34 Bayesian Bandits

In a Bayesian bandit problem, before learning starts, a bandit environment is chosen at random from a distribution. After this, the interaction between the learner and the environment continues as before. The goal is to design a computationally efficient strategy that keeps the expected regret small, where now the expectation is also taken over the random choice of the environment. To distinguish this regret from the previous one, we call it the Bayesian regret.

If the environment-generating distribution is unknown, designing a strategy to minimize the Bayesian regret is not much different than the problem studied before: One can study either the worst-case, or the instance dependent regret, or both, where now the instances are distributions over environments. We will not consider this problem here.

The nature of the problem changes when the environment-generating distribution is given as input. In this case, the problem loses its statistical character and becomes purely computational: That of computing the actions that an optimal, or near-optimal strategy will take. One analog of this question in the non-Bayesian setting is the computation of a near-minimax strategy, which we tackled in previous chapters. In this chapter we will build the foundation for Bayesian bandits, and, more broadly for Bayesian learning and then show some special cases when the Bayes optimal strategy can in fact be computed with reasonable accuracy.

So far we have not justified why one should be interested in Bayesian bandits and Bayesian strategies. No justification is needed when the environment is indeed a random choice from a known distribution. A second justification is that Bayesian strategies often enjoy small regret in all the environments in the support of the environment-generating distribution. We will see an example for this in the next chapter. Thus, Bayesian methods can be useful beyond their defining application. In fact, there is a close connection between Bayesian optimal strategies and minimax regret.

In statistics or machine learning, some people distinguish between the **Bayesian and the frequentist school of thought** and there is much that can be heard about the debate between these schools. In our context, this could refer to preference towards a Bayesian treatment of bandits that we consider in this and the next chapter, as opposed to the rest of the book where minimax and instance-by-instance optimality is considered. However, this debate does not interest us greatly: We prefer to think about the merits and flaws of problem definitions and solution methods regardless of the label on them. Bayesian approaches to bandits have their strengths and weaknesses and we hope to do them a modicum of justice here.

34.1 Bayesian regret and Bayes optimality

Let $\mathcal{E}$ be a set of finite-armed stochastic bandits. Recall the regret of policy π in environment $\nu \in \mathcal{E}$ over n rounds is

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

Now let $\mathcal{G}$ be a σ-algebra over $\mathcal{E}$ so that $(\mathcal{E}, \mathcal{G})$ is a measurable space and let $\mathbb{Q}$ be a probability measure on this space called the **prior**. The **Bayesian regret** of policy π is the expected regret of π over all environments in $\mathcal{E}$ with respect to the prior $\mathbb{Q}$:

$$\mathrm{BR}_n(\pi, \mathbb{Q}) = \int R_n(\pi, \nu)\, d\mathbb{Q}(\nu) .$$

Implicit in this definition is the assumption that $\mathcal{G}$ is sufficiently rich that $R_n(\pi, \nu)$ is $\mathcal{G}$-measurable, which is true for all reasonable choices of $\mathcal{G}$ and policies π. Given a prior and policy the Bayesian regret is just a number. The Bayesian optimal value is $\mathrm{BR}_n^*(\mathbb{Q}) = \inf_\pi \mathrm{BR}_n(\pi, \mathbb{Q})$ and the optimal policy is

$$\pi^* = \mathrm{argmin}_\pi\, \mathrm{BR}_n(\pi, \mathbb{Q}) . \tag{34.1}$$

In all generality there is no gurantee that the optimal policy exists, but the positivity of the Bayesian regret ensures that for any $\varepsilon > 0$ there exists a policy π with $\mathrm{BR}_n(\pi, \mathbb{Q}) \leq \mathrm{BR}_n^*(\mathbb{Q}) + \varepsilon$.

The fact that $R_n(\pi, \nu) \geq 0$ for all ν and π means the Bayesian regret is always nonnegative. Perhaps less obviously, the Bayesian regret of the Bayesian optimal policy can be strictly greater than zero (Exercise 34.1).

The Bayesian regret of an algorithm is strictly less informative than the frequentist regret. By this we mean that a bound on $\mathrm{BR}_n(\pi, \mathbb{Q})$ cannot usually be used to obtain a meaningful bound on $R_n(\pi, \nu)$, while if $R_n(\pi, \nu) \leq f(\nu)$ for a measurable function f, then clearly $\mathrm{BR}_n(\pi, \mathbb{Q}) \leq \mathbb{E}[f(\nu)]$. This is not an argument against using a Bayesian algorithm, but rather an argument to analyze the frequentist regret of Bayesian algorithms.

34.2 Bayesian optimal regret for finite-armed bandits

In the standard finite-armed bandit model, the Bayesian optimal policy cannot be computed efficiently. Nevertheless, one can investigate the value of the Bayesian optimal regret by proving upper and lower bounds. One way to upper bound the Bayesian optimal regret is to integrate the frequentist regret bound of one of the algorithms from Part II.

For simplicity we restrict our attention to Bernoulli bandits, but the arguments generalize more broadly. Let $\mathcal{E} = \mathcal{E}_{\mathcal{B}}^K$ be the set of K-armed Bernoulli bandits. Bandits in this class are characterized by their mean vectors so we identify $\mathcal{E}$ with $[0,1]^K$ and define the prior $\mathbb{Q}$ on $([0,1]^K, \mathfrak{B}([0,1]^K))$. The Bayesian optimal regret is necessarily smaller than the minimax regret, which by Theorem 9.1 means that

$$\mathrm{BR}_n^*(\mathbb{Q}) \leq C\sqrt{Kn}\,,$$

where $C > 0$ is a universal constant. The proof of the lower bound in Exercise 15.1 shows that for each n there exists a prior $\mathbb{Q}$ for which

$$\mathrm{BR}_n^*(\mathbb{Q}) \geq c\sqrt{Kn}\,,$$

where $c > 0$ is a universal constant. In fact one can find a single prior such that this nearly holds for all n. We ask you to prove the following theorem in Exercise 34.3.

THEOREM 34.1 *For any prior $\mathbb{Q}$,*

$$\limsup_{n\to\infty} \frac{\mathrm{BR}_n^*(\mathbb{Q})}{n^{1/2}} = 0\,.$$

Furthermore, there exists a prior $\mathbb{Q}$ such that for all $\varepsilon > 0$,

$$\liminf_{n\to\infty} \frac{\mathrm{BR}_n^*(\mathbb{Q})}{n^{1/2-\varepsilon}} = \infty\,.$$

The above theorem has a worst case flavor in the sense that the upper bound holds for all priors and the lower bound is given for a specific prior. Some could say that the prior that yields the lower bound is quite unnatural because it assigns the overwhelming majority of its mass to bandits with small suboptimality gaps. For more 'typical' priors, the Bayesian optimal regret satisfies $\mathrm{BR}_n^*(\mathbb{Q}) = \Theta(\log^2(n))$. See the bibliographic remarks for pointers to the literature.

34.3 Bayesian learning, posterior distributions (†)

So far we introduced the Bayesian regret as an average of the frequentist regret and used the results from previous chapters to bound the Bayesian regret. For the rest of the chapter we immerse ourselves in the Bayesian viewpoint by analyzing two special cases where the Bayesian optimal policy can be computed with reasonable accuracy. Before getting into the details we discuss briefly the ideas underlying Bayesian learning, which gives us a chance to also discuss some measure-theoretic aspects of Bayesian learning.

Starting gently, suppose you are given a bag containing two marbles. A trustworthy source tells you the bag contains either *(a)* two white marbles (WW) or *(b)* a white marble and a black marble (WB). You are allowed to choose a marble from the bag (without looking) and observe its color, which we abbreviate by 'select white' (SW) or 'select black' (SB). The question is how to update your 'beliefs' about the contents of the bag having observed one of the marbles. The Bayesian way to tackle this problem starts by choosing a probability distribution on the space of hypotheses, which, incidentally, is also called the prior. This distribution is usually supposed to reflect your prior belief about which hypothesis is more probable. In the lack of extra knowledge, for the sake of symmetry, it seems reasonable to choose $\mathbb{P}(\text{WW}) = 1/2$ and $\mathbb{P}(\text{WB}) = 1/2$. The next step is to think about the likelihood of the possible outcomes under each hypothesis. Assuming that the marble is selected blindly (without peeking into the bag) and the marbles in the bag are well shuffled, these are

$$\mathbb{P}(\text{SW} \mid \text{WW}) = 1 \qquad \text{and} \qquad \mathbb{P}(\text{SW} \mid \text{WB}) = 1/2\,.$$

The conditioning here indicates that we are including the hypotheses as part of the probability space, which is a distinguishing feature of the Bayesian approach. With this formulation we can apply Bayes' law (Eq. 2.2) to show that

$$\begin{aligned}\mathbb{P}(\text{WW} \mid \text{SW}) &= \frac{\mathbb{P}(\text{SW} \mid \text{WW})\mathbb{P}(\text{WW})}{\mathbb{P}(\text{SW})} = \frac{\mathbb{P}(\text{SW} \mid \text{WW})\mathbb{P}(\text{WW})}{\mathbb{P}(\text{SW} \mid \text{WW})\mathbb{P}(\text{WW}) + \mathbb{P}(\text{SW} \mid \text{WB})\mathbb{P}(\text{WB})} \\ &= \frac{1 \times \frac{1}{2}}{1 \times \frac{1}{2} + \frac{1}{2} \times \frac{1}{2}} = \frac{2}{3}\,.\end{aligned}$$

Of course $\mathbb{P}(\text{WB} \mid \text{SW}) = 1 - \mathbb{P}(\text{WW} \mid \text{SW}) = 1/3$. Thus, while in the lack of observations, 'a priori', both hypotheses are equally likely, having observed a white marble, the probability that the bag originally contained two white marbles (and thus the bag has a white marble remaining in it) jumps to 2/3. An alternative calculation shows that $\mathbb{P}(\text{WW} \mid \text{SB}) = 0$, which makes sense because choosing a black marble rules out the hypothesis that the bag contains two white marbles. The conditional distribution $\mathbb{P}(\cdot \mid \text{SW})$ over the hypotheses is called the **posterior** distribution and represents the Bayesian's belief in each hypothesis after selecting a white marble.

