

## 27. Multi-Armed Bandit Problems

*There's one thing I'm really good at, and that's hitting the ball over a net, in a box. I'm excellent.*

—Serena Williams

In this chapter, we add uncertainty to the problem of learning the best alternative to create a class of models known as *multi-armed bandit problems*. In bandit problems, rewards from alternatives are distributions rather than fixed amounts. Bandit problems apply to a wide variety of real-world situations. Any choice among actions that has an uncertain payoff—pharmaceutical drug trials, choice of where to place advertisements, choice among technologies, decisions as to whether to allow laptops in the classroom—can be modeled as bandit problems; so too can the problem of choosing a profession at which we can excel.<sup>1</sup>

A person facing a bandit problem must experiment with alternatives to learn the payoff distributions. This feature of bandit problems creates a trade-off between exploration (searching for the best alternative) and exploitation (choosing the alternative that has performed best so far). Finding an optimal balance in the explore-exploit trade-off requires sophisticated rules and behaviors.<sup>2</sup>

The chapter consists of two parts followed by a discussion on the value of applying models. In the first part of the chapter, we describe a special class of Bernoulli bandit problems, in which each alternative is a Bernoulli urn with unknown proportions of gray and white balls. We describe and compare heuristic solutions, and then show how these solutions can improve comparison tests of drug treatments, advertising plans, and teaching strategies. In the second part, we describe a more general model, where the reward distributions can take any form and the decision-maker has a prior distribution over their types. In that part, we also show how to solve for the Gittins index, which determines the optimal choice.

### **Bernoulli Bandit Problems**


We begin with a subclass of bandit problems in which each alternative has a fixed probability of producing a successful outcome. This first class of bandit problems is equivalent to deciding among a set of Bernoulli urns, each containing different proportions of gray and white balls. Therefore, we refer to these as *Bernoulli bandit problems*. These are also called *frequentist problems* because the decision-maker knows nothing about the distributions. As the decision-maker tries alternatives (explores), she learns about those distributions.


## Bernoulli Bandit Problems

Each of a collection of **alternatives**  $\{A, B, C, D, \dots, N\}$  has an unknown probability of producing a successful outcome,  $\{p_A, p_B, p_C, p_D, \dots, p_N\}$ . In each period, the decision-maker chooses an alternative,  $K$ , and receives a successful outcome with probability  $p_K$ .

For example, suppose a chimney cleaning company has a list of phone numbers of recent home buyers. The company tests three sales pitches: the scheduled appointment approach (“Hello, I’m calling to arrange a time for your annual chimney cleaning”), the concerned questioner approach (“Hello, did you know a dirty chimney can be a fire hazard?”), and the personal touch approach (“Hello, my name is Hildy, and I started this chimney sweeping company with my father fourteen years ago”).

Each sales pitch has an unknown probability of success. Suppose that the company first tries the scheduled appointment approach, and it fails. It then tries the concerned questioner approach and gains a client. That approach also works on the next call but then fails on the next three calls. After that, the company tries the third approach, which works on the first call but fails on the next four. After ten calls, the second approach has the highest success rate, but the first approach was only tried one time. The decision-maker faces a choice between exploiting (choosing the alternative that has worked best) or exploring (returning to the other two alternatives to get more information). This same problem is faced by a hospital selecting among surgical procedures and a pharmaceutical company testing various drug protocols. Each protocol has an unknown probability of success.

To gain insight into the explore-exploit trade-off, we compare two heuristics. The first, *sample-then-greedy*, tries each alternative a fixed number of times,  $M$ , and thereafter chooses the alternative with the highest average payoff. To determine the size of  $M$ , we can refer back to the Bernoulli urn model and the square root rules. The standard deviation of the mean proportion is bounded above by . If each alternative is tested 100 times, the standard deviation of the mean proportion will equal 5%. If we apply a two-standard-deviation rule to identify a significant difference, we can confidently distinguish between proportions that differ by 10%. If one alternative produced successful outcomes 70% of the time and another produced successes 55% of the time, we could place more than 95% confidence on the first being better.

