CHAPTER 4

RANKING AND SELECTION

Ranking and selection problems arise in many settings. We may have to choose a type of cholesterol drug, the parameters of a chemical process (e.g., temperature, relative mix of certain inputs), or the design of a manufacturing process (choice of equipment, size of buffers, routing of jobs). Testing an alternative might involve running a time-consuming computer simulation, or it might require a physical experiment. We assume that in either case, the measurement involves some noise, which means we may need to repeat the experiment several times. We assume that we have a budget that determines how many times we can perform these experiments, after which we have to take the best design and live with it.

We assume that we have a finite set of alternatives \(X = (1, 2, \ldots, M)\) where \(M\) is “not too large.” We then assume that we have a budget for evaluating these alternatives, and when our budget is exhausted, we have to choose the alternative that appears to be best. For this reason, this is called offline learning, since the cost of a measurement is distinct from the cost of using the resulting design in production.

The assumption of a “finite” set of alternatives is fairly important, and will prove to be somewhat restrictive. It will generally be the case that we plan on testing all the alternatives at least once, although this is not essential. We may use our prior knowledge about the performance of an alternative to decide that it is so poor that
we do not even have to try it. But there are many applications where the number of alternatives is much larger than what the measurement budget will allow us to evaluate. In these cases, we will introduce extensions to the basic model.

4.1 THE MODEL

Recall the Bayesian philosophy that any unknown number is a random variable. We start by assuming that we have a prior distribution of belief about the performance $\mu_x$ of each alternative $x \in \mathcal{X}$ which is normally distributed with mean $\theta_0^x$ and precision $\beta_0^x$. Throughout this book, it is important to recognize that when we talk about the true mean $\mu$ (or $\mu_x$), that $\mu$ is a random variable. Whenever we solve a problem, there is an underlying truth which is unknown to us, which can be represented as a realization of the random variable $\mu$.

At the $n$th iteration (starting with $n = 0$), we choose $x^n$ and observe $W^{n+1}_{x^n}$. We then use this new information to update our belief about $\mu_x$. We do not update priors that are not observed, so our updating equations now look like

$$
\theta_{x}^{n+1} = \begin{cases} 
\frac{\beta_0^x \theta_0^x + \beta^{W_{x^n+1}}}{\beta_0^x + \beta^{W_{x^n+1}}}, & \text{if } x^n = x, \\
\theta_0^x, & \text{otherwise,}
\end{cases}
$$

(4.1)

$$
\beta_{x}^{n+1} = \begin{cases} 
\beta_0^x + \beta^W, & \text{if } x^n = x, \\
\beta_0^x, & \text{otherwise.}
\end{cases}
$$

(4.2)

We are going to make a series of observations $W_1^x, W_2^x, \ldots, W_{N-x^n-1}^x$. These observations depend in part on our decisions of what to measure, and then in part on the random outcome of a measurement. In addition, the probability distribution that describes a measurement is based on the underlying truth, which is also unknown to us. It is useful in our presentation to think of a set $\Psi$ of potential outcomes of $\mu$. The element $\psi \in \Psi$ might be thought of as a state of the world, where $\mu(\psi)$ is the truth when the state of the world is $\psi$. This is easiest to visualize if you assume that there are finitely many states of the world, but everything we are going to do works fine if the set of all possible true values is continuous (for example, if our distribution of belief is normal).

For now, we assume that we have some rule or decision function $X^\pi(K^n)$ which depends on our state of knowledge $K^n = (\theta_x^K, \beta_x^K)_{x \in \mathcal{X}}$ and returns the alternative to be tested in iteration $n + 1$. We refer to $X^\pi(K)$ as a policy indexed by $\pi \in \Pi$. We will discuss the precise meaning of $\pi$ and $\Pi$ later. We refer to the decision function $X^\pi$ and the policy $\pi$ interchangeably.

We use the notation $\omega$ to denote a sample realization of all the random quantities in our problem. Think of a matrix $W(\omega)$ of numbers, such that the number in the $n$th row and $x$th column represents a sample realization $W_{x^n}^n(\omega)$ of the measurement $W_{x^n}^n$. Figure 4.1 illustrates this idea with three sample realizations of $W_{x^n}^n$ (three different values of $\omega$) for $n = 1, \ldots, 10$ and $x = 1, 2, 3$. It is useful to think of generating
Table 4.1 Three sample realizations of three alternatives over 10 observations.