A rigorous treatment of posterior distributions

A more sophisticated approach is necessary when the hypothesis and/or outcome are not discrete. In less mathematical texts the underlying details are often (quite reasonably) swept under the rug for the sake of clarity. Besides the desire for generality there are two reasons not to do this. First, having spent the effort developing the necessary tools in Chapter 2, it would seem a waste not to use them now. And second, the subtle issues that arise highlight some of the philosophical differences between the Bayesian and frequentist viewpoints that may worth illuminating. As we shall see, there is a real gap between these viewpoints!

Let $(\Omega, \mathcal{F})$, $(\Theta, \mathcal{G})$, $(\mathcal{X}, \mathcal{H})$ be measurable spaces $\{\mathbb{P}_\theta : \theta \in \Theta\}$ be a probability kernel from $(\Theta, \mathcal{G})$ to $(\Omega, \mathcal{F})$, and let $X : \Omega \to \mathcal{X}$ be a $\mathcal{F}/\mathcal{H}$-measurable map. **Posterior distributions** account for changes in beliefs, described by a probability distribution over the **parameter space** $(\Theta, \mathcal{G})$, given the **model** $\{\mathbb{P}_\theta : \theta \in \Theta\}$, in the face of observing a **realization** $x \in \mathcal{X}$ of the the random element X.

By the assumption that $\{\mathbb{P}_\theta\}$ is a probability kernel, we can define the joint probability measure $\mathbb{P}$ on $(\Theta \times \Omega, \mathcal{G} \otimes \mathcal{F})$ by

$$\mathbb{P}\left(\theta \in A, \omega \in B\right) = \int_A \mathbb{P}_\theta(B)\, d\mathbb{Q}(\theta)\,.$$

Suppose that (θ, ω) is sampled from the joint distribution $\mathbb{P}$ and $X(\omega) = x$ is observed. The posterior should be a measure on $(\Theta, \mathcal{G})$ that depends on the observed data. In other words, it should be a probability kernel from $(\mathcal{X}, \mathcal{H})$ to $(\Theta, \mathcal{G})$. Without much thought we might try and apply Bayes' law (Eq. 2.2) to claim that the posterior distribution having observed $X(\omega) = x$ should be a measure on $(\Theta, \mathcal{G})$ given by

$$\mathbb{Q}(A \mid X = x) = \mathbb{P}\left(\theta \in A \mid X = x\right) = \frac{\mathbb{P}\left(X = x \mid \theta \in A\right) \mathbb{P}\left(\theta \in A\right)}{\mathbb{P}\left(X = x\right)}\,, \tag{34.2}$$

where, by abusing the notation, we reused θ to denote the Θ-valued random element defined by $(\theta, \omega) \mapsto \theta$. The problem with the "definition" in (34.2) is that $\mathbb{P}\left(X = x\right)$ can have measure zero and then $\mathbb{P}\left(\theta \in A \mid X = x\right)$ is not defined. This is not an esoteric problem. When $\Theta = \Omega = \mathbb{R}$ with the usual Borel σ-algebras, $\mathbb{P}_\theta = \mathcal{N}(\theta, 1)$ is the Gaussian family, say, $\mathbb{Q} = \mathcal{N}(0, 1)$ and $X(\omega) = \omega$, then $\mathbb{P}\left(X = x\right) = 0$ for all x. Having read Chapter 2, the next attempt might be to define $\mathbb{Q}(A \mid X)$ as a $\sigma(X)$-measurable random variable defined using conditional expectations: For $A \in \mathcal{G}$,

$$\mathbb{Q}(A \mid X) = \mathbb{P}\left(\theta \in A \mid X\right) = \mathbb{E}[\, \mathbb{I}_{A \times \Omega} \mid X]\,,$$

where we abused θ again. Recall that $\mathbb{E}[\, \mathbb{I}_{A \times \Omega} \mid X]$ is a $\sigma(X)$-measurable random variable that is uniquely defined except for a set of measure zero. The nonuniqueness means that $\mathbb{Q}(A \mid X)$ is actually a version of $\mathbb{P}\left(\theta \in A \mid X\right)$. This leaves us with the question of which version should really be specified? For most applications of probability theory, the choice of conditional expectation does not matter. However, as it will be illustrated by means of an example soon, this is not true here! An issue that is even more annoying than nonuniqueness is that $\mathbb{Q}(\cdot \mid X)$, as defined above, need not be a measure and the basic theorems on the existence of conditional expectations do not guarantee that such a choice exists. Provided that $(\Theta, \mathcal{G})$ is not too big, however, one can guarantee the existence of a conditional expectation satisfying the conditions of a Markov kernel:

THEOREM 34.2 *If $(\Theta, \mathcal{G})$ is a Borel space then there exists a probability kernel*

$\mathbb{Q} : \mathcal{X} \times \mathcal{G} \to [0,1]$ *such that* $\mathbb{Q}(A \mid X) = \mathbb{P}(\theta \in A \mid X)$ $\mathbb{P}_X$ *almost surely for all* $A \in \mathcal{G}$. *Furthermore,* $\mathbb{Q}(\cdot \mid X)$ *is* $\mathbb{P}_X$ *almost surely unique.*

The notation for a probability kernel differs from the notation introduced in Chapter 3 because here we want to emphasize the fact that $\mathbb{Q}(A \mid X)$ is derived by conditioning. There may also be some confusion about the usage of $\mathbb{P}(\theta \in A \mid X) = \mathbb{E}[\mathbb{I}_{A\times\Omega} \mid X]$, which by definition is a $\sigma(X)$-measurable random variable on $\Theta \times \Omega$. Because it is $\sigma(X)$-measurable, however, by Lemma 2.1 there exists a $\mathcal{H}/\mathfrak{B}(\mathbb{R})$-measurable function $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{P}(\theta \in A \mid X) = f \circ X$ so that really $\mathbb{P}(\theta \in A \mid X)$ can be viewed as a function from $\mathcal{X}$. The theorem above shows that this function can be chosen so that $\mathbb{P}(\theta \in \cdot \mid X)$ is also a measure.

Theorem 34.2 shows the posterior exists, but does not suggest a useful way of finding it. In many practical situations the posterior can be calculated using densities. Given $\theta \in \Theta$ let p_θ be the Radon-Nikodym derivative of $\mathbb{P}_\theta$ with respect to some measure μ and let $q(\theta)$ be the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to another measure ν. Provided all terms are appropriately measurable and nonzero, then

$$q(\theta \mid X) = \frac{p_\theta(X) q(\theta)}{\int_\Theta p_\theta(X) q(\theta) d\nu(\theta)}$$

is the Radon-Nikodym derivative of $\mathbb{Q}(\cdot \mid X)$ with respect to ν. In other words, for any $A \in \mathcal{G}$ it holds that $\mathbb{Q}(A \mid X) = \int_A q(\theta \mid X) d\nu(\theta)$. This corresponds to the usual manipulation of densities when μ and ν are the Lebesgue measures.

The next example serves as an illustration of the issues related to the nonuniqueness of the posterior:

EXAMPLE 34.1 Let $\Theta = [0,1]$ and $\mathbb{Q}$ be the uniform measure on Θ and $\mathbb{P}_\theta = \delta_\theta$ be the Dirac measure on $[0,1]$ at θ and let $X : [0,1] \to [0,1]$ be the identity ($X(x) = x$). The following posterior satisfies the conditions of Theorem 34.2 for any countable set $C \subset [0,1]$ and probability measure μ on $([0,1], \mathfrak{B}(\mathbb{R}))$:

$$\mathbb{Q}(A \mid X = x) = \begin{cases} \delta_x(A)\,, & \text{if } x \notin C\,; \\ \mu(A)\,, & \text{if } x \in C\,. \end{cases}$$

A true Bayesian is probably unconcerned. If θ is sampled from the prior $\mathbb{Q}$, then the event $\{X \in C\}$ has measure zero and there is little cause to worry about events that happen with probability zero. But for a frequentist using Bayesian techniques for inference this actually matters. If θ is not sampled from $\mathbb{Q}$, then nothing prevents the situation that $\theta \in C$ and the nonuniqueness of the posterior is an issue. Probability theory does not provide a way around this issue.