The second heuristic, an *adaptive exploration rate heuristic*, allocates ten initial trials to each alternative. The next twenty trials are allocated in proportions corresponding to the success rates. If in the first ten trials one alternative produced six successes and the other produced only two, then the first alternative would receive three-fourths of the next twenty trials. The second set of twenty trials could also be allocated according to the ratio of the squared success probabilities. If successes continued in the same proportions, the better alternative would then receive , or 90%, of the third set of twenty trials. For each successive set of twenty trials the exponent of the probabilities could be increased at some rate. By increasing the rate of exploitation over time, the second algorithm improves on the first. If one alternative had a much higher probability of success than another, say 80% to 10%, the algorithm would not waste a hundred trials on the second alternative. On the other hand, if the two probabilities of success were close, the algorithm would continue to experiment.<sup>3</sup>

Adherence to the *sample-then-greedy* heuristic not only is inefficient, it can even be unethical. When Robert Bartlett tested an artificial lung, its success rate far surpassed those of the other alternatives. Continuing to test the other alternatives when the artificial lung performed best would have resulted in unnecessary deaths. Bartlett stopped experimenting with the other alternatives. Everyone was given the artificial lung. In fact, that can be shown to be an optimal rule: if an alternative is always successful, keep

choosing that alternative. Experimentation can have no value because no other alternative could perform better.

## Bayesian Multi-Armed Bandit Problems

In a *Bayesian bandit problem*, the decision-maker has prior beliefs over the reward distributions of the alternatives. Given these prior beliefs, a decision maker can quantify the trade-off between exploration and exploitation and (in theory) make optimal decisions in each period. However, except for the simplest of bandit problems, determining the optimal action requires rather tedious calculations. In real-world applications, these exact calculations may be impractical, obliging decision makers to rely on approximations.

## Bayesian Multi-Armed Bandit Problems

A collection of **alternatives**  $\{A, B, C, D, \dots, N\}$  have associated **reward distributions**  $\{f(A), f(B), f(C), f(D), \dots, f(N)\}$ . The decision-maker has prior beliefs over each distribution. In each period, the decision-maker chooses an alternative, receives a reward, and calculates new beliefs based on the reward.

Determining the optimal action relies on a four-step process. First, we calculate the expected immediate reward from each alternative. Second, for each alternative, we update our beliefs about the reward distribution. Third, based on our new beliefs about reward distributions, we determine the best possible actions in all subsequent periods based on what we know. Last, we add the expected reward from the action in the next period to the expected rewards from the optimal future actions. That sum is known as the *Gittins index*. In each period, the optimal action has the largest Gittins index.

Notice that the calculation of the index quantifies the value of exploration. If we try an alternative, the Gittins index does *not* equal the expected reward. It equals the sum of all future rewards assuming we take optimal actions given what we have learned. Computing a Gittins index is difficult. For a (relatively) simple example, suppose there exists a safe alternative that is certain to pay \$500 and a risky alternative that with a 10%






probability will always pay \$1,000. The remaining 90% of the time, the risky alternative pays nothing.

To calculate the Gittins index for the risky alternative, we first ask what could happen: either it always pays \$1,000 or it always pays nothing. We then think through how each outcome would influence our beliefs. If we knew the risky alternative paid \$1,000, we would always choose it. If we knew that the risky arm paid nothing, we would always choose the safe arm in the future.

It follows that the Gittins index for the risky arm corresponds to a 10% probability of a reward of \$1,000 in each period and a 90% probability of a reward of \$500 in every period but the first. For a situation in which we get to choose an alternative many times, this averages out to approximately \$550 per period. The risky alternative is therefore the better choice.<sup>4</sup>



## The Gittins Index: Example

To show how to compute Gittins indices, we consider the following example with two alternatives. Alternative *A* produces a certain reward in  $\{0, 80\}$  with 0 and 80 equally likely. Alternative *B* produces a certain reward in  $\{0, 60, 120\}$  with each equally likely. We assume that the decision-maker wants to maximize reward over ten periods.

**Alternative A:** With probability  the reward equals 0, so alternative *B*, which has an expected reward of 60, will be chosen in all remaining periods. This produces an expected reward of 540 (9 times 60). With probability  the reward equals 80. The optimal choice in the second period even with this outcome is to choose alternative *B* in the second period. With probability , *B* produces a reward of 120, so the total payoff equals 1,160 (80 plus 9 times 120). With probability , *B* produces a reward of 60. In that case, alternative *A* is optimal choice in all subsequent periods generating a total payoff equal to 780 (60 plus 9 times 80). Finally, with probability , *B* produces a reward of 0. In this case as well, alternative *A* is optimal choice in all subsequent periods. The total payoff will 720 (9 times 80).

Combining all three possibilities, it follows that the Gittins index in period one for alternative *A* equals the following:

image

**Alternative B:** With probability  the reward equals 120. If that occurs, the optimal choice in all future periods will also be *B*. Over ten periods the total reward will equal 1,200. If the reward equals zero, then the optimal choice in all future periods will be alternative *A*, which has an expected reward of 40. The expected total reward will equal 360 (9 times 40). If the reward equals 60, then the decision-maker could choose alternative *B* in all future periods, for a total return of 600. However, if she chooses *A* in the second period, half of the time it will always produce a reward of 80, for a total return of 780 (60 plus 9 times 80). The other half of the time it produces a reward of zero, and the optimal choice in all subsequent periods will be *B*, which produces a reward of 60, resulting in a total reward of 540 (9 times 60). It follows that the expected reward from making optimal choices after choosing *A* in the second period equals 660 ().

