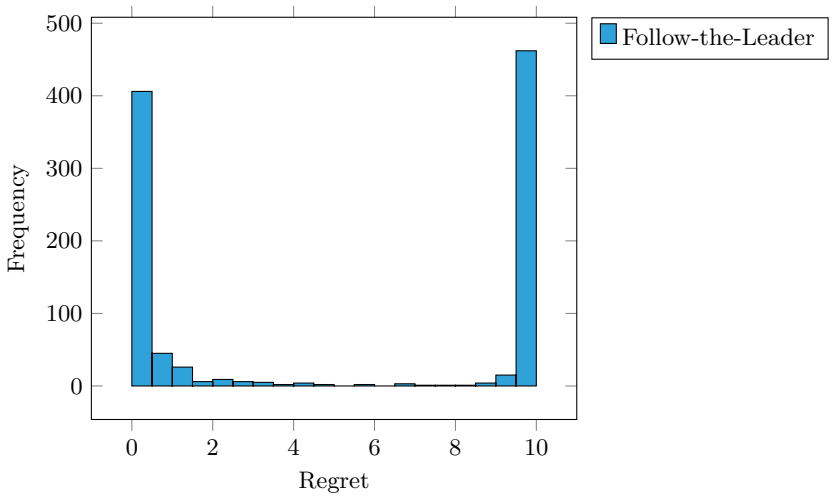


Figure 4.2 Histogram of regret for Follow-the-Leader over 1000 trials on Bernoulli bandit with means $\mu_1 = 0.5, \mu_2 = 0.6$

```
def FollowTheLeader(bandit, n):
  # implement the Follow-the-Leader algorithm by replacing
  # the code below that just plays the first arm in every round
  for t in range(n):
    bandit.pull(0)
```

Depending on the literature you are reading, Follow-the-Leader may be called 'stay with the winner' or the 'greedy algorithm'.

4.5 Consider a Bernoulli bandit with two arms and means $\mu_1 = 0.5$ and $\mu_2 = 0.6$.

(a) Using a horizon of $n = 100$, run 1000 simulations of your implementation of Follow-the-Leader on the Bernoulli bandit above and record the (random) regret, $n\mu^* - S_n$, in each simulation.
(b) Plot the results using a histogram. Your figure should resemble Fig. 4.2.
(c) Explain the results in the figure.

4.6 Consider the same Bernoulli bandit as used in the previous question.

(a) Run 1000 simulations of your implementation of Follow-the-Leader for each horizon $n \in \{100, 200, 300, \ldots, 1000\}$.
(b) Plot the average regret obtained as a function of n (see Fig. 4.3). Because the average regret is an estimator of the expected regret, you should generally include error bars to indicate the uncertainty in the estimation.

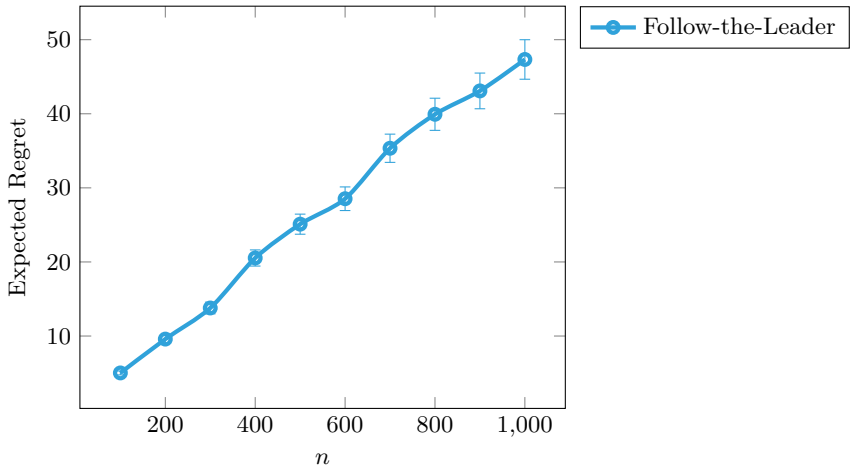


Figure 4.3 Histogram of regret for Follow-the-Leader over 1000 trials on a Bernoulli bandit with means $\mu_1 = 0.5, \mu_2 = 0.6$

(c) Explain the plot. Do you think Follow-the-Leader is a good algorithm? Why/why not?

4.7 Prove Lemma 4.1.

All items follow from Lemma 4.2.

4.8 Suppose ν is a finite-armed stochastic bandit and π is a policy such that

$$\lim_{n\to\infty} \frac{R_n(\pi,\nu)}{n} = 0\,.$$

Let $T^*(n) = \sum_{t=1}^{n} \mathbb{I}\{\mu_{A_t} = \mu^*\}$ be the number of times the optimal arm is chosen. Prove or disprove each of the following statements:

(a) $\lim_{n\to\infty} \mathbb{E}[T^*(n)]/n = 1$.
(b) $\lim_{n\to\infty} \mathbb{P}(\mu^* - \mu_{A_t} > 0) = 0$.

4.9 [One-armed bandits] This exercise is concerned with a very simple model called the **one-armed bandit**. A bar contains a single slot machine. Playing costs \$1 and the payoff is either \$2 or \$0 with probabilities p and $1-p$ respectively. Of course you do not know p and – unlike in the real world – we will assume it could reasonably take on any value in $[0, 1]$. The game proceeds over n rounds, where in each round you choose either to play the machine or do nothing. If you do nothing, then your reward is $X_t = 0$. If you play the machine, then your reward is $X_t = 1$ with probability p and $X_t = -1$ otherwise. A policy in this case chooses either PLAY or DONOTHING based on the history. The expected regret

of policy π is given by

$$R_n(p,\pi) = n \max\{0, 2p-1\} - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right],$$

where $X_1, \ldots, X_n$ are the random rewards earned by π.

(a) Describe an optimal policy when p is known (your policy should depend on p).

(b) A policy is called a **retirement policy** if it chooses to play the machine until some (possibly random) time and then does nothing until the game ends. Prove that if n is known, then for any policy π there exists a retirement policy π' such that

$$R_n(p,\pi') \le R_p^{\pi}(n) \text{ for all } p\,.$$

(c) Prove that if n is not known, then all retirement policies have linear regret for some $p \in [0,1]$ as n tends to infinity.

For (b) specify what the policy π' does given that it has access to the policy π. One easy way of doing this is assuming that π has a memory of past observations it has two subroutines; one for getting the next action and one for feeding π with the next observation. Show in a pseudocode how π' would use π through these two subroutines and argue that π' is indeed a retirement policy. For (c) use that in a stochastic bandit problem the regret can be written as $R_n = \sum_{i:\Delta_i>0} \Delta_i \mathbb{E}\left[T_i(n)\right]$ where Δ_i are the action gaps and $T_i(n)$ is the number of times arm i is chosen.