⚠ When using Bayesian techniques for inference in a frequentist setting one should be careful to specify the version of the posterior being used. This is important because in the frequentist viewpoint θ is not part of the probability space and results are proven for $\mathbb{P}_\theta$. By contrast, the all-in Bayesian includes θ in the probability space and need not worry about events with negligible prior probability.

34.4 Conjugate priors and the exponential family (†)

One of the strengths of Bayesian approach is the ability to explicitly specify and incorporate prior beliefs into the uncertainty models in a natural way via the prior. When it comes to Bayesian algorithms, this advantage is belied a little by the competing necessity of choosing a prior for which the posterior can be efficiently computed. The ease of computing the posterior depends on the interplay between the prior and the model. Given the importance of computation, it is hardly surprising that researchers have worked hard to find models and priors that behave well together. A prior and model are called **conjugate** if the posterior has the same parametric form as the prior.

Gaussian model/Gaussian prior
Suppose that $(\Theta, \mathcal{G}) = (\Omega, \mathcal{F}) = (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ and $X : \Omega \to \Omega$ is the identity and $\mathbb{P}_\theta$ is Gaussian with mean θ and known **signal variance** σ_S^2. If the prior $\mathbb{Q}$ is Gaussian with mean μ_P and **prior variance** σ_P^2, then the posterior distribution having observed $X = x$ is

$$\mathbb{Q}(\cdot \mid X = x) = \mathcal{N}\left(\frac{\mu_P/\sigma_P^2 + x/\sigma_S^2}{1/\sigma_P^2 + 1/\sigma_S^2}, \left(\frac{1}{\sigma_S^2} + \frac{1}{\sigma_P^2}\right)^{-1}\right).$$

We leave the proof of this fact to the reader (Exercise 34.2). The limiting regimes as the prior/signal variance tend to zero or infinity are quite illuminating. For example, as $\sigma_P^2 \to 0$ the posterior tends to a Gaussian $\mathcal{N}(\mu_P, \sigma_P^2)$, which is equal to the prior and indicates that no learning occurs. This is consistent with intuition: If the prior variance is zero, then the statistician is already certain of the mean and no amount of data can change their belief. On the other hand, as σ_P^2 tends to infinity we see the mean of the posterior has no dependence on the prior mean, which means that all prior knowledge is washed away with just one sample. We encourage you to examine what happens when $\sigma_S^2 \to \{0, \infty\}$.

Notice how the model has fixed σ_S^2, suggesting that the model variance is known. We made this kind of assumption very often in the book so far, but the Bayesian can incorporate their uncertainty over the variance. In this case the model parameters are $\Theta = \mathbb{R} \times [0, \infty)$ and $\mathbb{P}_\Theta = \mathcal{N}(\theta_1, \theta_2)$. But is there a conjugate prior in this case? Already things are getting complicated, so we will simply let you know that the family of Gaussian-inverse-gamma distributions is conjugate.

Bernoulli model/beta prior
Suppose that $\Theta = [0,1]$ and $\mathbb{P}_\theta = \mathcal{B}(\theta)$ is Bernoulli with parameter θ. In this case it turns out that the family of beta distributions is conjugate, which for parameters $\theta = (\alpha, \beta) \in (0, \infty)^2$ is given in terms of its probability density function with respect to the Lebesgue measure:

$$p_{\alpha,\beta}(x) = x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,, \tag{34.3}$$

where $\Gamma(x)$ is the Gamma function. Then the posterior having observed $X \in \{0,1\}$ is also a beta distribution with parameters $(\alpha + X, \beta + 1 - X)$.

Exponential families
Both the Gaussian and Bernoulli families are examples of a more general concept. Let μ be a measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ and $T, \eta : \mathbb{R} \to \mathbb{R}$ where T is called the **sufficient statistic** and define measure $\mathbb{P}_\theta$ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ in terms of its Radon-Nikodym derivatives with respect to μ:

$$\frac{d\mathbb{P}_\theta}{d\mu}(x) = \exp\left(\eta(\theta)T(x) - A(\theta)\right)\,,$$

where $A(\theta) = \log \int_{\mathbb{R}} \exp(\eta(\theta)T(x))d\mu(x)$ is the **log-partition** function. To expand the definition of P_θ, let $\Theta = \operatorname{dom}(A) = \{\theta : A(\theta) < \infty\}$ be the domain of A. Then, for $\theta \in \operatorname{dom}(A)$, P_θ is the measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ defined by

$$P_\theta(B) = \int_B \frac{dP_\theta}{d\mu}(x)d\mu(x)\,.$$

The collection $\{P_\theta : \theta \in \Theta\}$ is called a **single parameter exponential family**.

EXAMPLE 34.2 Let $\sigma^2 > 0$ and $\mu = \mathcal{N}(0, \sigma^2)$ and $\eta(\theta) = \frac{\theta}{\sigma}$ and $T(x) = \frac{x}{\sigma}$. An easy calculation shows that $A(\theta) = \theta^2/(2\sigma^2)$, which has domain $\Theta = \mathbb{R}$ and $P_\theta = \mathcal{N}(\theta, \sigma^2)$.

EXAMPLE 34.3 Let $\mu = \delta_0 + \delta_1$ be the sum of Dirac measures and $T(x) = x$ and $\eta(\theta) = \theta$. Then $A(\theta) = \log(1 + \exp(\theta))$ and $\Theta = \mathbb{R}$ and $P_\theta = \mathcal{B}(\sigma(\theta))$ where $\sigma(\theta) = \exp(\theta)/(1 + \exp(\theta))$ is the sigmoid function.

EXAMPLE 34.4 The same family can be parameterized in many different ways. Let $\mu = \delta_0 + \delta_1$ and $T(x) = x$ and $\eta(\theta) = \log(\theta/(1-\theta))$. Then $A(\theta) = -\log(1-\theta)$ and $\Theta = (0,1)$ and $P_\theta = \mathcal{B}(\theta)$.

Exponential families have many nice properties. Of most interest to us here is the existence of conjugate priors. Suppose that $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is a single parameter exponential family determined by functions η and T with $T(x) = x$ assumed to be the identity. Let $x_0, n_0 \in \mathbb{R}$ and define prior measure $\mathbb{Q}$ on $(\Theta, \mathfrak{B}(\Theta))$ in terms of its density $q = d\mathbb{Q}/d\lambda$ with λ the Lebesgue measure:

$$q(\theta) = \frac{\exp\left(n_0 x_0 \eta(\theta) - n_0 A(\theta)\right)}{\int_\Theta \exp\left(n_0 x_0 \eta(\theta) - n_0 A(\theta)\right) d\theta}\,, \tag{34.4}$$

where we assume the existence and strict positivity of the integral in the denominator. Suppose we observe $X = x$. Then the posterior has density with respect to the Lebesgue measure given by

$$q(\theta \mid x) = \frac{\exp\left(\eta(\theta)(x + n_0 x_0) - (1+n_0)A(\theta)\right)}{\int_\Theta \exp\left(\eta(\theta)(x + n_0 x_0) - (1+n_0)A(\theta)\right) d\lambda(\theta)} \,.$$

What this means is that after observing the value x, the posterior takes the form of the prior except that the parameters (x_0, n_0) associated with the prior get updated to $((n_0 x_0 + x)/(n_0+1), n_0+1)$: The posterior is both easy to represent and easy to maintain. To see how exponential families recovers previous examples consider the Bernoulli case of Example 34.4. Since

$$\exp(n_0 x_0 \eta(\theta) - n_0 A(\theta)) = \left(\frac{\theta}{1-\theta}\right)^{n_0 x_0} (1-\theta)^{n_0} = \theta^{n_0 x_0}(1-\theta)^{n_0(1-x_0)} \,,$$

we see that the prior from (34.4) is a beta distribution with parameters $\alpha = 1 + n_0 x_0$ and $\beta = 1 + n_0(1 - x_0)$ as can be seen from (34.3). As expected, the posterior update also works as described earlier.

There are important parametric families with conjugate priors that are not exponential families. One example is the uniform family $\{\mathcal{U}(a,b) : a < b\}$, which is conjugate to the Pareto family.