Combining all the possibilities, it follows that the Gittins index in period one for alternative *B* equals the following:

image

Given these calculations, alternative *B* is the optimal first period choice. The optimal long run choice depends upon what is learned in the first period. If alternative *B* produces an outcome of 120, we stick with *B* forever.

The analysis shows that when taking an action, we care more about the probability that an alternative will be the best than about its expected reward. Moreover, if an alternative produces a very high reward, we should be more likely to select it in the future. In contrast, if it produces an average reward, even a reward above the expected reward of another alternative, we may be less likely to stick with the alternative. That is particularly true in early periods, where we want to look for high-reward alternatives. These insights hold across the many applications discussed. Provided there are not

risks or high costs associated with actions, the model tells us to explore potentially high-reward actions even if they are low probability.

## Summary

A key takeaway from this book is that by learning models a person can make better decisions. We can see that in stark relief by comparing what people should do in the bandit problem with what people actually do. Most people do not try to estimate a Gittins index when confronted with a bandit problem. They fail to do so, in part, because they do not keep data. Only recently, for example, have doctors begun to keep data on the efficacy of the many procedures—the efficacy of the various types of artificial joints or, say, the advantages of stents. Without that data, a doctor cannot determine which action has the highest expected payoff.

Doctors, and everyone else, need data to apply the lessons produced by the model. So if you wanted to learn whether taking a walk before dinner or after dinner resulted in better sleep patterns, you would need to keep track of how well you had slept, and by using a sophisticated heuristic, you could learn which probably works best. That may seem like a lot of effort. And it is, but less so now. New technology enables us to gather data on sleep patterns, pulse rate, weight, and even mood.

Most of us will not gather the necessary data and compute Gittins indices for life choices like when to exercise. The point is only that we could, and if we did, we would see improvement in life choices—in our sleep patterns and general health. Psychologist Seth Roberts performed self-study for twelve years and found that standing at least eight hours a day improved his sleep (though he slept less) and that standing along with getting morning light reduced his upper respiratory infections.<sup>5</sup> We may lack his dedication to self-experimentation. By not keeping data and comparing outcomes, we may go through life skipping breakfast when we would have been better off having grapefruit.

On high-stakes business, policy, and medical decisions where data are easier to gather, applying bandit models is common practice. Businesses, governments, and nonprofits experiment with alternatives and then exploit

those that perform best. In practice, the alternatives may not remain fixed. A government mailer to increase participation in a farm subsidy program may be altered from year to year—say, swapping out a picture of a man with a picture of a woman.<sup>6</sup> This type of continued experimentation can be captured by the models we take up in the next chapter: rugged-landscape models.

## Presidential Elections

We now apply three models to analyze outcomes in presidential elections: the spatial model, the category model, and the multi-armed bandit model.

**Spatial model:** To attract voters, candidates compete in an ideological issue space. Thus, we should expect candidates to tend toward moderate positions, elections to be close, and the winning sequence of parties to be random. Presidential elections are, with a few exceptions, close. To test if the sequence of winners is random, we first construct a time series of the thirty-eight winning parties from 1868 through 2016.

*RRRRDRDRRRRDDRDRRRDDDDDRDRDRRRDDRRDDR*

We can then measure the *block entropy* of subsequences of various lengths. Subsequences of length 1 have entropy 0.98. Subsequences of length 4 have entropy of 3.61. Statistical tests show that we cannot reject that the sequence is random. For comparison, a random sequence of length 38 had block 1 entropy of 1.0 and block 4 entropy of 3.58.

**Category model:** If we think of each state as a category and assume heterogeneity across states, the spatial model implies that once the candidates choose initial positions, some states will not be competitive. The model predicts fierce competition in a handful of moderate states. In 2012, Obama and Romney spent over 96% of their television advertising budgets in ten states. Each spent nearly half of their advertising budgets in three moderate states: Florida, Virginia, and Ohio. In 2016, Clinton and Trump



also spent more than half of their television dollars in three moderate states: Florida, Ohio, and North Carolina.<sup>7</sup>

**Multi-armed bandit model (retrospective voting):** Voters will be more likely to reelect a party that produces good outcomes. Voting for effective parties corresponds to pulls of a lever that generate a high payoff. A strong economy should benefit the incumbent party. Evidence shows that voters are more likely to reelect the party in power when the economy performs well. The effect is larger for the incumbent candidate than for a nonincumbent candidate from the party in power.<sup>8</sup>