<table>
<thead>
<tr>
<th>n</th>
<th>$W^n_1$</th>
<th>$W^n_2$</th>
<th>$W^n_3$</th>
<th>$W^n_1$</th>
<th>$W^n_2$</th>
<th>$W^n_3$</th>
<th>$W^n_1$</th>
<th>$W^n_2$</th>
<th>$W^n_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>37.4</td>
<td>61.4</td>
<td>59.2</td>
<td>38.1</td>
<td>66.9</td>
<td>55.7</td>
<td>33.4</td>
<td>66.5</td>
<td>40.1</td>
</tr>
<tr>
<td>2</td>
<td>24.3</td>
<td>65.2</td>
<td>56.0</td>
<td>28.0</td>
<td>57.6</td>
<td>59.3</td>
<td>25.4</td>
<td>60.1</td>
<td>59.5</td>
</tr>
<tr>
<td>3</td>
<td>30.5</td>
<td>64.5</td>
<td>56.5</td>
<td>38.9</td>
<td>59.3</td>
<td>50.2</td>
<td>22.7</td>
<td>59.5</td>
<td>48.7</td>
</tr>
<tr>
<td>4</td>
<td>20.5</td>
<td>63.2</td>
<td>55.9</td>
<td>34.1</td>
<td>57.7</td>
<td>57.1</td>
<td>24.1</td>
<td>59.8</td>
<td>58.8</td>
</tr>
<tr>
<td>5</td>
<td>29.1</td>
<td>64.5</td>
<td>44.8</td>
<td>36.3</td>
<td>60.1</td>
<td>56.4</td>
<td>37.3</td>
<td>65.1</td>
<td>57.1</td>
</tr>
<tr>
<td>6</td>
<td>36.7</td>
<td>62.1</td>
<td>59.6</td>
<td>37.0</td>
<td>53.6</td>
<td>42.9</td>
<td>20.7</td>
<td>51.8</td>
<td>42.9</td>
</tr>
<tr>
<td>7</td>
<td>27.3</td>
<td>53.5</td>
<td>44.8</td>
<td>27.7</td>
<td>60.9</td>
<td>48.9</td>
<td>32.0</td>
<td>53.4</td>
<td>53.3</td>
</tr>
<tr>
<td>8</td>
<td>26.8</td>
<td>57.7</td>
<td>48.1</td>
<td>28.8</td>
<td>68.6</td>
<td>54.4</td>
<td>34.3</td>
<td>53.0</td>
<td>48.6</td>
</tr>
<tr>
<td>9</td>
<td>40.0</td>
<td>58.6</td>
<td>42.8</td>
<td>39.6</td>
<td>52.8</td>
<td>53.5</td>
<td>31.1</td>
<td>64.0</td>
<td>59.9</td>
</tr>
<tr>
<td>10</td>
<td>21.4</td>
<td>51.6</td>
<td>56.4</td>
<td>33.1</td>
<td>54.5</td>
<td>42.7</td>
<td>33.4</td>
<td>56.7</td>
<td>42.7</td>
</tr>
</tbody>
</table>

All these realizations, but then we are going to choose a measurement policy $\pi$ that determines which of these realizations that we are actually going to see. That is, we only get to see $W^n_x$ if we choose to observe $x = X^{\pi}(K^n)$.

Since the true values $\mu_x$ are also random variables, we can let $\psi$ be a sample realization of the truth. That is, $\mu(\psi)$ is a particular set of truth values $\mu_1(\psi), ..., \mu_M(\psi)$. Our Bayesian model makes the fundamental assumption that $\mu_x \sim N(\theta_0^x, \beta_0^x)$, that is, our prior distribution is assumed to be accurate on average. Therefore, the sample path $\psi$ is generated from the prior distribution.

With this, we finally can compute a sample realization of the value of a policy using

$$F^n(\mu(\psi), W(\omega)) = \max_{x \in X} \theta^N_x(\psi, \omega, \pi).$$

(4.3)

In other words, we take our estimate of $\theta^N_x$ for each $x$ after $N$ measurements and then choose the alternative with the highest estimate. The estimates $\theta^N_x$ depend on both types of sample realizations, $\omega$ and $\psi$. Our beliefs change over time in a way that depends on the exact numbers $W^1(\omega), ..., W^N(\omega)$. At the same time, the probability distribution of the observations (the likelihood of observing, say, $\omega_i$ instead of $\omega_i$) is determined by the truth $\mu(\psi)$.

Moreover, the measurement policy $\pi$ determines which observations we actually see: recall again that $W^n = (W^n_x)_{x \in X}$ is the vector of outcomes for all alternatives, but that we observe only one component of this vector. Thus, the precise distribution of $F^n$ depends on $\pi$. We did not observe this distinction in Chapter 2, where we were given an observation. Now, we have the ability to choose what we want to observe.