34.5 Posterior distributions in bandits

Adapting the tools of the previous two sections to bandits is straightforward. Let $\mathcal{E}$ be a set of K-armed stochastic bandits and $\mathcal{G}$ be a σ-algebra on $\mathcal{E}$ and $\mathbb{Q}$ be a prior probability measure on $(\mathcal{E}, \mathcal{G})$. Given a bandit $\nu \in \mathcal{E}$ let $\mathbb{P}_\nu$ be the product measure on $(\mathbb{R}^K, \mathfrak{B}(\mathbb{R}^K))$ that corresponds to the reward distributions. Let $\bar{\mathbb{P}} = (\mathbb{P}_\nu)_\nu$ which is assumed to be a Markov kernel from $(\mathcal{E}, \mathcal{G})$ to $(\mathbb{R}^K, \mathfrak{B}(\mathbb{R}^K))$. Fix a policy $\pi = (\pi_1, \dots, \pi_n)$ and let

$$\Omega = \{(a_1, x_1, \dots, a_n, x_n) \ : a_t \in [K] \text{ and } x_t \in \mathbb{R}^K\} \qquad \text{and} \qquad \mathcal{F} = (2^{[K]} \otimes \mathfrak{B}(\mathbb{R}))^n \,. \tag{34.5}$$

Then define joint probability space $(\mathcal{E} \times \Omega, \mathcal{G} \otimes \mathcal{F}, \mathbb{P})$, where

$$\begin{aligned} d\mathbb{P}(\nu, a_1, x_1, \dots, a_n, x_n) = d\mathbb{Q}(\nu) d\pi_1(a_1) dP_\nu(x_1|a_1) d\pi_2(a_2|a_1, x_1) \\ \dots\ d\pi_n(a_n|a_1, x_1, \dots, a_{n-1}, x_{n-1}) dP_\nu(x_n|a_n) \,. \end{aligned}$$

The coordinate projections are

$$\begin{aligned} \nu((\nu, a_1, x_1, \dots, a_n, x_n)) &= \nu \\ A_t((\nu, a_1, x_1, \dots, a_n, x_n)) &= a_t \qquad \text{and} \\ X_t((\nu, a_1, x_1, \dots, a_n, x_n)) &= x_t \,, \end{aligned} \tag{34.6}$$

which means that $\nu \in \mathcal{E}$, $A_t \in [K]$ and $X_t \in \mathbb{R}^K$ are random elements. Then let $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ be the σ-algebra generated by the observations of the learner after t rounds. The posterior after t rounds is a probability kernel from $(\Omega, \mathcal{F}_t)$ to $(\mathcal{E}, \mathcal{G})$ denoted by $\mathbb{Q}_t(\cdot) = \mathbb{Q}(\cdot \mid A_1, X_1, \dots, A_t, X_t)$ that for all $B \in \mathcal{G}$ satisfies

$$\mathbb{Q}_t(B) = \mathbb{E}[\,\mathbb{I}_B(\nu) \mid A_1, X_1, \dots, A_t, X_t\,] \quad \text{a.s.}$$

Theorem 34.2 guarantees the existence of the posterior as long as $(\mathcal{E}, \mathcal{G})$ is a Borel space, but the abstract definition is not very useful for explicit calculations. By making mild assumptions the posterior can be written in terms of densities. Assume there exists a σ-finite measure λ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $P \ll \lambda$ for all reward distributions P used by the bandits in $\mathcal{E}$. Recall from Chapter 15 that the Radon-Nikodym derivative of $\mathbb{P}_{\nu\pi}$ with respect to $(\rho \times \lambda)^n$ is

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi_t(a_t \mid a_1, x_1, \dots, a_{t-1} x_{t-1}) p_{\nu a_t}(x_t)\,, \tag{34.7}$$

where $p_{\nu a}$ is the density with respect to λ of the reward distribution for the ath arm of ν. Then the posterior after t rounds is given by

$$\begin{aligned}\mathbb{Q}_t(\,B \mid a_1, x_1, \dots, a_t, x_t\,) &= \frac{\int_B p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) d\mathbb{Q}(\nu)}{\int_{\mathcal{E}} p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) d\mathbb{Q}(\nu)} \\ &= \frac{\int_B \prod_{s=1}^t p_{\nu a_s}(x_s) d\mathbb{Q}(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) d\mathbb{Q}(\nu)}\,, \end{aligned} \tag{34.8}$$

where the second equality follows from Eq. (34.7). Of course, we are assuming here that all quantities are well defined. In particular, the integral in the denominator must be positive almost surely and $p_{\nu a}(x)$ should be measurable as a function of ν for all x.

EXAMPLE 34.5 Let $\mathcal{E} = \mathcal{E}_{\mathcal{B}}^K$ be the set of all Bernoulli bandits with K arms. Bandits in $\mathcal{E}$ are characterized by their mean vectors in $[0,1]^K$ so it suffices to choose our prior on $[0,1]^K$ with the Lebesgue σ-algebra. A natural prior is a product of Beta priors with parameters $\alpha, \beta > 0$ defined in terms of its density with respect to the Lebesgue measure λ by

$$q(\theta) \propto \prod_{i=1}^K \theta_i^{\alpha-1}(1-\theta_i)^{\beta-1}\,.$$

Recall that $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_t = i\}$ and let $S_i(t) = \sum_{s=1}^t \mathbb{I}\{A_t = i\} X_t$. The posterior is also given in terms of its density with respect to the Lebesgue measure.

$$q(\theta \mid A_1, X_1, \dots, A_t, X_t) \propto \prod_{i=1}^K \theta_i^{\alpha + S_i(t) - 1}(1-\theta_i)^{\beta + T_i(t) - S_i(t) - 1}\,.$$

This means the posterior is also the product of beta distributions, each updated according to the observations from the relevant arm.

34.6 One-armed bandits

We return to the one-armed bandit problem that appeared in various exercises in earlier chapters. In each round t the learner chooses action $A_t \in \{1, 2\}$. The reward when choosing the first action is X_t where $X_1, \dots, X_n$ is a sequence of independent and identically distributed random variables with unknown distribution ν and mean $\mu \in \mathbb{R}$. The reward when choosing the second action is a deterministic known value $\mu_\circ \in \mathbb{R}$. In Exercise 4.9 we defined a **retirement policy** for one-armed bandits as a policy that chooses $A_t = 1$ (experimenting) until some random time and subsequently $A_t = 2$. There you showed that provided the horizon n is known in advance, then there is no reason to consider policies of any other kind (the problem was posed just for a specific distribution ν, but it holds with full generality). So, given a fixed horizon n, the problem reduces to finding the 'best' retirement policy.

Let $\mathcal{F}_0$ be the trivial σ algebra, and let $\mathcal{F}_t = \sigma(X_1, \dots, X_t)$ so that a retirement policy is a stopping time $0 \le \tau \le n$ with respect to filtration $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$. Given ν, the regret of the policy induced by $\mathbb{F}$-stopping time τ is

$$\begin{aligned} R_n(\tau, \nu) &= n \max\{\mu, \mu_\circ\} - \mathbb{E}\left[\sum_{t=1}^{\tau} X_t + \sum_{t=\tau+1}^{n} \mu_\circ\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{\tau} \max\{0, \mu_\circ - \mu\} + \sum_{t=\tau+1}^{n} \max\{\mu - \mu_\circ, 0\}\right] . \end{aligned}$$

When necessary, we will call this the frequentist regret to distinguish it from the Bayesian regret. The regret is minimized by deterministic stopping time $\tau = n$ if $\mu > \mu_\circ$ and $\tau = 0$ otherwise. As usual the problem is that μ is unknown.

Frequentist regret and policy

If we assume that ν is 1-subgaussian, then the techniques of Part II can be applied to derive a stopping time such that

$$R_n(\tau, \nu) \le \begin{cases} \Delta\,, & \text{if } \Delta \ge 0\,; \\ \min\{\Delta + C\log(n)/\Delta,\, n\Delta\}\,, & \text{otherwise}\,, \end{cases} \tag{34.9}$$

where $C > 0$ is a universal constant and $\Delta = |\mu - \mu_\circ|$. An example stopping time for which this holds is

$$\tau = n \wedge \min\left\{t \in [n] : \hat{\mu}_t + \sqrt{\frac{2\log(n^2)}{t}} \le \mu_\circ\right\} , \tag{34.10}$$

where $\hat{\mu}_t = \frac{1}{t}\sum_{s=1}^t X_s$. We leave it to the reader to establish that Eq. (34.9) indeed holds for the retirement policy using the above stopping time (Exercise 34.9).

Bayesian regret and policy

Moving on to the Bayesian framework, we now suppose that $\nu = \nu_\theta$ where $\theta \in \Theta$ and $\{\mathbb{P}_{\nu_\theta} : \theta \in \Theta\}$ is a probability kernel from Borel space $(\Theta, \mathcal{G})$ to $(\Omega, \mathcal{F})$, where

$(\Omega, \mathcal{F})$ is the measurable space that holds the random variables $(X_1, \dots, X_n)$. Then let $\mathbb{Q}$ be a prior measure on $(\Theta, \mathcal{G})$ and $\mathbb{P}$ be the measure on $(\Theta \times \Omega, \mathcal{G} \otimes \mathcal{F})$ given by

$$\mathbb{P}(\theta \in A, \omega \in B) = \int_A \mathbb{P}_{\nu_\theta}(B) d\mathbb{Q}(\theta)\,.$$

Unless otherwise specified, expectations for the remainder of the section are with respect to $\mathbb{P}$.

⚠ In the model used in the frequentist setting the variables $X_1, X_2, \dots, X_n$ are independent and identically distributed with respect to $\mathbb{P}_\nu$. Having incorporated θ into the probability space this is not true anymore. However, up to the usual 'almost surely' exceptions they are still conditionally independent and identically distributed given θ.

The posterior after t observations is a probability kernel $\mathbb{Q}_t$ from $(\mathbb{R}^t, \mathfrak{L}(\mathbb{R}^t))$ to $(\Theta, \mathcal{G})$ such that $\mathbb{Q}_t(A) = \mathbb{E}[\mathbb{I}(\theta \in A) \mid \mathcal{F}_t]$ almost surely, where we have abused θ, as explained after Eq. (34.2). We abbreviate $\mathbb{E}_t[\cdot] = \mathbb{E}_{[}\cdot \mid \mathcal{F}_t]$. The Bayesian regret of the retirement policy determined by τ is

$$\mathrm{BR}_n(\tau, \mathbb{Q}) = \mathbb{E}\left[R_n(\tau, \nu_\theta)\right]\,.$$

The Bayesian optimal policy (if it exists) is a retirement policy (answer why in Exercise 34.4) that minimizes this quantity,

$$\tau^* \in \mathrm{argmin}_\tau\, \mathrm{BR}_n(\tau, \mathbb{Q})$$

The key idea to finding τ^* is to rewrite the optimization problem in terms of an **optimal stopping problem**. Let $U_t = \sum_{s=1}^t X_s + (n-t)\mu_\circ$ be the cumulative reward received by a learner that stops at time t (for $t = 0$, $U_0 = n\mu_\circ$). Then, it is easy to see that

$$\tau^* \in \mathrm{argmax}_\tau\, \mathbb{E}[U_\tau]\,. \tag{34.11}$$

Because U_t is $\mathcal{F}_t$-measurable, this problem is called a **standard optimal stopping problem**. For standard problems in discrete time with a finite horizon the solution can be found using **backwards induction**. Intuitively, having observed $X_1, \dots, X_t$ the optimal policy will retire if U_t is larger than the expected return from the optimal stopping policy that stops at time $t+1$ or later. This suggests defining things backwards from n by

$$V_n = U_n \qquad \text{and} \qquad V_t = \max\{U_t, \mathbb{E}_t[V_{t+1}]\} \quad \text{for } t < n\,.$$

The process $(V_t)_t$ is called the **Snell envelope** and the optimal stopping time stops at the earliest time such that $U_t \geq V_t$. This intuitive fact is captured in the following theorem, whose proof is left as a not necessarily easy exercise (Exercise 34.5).

THEOREM 34.3 *Assuming that* $\sup_\tau \mathbb{E}[|U_\tau|] < \infty$, *then* $\tau^* = \min\{t : U_t \geq V_t\}$ *satisfies Eq.* (34.11).

The optimal policy in Theorem 34.3 only depends on the ordering of U_t and V_t, which by subtracting the cumulative observed reward allows us to rewrite the optimal stopping time in a more convenient form. Define $W_n = 0$ and for $t < n$ let $W_t = V_t - \sum_{s=1}^t X_s$, which satisfies

$$\begin{aligned} W_t &= \max\left((n-t)\mu_\circ,\, \mathbb{E}_t[V_{t+1}] - \sum_{s=1}^t X_s\right) \\ &= \max\left((n-t)\mu_\circ,\, \mathbb{E}_t\left[W_{t+1}\right] + \mathbb{E}_t[X_{t+1}]\right) . \end{aligned} \tag{34.12}$$

Then the optimal policy is

$$\begin{aligned} \tau^* = \min\{t : U_t \geq V_t\} &= \min\left\{t : U_t - \sum_{s=1}^t X_s \geq V_t - \sum_{s=1}^t X_s\right\} \\ &= \min\left\{t : (n-t)\mu_\circ \geq \mathbb{E}_t\left[W_{t+1}\right] + \mathbb{E}_t[X_{t+1}]\right\} . \end{aligned}$$

Theorem 34.3 and the above display characterize the optimal stopping rule in a straightforward way. The difficulty is that $\mathbb{E}_t[W_{t+1}]$ is usually a complicated object. We now give two examples where $\mathbb{E}_t[W_{t+1}]$ has a simple representation that means computing the optimal stopping rule is practical.

Bernoulli rewards

Let $\Theta = [0, 1]$ and $\mathcal{F}$ be the standard Borel σ-algebra and $\nu_\theta = \mathcal{B}(\theta)$ be Bernoulli with bias θ. In the previous section we showed that the beta prior and Bernoulli family are conjugates so we will choose the prior to be $\mathbb{Q} = \text{Beta}(\alpha, \beta)$ for some $\alpha, \beta > 0$. A calculation shows that

$$\mathbb{E}[X_{t+1} \mid \mathcal{F}_t] = \frac{\alpha + S_t}{\alpha + \beta + t} = p_t(S_t)\,,$$

where $p_t(s) = (\alpha + s)/(\alpha + \beta + t)$. This greatly simplifies matters because W_t can be written as a function of just $S_t \in \{0, 1, \ldots, t\}$.

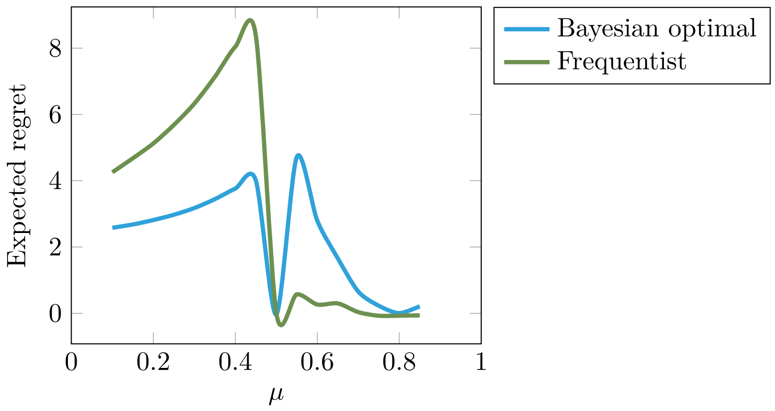
$$\begin{aligned} W_t(s) &= \max\left((n-t)\mu_\circ,\, \mathbb{E}\left[W_{t+1} \mid S_t = s\right] + \mathbb{E}_t[X_{t+1} \mid S_t = s]\right) \\ &= \max\left((n-t)\mu_\circ,\, p_t(s)W_{t+1}(s+1) + (1 - p_t(s))W_{t+1}(s) + p_t(s)\right) . \end{aligned}$$

So the optimal policy can be computed by evaluating $W_t(s)$ for all $s \in \{0, \ldots, t\}$ starting with $t = n - 1$, then $n - 2$ and so-on until $t = 0$. The total computation for this backwards induction is $O(n^2)$ and the output is a policy that can be implemented over all n rounds. In contrast, the stopping rule proposed in Eq. (34.10) requires only $O(n)$ computations, so the overhead is quite severe. The improvement is also not insignificant as illustrated by the following experiment.

The horizon is set to $n = 500$ and $\mu_\circ = 1/2$. The stopping times we compare are the Bayesian optimal policy with a Beta$(1, 1)$ prior and the 'frequentist' stopping time given by

$$\tau_{\mathrm{D}} = n \wedge \min\left\{t \geq 1 : \hat{\mu}_t < \mu_\circ \text{ and } d(\hat{\mu}_t, \mu_\circ) \geq \frac{\log(n/t)}{t}\right\}, \tag{34.13}$$

where $d(p, q) = \mathrm{D}(\mathcal{B}(p), \mathcal{B}(q))$ is the relative entropy between Bernoulli distributions with parameters p and q respectively. The plot below shows the expected regret for different values of μ. As you can see, the results are not a clear win in favour of the Bayesian optimal policy. The asymmetric behaviour of the frequentist policy is explained by the conservatism of the confidence interval in Eq. (34.13), which makes it stop consistently later than its Bayesian counterpart. In a sense this is an advantage of the Bayesian approach, where the prior encodes the objective and the policy automatically optimises the criteria. Because the Beta$(1, 1)$ prior is symmetric about $1/2$ it should not surprise us that the regret is approximately symmetric, but not completely so because the one-armed bandit problem is asymmetric. The Bayesian optimal regret for this problem is approximately $\mathrm{BR}_n^* \approx 2$ while the frequentist algorithm has Bayesian regret of $\mathrm{BR}_n \approx 4.5$.



Gaussian rewards

The Gaussian case is more delicate because W_t does not have a discrete representation. To make things concrete assume that $\nu_\theta = \mathcal{N}(\theta, 1)$ is Gaussian with unit variance and the prior $\mathbb{Q}$ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ is also Gaussian with mean $\mu_P \in \mathbb{R}$ and variance $\sigma_P^2 > 0$. By the results in Section 34.4 the posterior $\mathbb{Q}_t$ is Gaussian with mean μ_t and variance σ_t^2 given by

$$\mu_t = \frac{\frac{\mu_P}{\sigma_P^2} + \sum_{s=1}^t X_s}{1 + \sigma_P^{-2}} \qquad \text{and} \qquad \sigma_t^2 = \left(t + \frac{1}{\sigma_P^2}\right)^{-1}.$$

Notice that σ_t^2 is independent of the observations so the posterior is determined entirely by its mean. Thus we can view W_t as a function from the posterior mean $\mu_t \in \mathbb{R}$ to $\mathbb{R}$, which by Eq. (34.12) is given by

$$W_t(\mu) = \max\left((n-t)\mu_\circ,\, \mu + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma_t^2}\right) W_{t+1}(\mu + x)dx\right) .$$

In general the integral on the right-hand side does not have a closed form solution, which forces the use of approximate methods. Fortunately W_t is a well behaved function and can be efficiently approximated.

LEMMA 34.1 *The following hold:*

(a) $W_t(\mu)$ *is monotone increasing in* μ*.*
(b) $W_t(\mu)$ *is convex.*
(c) $\lim_{\mu\to\infty} W_t(\mu)/\mu = n - t$ *and* $\lim_{\mu\to-\infty} W_t(\mu) = (n-t)\mu_\circ$*.*

Proof The first two follow from the calculus of monotone/convex functions while the third we leave as an exercise to the reader. □

There are many ways to approximate a function, but the important point here is we want an approximation of W_t such that the integral in the recursive definition can be computed efficiently. Given the properties in Lemma 34.1 a natural choice is to approximate W_t using piecewise quadratic functions. Let $\tilde{W}_{n+1}(\mu) = 0$ and

$$\bar{W}_t(\mu) = \max\left\{(n-t)\mu_\circ,\, \mu + \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma_t^2}\right) \tilde{W}_{t+1}(\mu + x)dx\right\} .$$

Then let $-\infty < x_1 \le x_2 \le \dots \le x_N < \infty$ and for $\mu \in [x_i, x_{i+1}]$ define $\tilde{W}_t(\mu) = a_i\mu^2 + b_i\mu + c_i$ to be the unique quadratic approximation of $\bar{W}_t(\mu)$ such that

$$\begin{aligned}
\tilde{W}_t(x_i) &= \bar{W}_t(x_i) \\
\tilde{W}_t(x_{i+1}) &= \bar{W}_t(x_{i+1}) \\
\tilde{W}_t((x_i + x_{i+1})/2) &= \bar{W}_t((x_i + x_{i+1})/2) .
\end{aligned}$$

For $\mu < x_1$ we approximate $W_t(\mu) = (n-t)\mu_\circ$ and for $\mu > x_N$ the linear approximation $\tilde{W}_t(\mu) = (n-t)\mu$ is reasonable by Lemma 34.1. The computation time for calculating the coefficients a_i, b_i, c_i for all t and $i \in [N]$ is $O(Nn)$.

34.7 Gittins index

Generalizing the analysis in the previous section to multiple actions is mathematically straightforward, but computationally intractable. The computational complexity of backwards induction increases exponentially with the number arms, which even for two actions makes this approach quite impractical.

An **index policy** is a policy that in each round computes a real-valued **index** for each arm and plays the arm with the largest index. Furthermore, the index of an arm should only depend on statistics collected for that arm (and perhaps the time horizon). For example, most variants of the upper confidence bound algorithm introduced in Part II are index policies. Sadly, however, the Bayesian optimal policy for finite-horizon bandits is not usually an index policy. John Gittins proved that if one is prepared to modify the objective to a special kind of infinite horizon problem, then the Bayesian optimal policy becomes an index policy.

A discounted retirement game

We start by describing the discounted setting with one action and then generalize to multiple actions. Let $(\mathcal{S}, \mathcal{G})$ be a measurable space called the **state space**. Then let μ be a Markov kernel from $(\mathcal{S}, \mathcal{G})$ to itself and $(\Omega, \mathcal{F}, \mathbb{P}_s)$ be a probability space and $S_1, S_2, \dots$ be a sequence of $\mathcal{F}/\mathcal{G}$-measurable random elements such that $\mathbb{P}_s(S_1 = s) = 1$ and $\mathbb{P}_s(S_{t+1} \in \cdot \mid S_t) = \mu(S_t, \cdot)$ almost surely. Finally let $r : \mathcal{S} \to [0, 1]$ be a known measurable function and $\gamma \in \mathbb{R}$. In each round t the learner observes the state S_t and chooses one of two options: *(a)* to retire, which ends the game. Or *(b)* pay a fixed cost of γ to receive a reward of $r(S_t)$ and continue for another round. The policy of a learner in this game corresponds to choosing a stopping time τ with respect to the filtration $\mathbb{F} = (\mathcal{F}_t)_t$ with $\mathcal{F}_t = \sigma(S_1, \dots, S_t)$, where $\tau = t$ means that the learner retires after observing S_t at the start of round t. The value of a retirement policy τ is given by

$$V^\tau(s; \gamma) = \mathbb{E}_s\left[\sum_{t=1}^{\tau-1} \alpha^{t-1}(r(S_t) - \gamma)\right],$$

where $\alpha \in (0, 1)$ is the **discount factor** and $\mathbb{E}_s$ is the expectation with respect to $\mathbb{P}_s$. This definition of the value function means a learner is encouraged to obtain large rewards earlier rather than later and is one distinction between this model and the finite-horizon model studied for most of this book. A brief discussion of discounting is left for the notes.

If $\tau = 1$ almost surely, then learner retires immediately without receiving any reward or paying any cost and $V^\tau(s; \gamma) = 0$. The **Gittins index** or **fair charge** of a state s is the largest value of γ for which the learner is indifferent between retiring immediately and playing for at least one round:

$$G^*(s) = \sup\left\{\gamma \in \mathbb{R} : \sup_{\tau > 1} V^\tau(s; \gamma) \geq 0\right\}, \tag{34.14}$$

where the inner supremum is taken over $\mathbb{F}$-stopping times τ with $\tau > 1$ almost surely. Straightforward manipulation (Exercise 34.6) shows that

$$G^*(s) = \sup_{\tau > 1} \frac{\mathbb{E}_s\left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t)\right]}{\mathbb{E}_s\left[\sum_{t=1}^{\tau-1} \alpha^t\right]}, \tag{34.15}$$

It is not immediately clear that a stopping time attaining the supremum in the definition exists. The following lemma shows that it does and gives an explicit form. The proof of this result is left as a technical challenge for the reader (Exercise 34.7).

LEMMA 34.2 *For each $s \in \mathcal{S}$ the following stopping times both attain the supremum in Eq.* (34.15).

(a) $\tau = \min\{t > 1 : G^*(S_t) < G^*(s)\}$.
(b) $\tau = \min\{t > 1 : G^*(S_t) \leq G^*(s)\}$.

The result is relatively intuitive. The Gittins index represents the price the learner should be willing to pay for the privilege of continuing to play. The optimal policy continues to play as long as the actual value of the game is not smaller than this price with an indifference region when the price is exactly equal to the value.

Discounted bandits and the index theorem

The generalization of the discounted retirement game to multiple arms is quite straightforward. There are now K independent Markov chains on the same state-space and in each round the learner first observes the state of all chains and chooses an action $A_t \in [K]$. The learner receives a reward from the corresponding chain, which then evolves randomly to a new state sampled from the probability kernel. The states for unplayed arms do not change and we assume that all chains evolve according to the same Markov kernel. The protocol is given in Fig. 34.1.

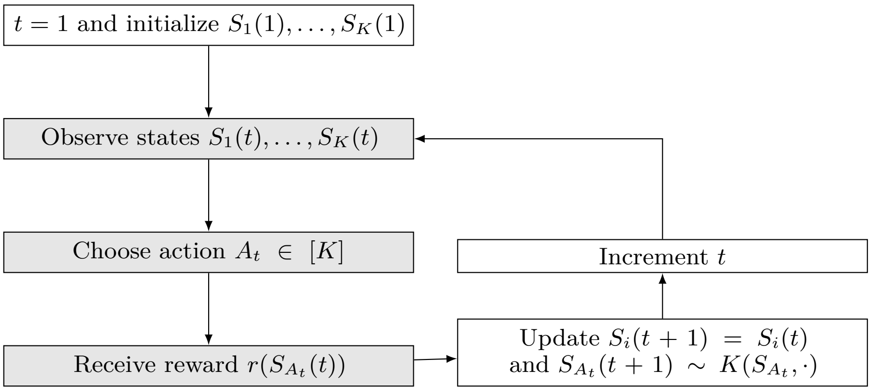

$t = 1$ and initialize $S_1(1), \dots, S_K(1)$
Observe states $S_1(t), \dots, S_K(t)$
Choose action $A_t \in [K]$
Increment t
Receive reward $r(S_{A_t}(t))$
Update $S_i(t+1) = S_i(t)$ and $S_{A_t}(t+1) \sim K(S_{A_t}, \cdot)$

Figure 34.1 Interaction protocol for discounted bandits.

The assumption that the Markov chains evolve on the same state-space with the same transition kernel is non-restrictive since the state-space can always be taken to be the union of K state-spaces and the transition kernel defined with K disconnected components.

Given a discount parameter $\alpha \in (0,1)$, the value of policy π is

$$V^\pi = \mathbb{E}\left[\sum_{t=1}^\infty \alpha^t r(S_{A_t}(t))\right] .$$

EXAMPLE 34.6 To see the relation to Bayesian bandits with discounted rewards consider the following setup. Let $\mathcal{S} = [0,\infty) \times [0,\infty)$ and $\mathcal{G} = \mathfrak{B}(\mathcal{S})$. Then let the initial state of each Markov chain be $S_i(1) = (1,1)$ and define probability kernel μ from $(\mathcal{S}, \mathcal{G})$ to itself by

$$\mu((x,y), A) = \frac{x}{x+y}\delta_{(x+1,y)}(A) + \frac{y}{x+y}\delta_{(x,y+1)}(A) .$$

The reward function is $r(x,y) = x/(x+y)$. The reader should check that this corresponds to a Bernoulli bandit with Beta(1,1) prior on the mean reward of each arm.

One of the most celebrated theorems in the study of bandits is that the optimal policy for this problem is to choose in each round the Markov chain with the largest Gittins index.

THEOREM 34.4 *Let π^* be the policy choosing $A_t = \operatorname{argmax}_i G^*(S_i(t))$. Then $V^{\pi^*} = \sup_\pi V^\pi$ where the supremum is taken over all policies.*

The remainder of the section is devoted to proving Theorem 34.4. The choice of actions produces an interleaving of the rewards generated by each Markov chain and it will be useful to have a notation for these interleavings. For each $i \in [K]$ let $g_i = (g_{it})_{t=1}^\infty$ be a real-valued sequence and $g = (g_1, \dots, g_K)$ be the tuple of these sequences. Given an infinite sequence $(a_t)_{t=1}^\infty$ with $a_t \in [K]$ define the interleaving sequence $I_1(g,a), I_2(g,a), \dots$ by

$$I_t(g,a) = g_{a_t, 1+n_{a_t}(t-1)} \qquad \text{with} \qquad n_i(t-1) = \sum_{s=1}^{t-1} \mathbb{I}\{a_s = i\} .$$

In the special case that g_i is monotone nonincreasing for each i there exists a largest interleaving $I^*(g) = I(g, a^*)$, where $a_t^* = \operatorname{argmax}_i g_{a,n_{t-1,i}}$. The following lemma follows from the Hardy–Littlewood inequality and we leave the proof as an exercise.

LEMMA 34.3 *If $g_{i1} \le g_{i2} \le \cdots$ for each i and $\alpha \in (0,1)$, then*

$$\sum_{t=1}^\infty \alpha^t I_t^*(g) = \sup_a \sum_{t=1}^\infty \alpha^t I_t(g,a) .$$

Proof of Theorem 34.4 Let $\underline{G}(t) = \min_{s \le t} G^*(S_{A_t}(s))$ and define an increasing sequence of stopping times $(\tau_k)_{k=0}^\infty$ by

$$\tau_0 = 1 \qquad \text{and} \qquad \tau_{k+1} = \min\left\{t > \tau_k : A_t \neq A_{\tau_k} \text{ or } \underline{G}(t) < \underline{G}(\tau_k - 1)\right\} .$$

For the Gittins index policy the τ_{k+1} is exactly the stopping time given in Lemma 34.2. Let $k \in \mathbb{N}$ and abbreviate $i = A_{\tau_k}$. Then

$$\begin{aligned}\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t r(S_i(t)) \,\middle|\, \mathcal{F}_{\tau_k}\right] &= \underline{G}(\tau_k)\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \,\middle|\, \mathcal{F}_{\tau_k}\right] \text{ a.s} \\ &= \mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \underline{G}(t)\alpha^t \,\middle|\, \mathcal{F}_{\tau_k}\right] \text{ a.s}\,,\end{aligned}$$

where the first equality follows from the definition of the stopping time and Eq. (34.15) and the second because the definition of the stopping time ensures that $\underline{G}(t) = \underline{G}(\tau_k)$ on $\{\tau_k, \dots, \tau_{k+1} - 1\}$. Let S_{iu} be the state of the ith Markov chain when $T_i(t-1) = u$ and $H_{iu} = \min_{v \le u} G^*(S_{iv})$. The key point is that the distribution of H does not depend on the choice of policy and clearly H_{iu} is monotone nonincreasing in u for each i. Substituting the previous display into the definition of the value function shows that

$$V^{\pi^*} = \mathbb{E}\left[\sum_{k=0}^{\infty} \sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \underline{G}(t)\right] = \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^t I_t(H, A)\right] = \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^t I_t^*(H)\right]$$

For policies other than the Gittins policy we note that

$$\mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t r(S_i(t)) \,\middle|\, \mathcal{F}_{\tau_k}\right] \le \mathbb{E}\left[\sum_{t=\tau_k}^{\tau_{k+1}-1} \alpha^t \underline{G}(t) \,\middle|\, \mathcal{F}_{\tau_k}\right] \text{ a.s}\,.$$

Summing over all k and taking expectation combined with Lemma 34.3 yields

$$V^{\pi} \le \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^t \underline{G}(t)\right] = \mathbb{E}\left[\sum_{t=1}^{n} I_t(H, A)\right] \le \mathbb{E}\left[\sum_{t=1}^{n} I_t^*(H)\right]\,. \qquad \square$$

34.8 Computing the Gittins index

We describe a simple approach that depends on the state space being finite. References to more general methods are given in the bibliographic remarks. Assume without loss of generality that $\mathcal{S} = \{1, 2, \dots, |\mathcal{S}|\}$ and $\mathcal{G} = 2^{\mathcal{S}}$. The matrix form of the transition kernel is $P \in [0,1]^{|\mathcal{S}| \times |\mathcal{S}|}$ and is defined by $P_{ij} = \mu(i, \{j\})$. We also let $r \in [0,1]^{|\mathcal{S}|}$ be the vector of rewards so that $r_i = r(i)$. The standard basis vector is $e_i \in \mathbb{R}^{|\mathcal{S}|}$ and $\mathbf{1} \in \mathbb{R}^{|\mathcal{S}|}$ is the vector with 1 in every coordinate. For $C \subset \mathcal{S}$ we let Q_C be the transition matrix with $(Q_C)_{ij} = P_{ij}\mathbb{I}_C(j)$. For each $i \in \mathcal{S}$ our goal is to find

$$G^*(i) = \sup_{\tau > 1} \frac{\mathbb{E}_i\left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t)\right]}{\mathbb{E}_i\left[\sum_{t=1}^{\tau-1} \alpha^t\right]}\,,$$

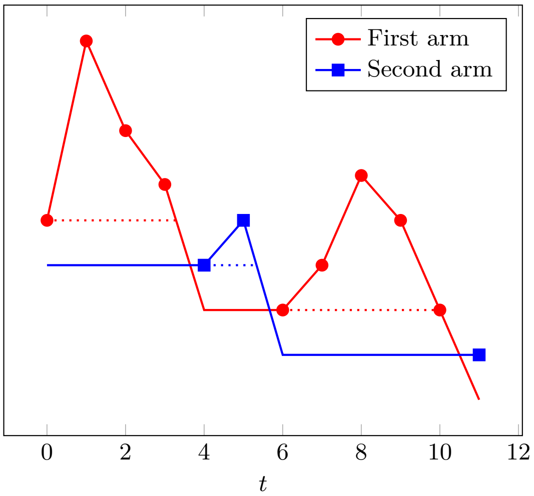


Figure 34.2 The evolution of the fair charge $G^*(S_i(t))$ and prevailing charge $\underline{G}(t)$ for a two-armed bandit. The solid lines indicate the fair charge for each arm, while dotted lines indicate the prevailing charge. Marks indicate which arm is played in each round.

where $\mathbb{E}_i$ is the expectation with respect the measure $\mathbb{P}_i$ for which the initial state is $S_1 = i$. The second part of Lemma 34.2 shows that the stopping time $\tau = \min\{t > 1 : G^*(S_t) \le G^*(i)\}$ attains the supremum in the above display. The set $C_i = \{j : G^*(j) > G^*(i)\}$ is called the continuation region and $S_i = \mathcal{S} \setminus C_i$ is the stopping region. Then the Gittins index can be calculated as

$$G^*(i) = \frac{\mathbb{E}_i\left[\sum_{t=1}^{\tau-1} \alpha^t r(S_t)\right]}{\mathbb{E}_i\left[\sum_{t=1}^{\tau-1} \alpha^t\right]} = \frac{\sum_{t=1}^{\infty} \alpha^t e_i^\top Q_{C_i}^{t-1} r}{\sum_{t=1}^{\infty} \alpha^t e_i^\top Q_{C_i}^{t-1} \mathbf{1}} = \frac{e_i^\top (I - \alpha Q_{C_i})^{-1} r}{e_i^\top (I - \alpha Q_{C_i})^{-1} \mathbf{1}} \,.$$

All this suggests an induction approach where the Gittins index is calculated for each state in decreasing order of their indices. To get started note that the maximum possible Gittins index is $\max_i r_i$ and that this is achievable for state $i = \operatorname{argmax}_j r_j$ with deterministic stopping time $\tau = 2$. Assume that $G^*(i)$ is known for the k states $C = \{i_1, i_2, \dots, i_k\}$ with the largest indices. Then i_{k+1} is given by

$$i_{k+1} = \operatorname{argmax}_{i \notin C} \frac{e_i^\top (I - \alpha Q_C)^{-1} r}{e_i^\top (I - \alpha Q_C)^{-1} \mathbf{1}} \,.$$

If Gauss–Jordan elimination is used for matrix inversion, then the computational complexity of this algorithm is $O(|\mathcal{S}|^4)$. A more sophisticated inversion algorithm would reduce the complexity to $O(|\mathcal{S}|^{3+\varepsilon})$ for some $\varepsilon \le 0.373$, but these are seldom practical. When α is relatively small the inversion can be replaced by directly calculating the sums to some truncated horizon with little loss in accuracy. There are many other ways to compute the Gittins index with better complexity guarantees. We refer the reader to the bibliographic remarks for references.

34.9 Notes

1 An advantage of Bayesian methods is that they automatically and optimally exploit the assumptions. For example, the Bayesian optimal policy for one-armed Bernoulli bandits that we analyzed empirically is essentially the same as its frequentist cousin, but with the tightest possible confidence bounds. This blessing can also be a curse. A policy that exploits its assumptions too heavily can be brittle when those assumptions turn out to be wrong. This can have a devastating effect in bandits where the cost of overly aggressive confidence intervals is large.

2 The issue of conditioning on measure zero sets has been described in many places. We do not know of a practical situation where things go awry. Sensible choices yield sensible posteriors. The curious reader could probably burn a few weeks reading through the literature on the **Borel–Kolmogorov paradox**.

3 Economists have long recognized the role of time in the utility people place on rewards. Most people view a promise of pizza a year from today as less valuable than the same pizza tomorrow. Discounting rewards is one way to model this kind of preference. The formal model is credited to renowned American economist Paul Samuelson [1937], who according to Frederick et al. [2002] had serious reservations about both the normative and descriptive value of the model. While discounting is not very common in the frequentist bandit literature, it appears often in reinforcement learning where it offers certain technical advantages [Sutton and Barto, 1998].

4 Theorem 34.4 only holds for geometric discounting. If α^t is replaced by $\alpha(t)$ where $\alpha(\cdot)$ is not an exponential, then one can construct Markov chains for which the optimal policy is not an index policy. The intuition behind this result is that when $\alpha(t)$ is not an exponential function, then the Gittins index of an arm can change even in rounds you play a different arm and this breaks the interleaving argument [Berry and Fristedt, 1985].

5 The previous note does not apply to one-armed bandits for which the interleaving argument is not required. Given a Markov chain $(S_t)_t$ and horizon n, the undiscounted Gittins index of state s is

$$G^*_n(s) = \sup_{\tau > 1} \frac{\mathbb{E}_s\left[\sum_{t=1}^{\tau-1} r(S_t)\right]}{\mathbb{E}_s[\tau - 1]}\,.$$

If the learner receives reward $\mu_\circ$ by retiring, then the Bayesian optimal policy is to retire in the first round t when $G^*_{n-t+1}(S_t) \leq \mu_\circ$. A reasonable strategy for undiscounted K-armed bandits is to play the arm A_t that maximizes $G^*_{n-t+1}(S_i(t))$. Although this strategy is not Bayesian optimal anymore, it nevertheless performs well in practice. In the Gaussian case it even enjoys frequentist regret guarantees [Lattimore, 2016c].

6 The form of the undiscounted Gittins index was analyzed asymptotically by Burnetas and Katehakis [1997b], who showed the index behaves like the

upper confidence bound provided by KL-UCB. This should not be especially surprising and explains the performance of the algorithm in the previous note. The asymptotic nature of the result does not make it suitable for proving regret guarantees, however.

7 We mentioned that computing the Bayesian optimal policy in finite horizon bandits is computationally intractable. But this is not quite true if n is small. For example, when $n = 50$ and $K = 5$ the dynamic program for computing the exact Bayesian optimal policy for Bernoulli noise and Beta prior has approximately 10^{11} states. A big number to be sure, but not so large that the table cannot be stored on disk. And this is without any serious effort to exploit symmetries. Perhaps for mission-critical applications with small horizon the benefits of exact optimality make the computation worth the hassle.

8 The algorithm in Section 34.8 for computing Gittins index is called **Varaiya's algorithm**. In the bibliographic remarks we give some pointers on where to look for more sophisticated methods. The assumption that $|\mathcal{S}|$ is finite is less severe than it may appear. When the discount rate is not too close to 1, then for many problems the Gittins index can be approximated by removing states that are not reachable from the start state before the discounting means they becomes close to irrelevant. When the state space is infinite there is often a topological structure that makes a discretization possible.

34.10 Bibliographical remarks

There are many texts on Bayesian statistics. It's hard not to recommend the book by Gelman et al. [2014] who is one of the main proponents of Bayesian methods. A more philosophical book that takes a foundational look at probability theory from a Bayesian perspective is by Jaynes [2003]. The careful definition of the posterior can be found in several places, but the recent book by Ghosal and van der Vaart [2017] does an impeccable job. A worthy mention goes to the article by Chang and Pollard [1997], which uses disintegration to formalise the "private calculations" that probabalists so frequently make before writing everything carefully using Nikodym derivatives and regular versions. Theorem 34.2 is well known. For a simple proof see Theorem 5.3 in the book by Kallenberg [2002]. The classic text on optimal stopping is by Robbins et al. [1971], while a more modern text is by Peskir and Shiryaev [2006], which includes a proof of Theorem 34.3 (see Thm. 1.2). For a detailed presentation of exponential families see the book by Lehmann and Casella [2006]. We are not aware of a reference for Theorem 34.1, but Lai [1987] has shown that for sufficiently regular priors and noise models the asymptotic Bayesian optimal regret is $\mathrm{BR}_n^* \sim c \log(n)^2$ for some constant $c > 0$ that depends on the prior/model. The Bayesian approach dominated research on bandits from 1960–1980, with Gittins' result (Theorem 34.4) the most celebrated [Gittins, 1979]. Gittins et al. [2011] has written a whole book on Bayesian bandits. Another book that focusses mostly on the Bayesian problem is by Berry and

Fristedt [1985]. Although it is now more than thirty years old this book is still a worthwhile read and presents many curious and unintuitive results about exact Bayesian policies. As far as we know the earliest fully Bayesian analysis is by Bradt et al. [1956], who studied the finite horizon Bayesian one-armed bandit problem, essentially writing down the optimal policy using backwards induction as presented here in Section 34.6. For more general approximation results there is the article by Burnetas and Katehakis [2003], which shows that under weak assumptions the Bayesian optimal strategy for one-armed bandits is asymptotically approximated by a retirement policy reminiscent of Eq. (34.13). The very specific approach to approximating the Bayesian strategy for Gaussian one-armed bandits is by one of the authors [Lattimore, 2016a], where a precise approximation for this special case is also given. There are at least four proofs of Gittins' theorem [Gittins, 1979, Whittle, 1980, Weber, 1992, Tsitsiklis, 1994]. All are summarized in the review by Frostig and Weiss [1999]. There is a line of work on computing and/or approximating the Gittins index, which we cannot do justice to. The approach presented here for finite state spaces is due to Varaiya et al. [1985], but more sophisticated algorithms exist with better guarantees. A nice survey is by Chakravorty and Mahajan [2014], but see also the articles by Chen and Katehakis [1986], Kallenberg [1986], Sonin [2008], Niño-Mora [2011], Chakravorty and Mahajan [2013]. There is also a line of work on approximations of the Gittins index, most of which are based on approximating the discrete time stopping problem with continuous time and applying free boundary methods [Yao, 2006, and references therein]. We mentioned restless bandits in Chapter 31 on nonstationary bandits, but they are usually studied in the Bayesian context Whittle [1988], Weber and Weiss [1990]. The difference is that now the Markov chain for all actions evolve regardless of the action chosen, but the learner only gets to observe the new state for the action they chose.

34.11 Exercises

34.1 Construct an example demonstrating that for some priors over finite-armed stochastic bandits the Bayesian regret is strictly positive: $\inf_\pi \mathrm{BR}_n(\pi, \mathbb{Q}) > 0$.

34.2 Evaluate the posteriors for each pair of conjugate priors in Section 34.4.

34.3 Prove Theorem 34.1.

For the first part you should use the existence of a policy for Bernoulli bandits such that

$$R_n(\pi, \nu) \le C \min\left\{\sqrt{Kn}, \frac{K\log(n)}{\Delta_{\min}(\nu)}\right\} ,$$

where $C > 0$ is a universal constant and $\Delta_{\min}(\nu)$ is the smallest positive suboptimality gap. Then let $\mathcal{E}_n$ be a set of bandits for which there exists a small enough positive suboptimality gap and integrate the above bound on $\mathcal{E}_n$ and $\mathcal{E}_n^c$. The second part is left as a challenge, though the solution is available.

34.4 Show that in the one-armed bandit of Section 34.6, the Bayesian optimal policy takes the form of a retirement policy.

34.5 Prove Theorem 34.3.

34.6 Prove that the definitions of the Gittins index given in Eq. (34.14) and Eq. (34.15) are equivalent.

34.7 Prove Lemma 34.2.

A proof of this result is given by Frostig and Weiss [1999]. A solution is also available.

34.8 Prove Lemma 34.3.

34.9 This question is about one armed bandits with 1-subgaussian rewards.

(a) Prove the bound in Eq. (34.9) holds for the retirement policy determined by the stopping time in Eq. (34.10).
(b) Explain why the policy has bounded regret when $\Delta \geq 0$.

34.10 Reproduce the experimental results in Section 34.6.