If we want to evaluate a policy, we need to compute $F^n(\mu(\psi), W(\omega))$ for different samples $W(\omega)$ and different truths $\mu(\psi)$, and take an average. Let $\omega_k$ be the $k$th sample of measurements where $k = 1, 2, \ldots, K$, and let $\psi_\ell$ be a particular truth
Table 4.2  Three sample realizations of both a truth $\mu$ and a set of sample realizations $W$ drawn from this truth.

<table>
<thead>
<tr>
<th></th>
<th>$\omega = \omega_1$</th>
<th>$\omega = \omega_2$</th>
<th>$\omega = \omega_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\mu_1(\omega_1)$</td>
<td>$\mu_2(\omega_1)$</td>
<td>$\mu_3(\omega_1)$</td>
</tr>
<tr>
<td></td>
<td>32.1</td>
<td>57.8</td>
<td>55.8</td>
</tr>
<tr>
<td>$n$</td>
<td>$W^n_1$</td>
<td>$W^n_2$</td>
<td>$W^n_3$</td>
</tr>
<tr>
<td>1</td>
<td>33.8</td>
<td>57.7</td>
<td>53.3</td>
</tr>
<tr>
<td>2</td>
<td>37.0</td>
<td>55.0</td>
<td>59.7</td>
</tr>
<tr>
<td>3</td>
<td>36.0</td>
<td>54.9</td>
<td>51.9</td>
</tr>
<tr>
<td>4</td>
<td>30.8</td>
<td>53.3</td>
<td>58.6</td>
</tr>
<tr>
<td>5</td>
<td>30.9</td>
<td>58.6</td>
<td>58.3</td>
</tr>
<tr>
<td>6</td>
<td>29.7</td>
<td>56.3</td>
<td>55.6</td>
</tr>
<tr>
<td>7</td>
<td>32.1</td>
<td>57.0</td>
<td>57.9</td>
</tr>
<tr>
<td>8</td>
<td>31.1</td>
<td>53.9</td>
<td>54.4</td>
</tr>
<tr>
<td>9</td>
<td>35.9</td>
<td>55.5</td>
<td>50.8</td>
</tr>
<tr>
<td>10</td>
<td>35.4</td>
<td>60.5</td>
<td>54.7</td>
</tr>
<tr>
<td></td>
<td>$\mu_1(\omega_2)$</td>
<td>$\mu_2(\omega_2)$</td>
<td>$\mu_3(\omega_2)$</td>
</tr>
<tr>
<td></td>
<td>32.6</td>
<td>62.7</td>
<td>60.5</td>
</tr>
<tr>
<td></td>
<td>36.0</td>
<td>65.0</td>
<td>58.3</td>
</tr>
<tr>
<td></td>
<td>30.0</td>
<td>57.7</td>
<td>57.8</td>
</tr>
<tr>
<td></td>
<td>30.7</td>
<td>59.4</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>26.3</td>
<td>62.4</td>
<td>60.9</td>
</tr>
<tr>
<td></td>
<td>31.6</td>
<td>57.8</td>
<td>59.9</td>
</tr>
<tr>
<td></td>
<td>29.1</td>
<td>59.6</td>
<td>59.0</td>
</tr>
<tr>
<td></td>
<td>31.5</td>
<td>65.9</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td>33.7</td>
<td>62.4</td>
<td>55.9</td>
</tr>
<tr>
<td></td>
<td>31.6</td>
<td>58.9</td>
<td>60.8</td>
</tr>
<tr>
<td></td>
<td>30.8</td>
<td>63.3</td>
<td>55.3</td>
</tr>
<tr>
<td></td>
<td>36.0</td>
<td>62.3</td>
<td>62.9</td>
</tr>
<tr>
<td></td>
<td>38.9</td>
<td>54.9</td>
<td>64.2</td>
</tr>
</tbody>
</table>

where $\ell = 1, 2, \ldots, L$ is a set of potential truths. We can compute an average using

$$
\bar{\bar{F}}^\pi = \frac{1}{L} \sum_{\ell=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} F^\pi(\mu(\psi_\ell), W(\omega_k)) \right).
$$

(4.4)

This hints at how we would like to state our objective in a formal way. Let $E_W$ be the expectation over all possible measurements. The measurements we observe depend on our measurement policy $\pi$, since this determines which elements of the vector $W$ that we actually observe. For this reason, we write the expectation over measurements as $E_W$. Let $E_\mu$ be the expectation over all possible truths. We can write our objective as

$$
\max_{\pi \in \Pi} E_\mu E_W F^\pi(\mu, W),
$$

(4.5)

where $F^\pi(\mu, W) = \max_{x \in X} \theta^N_x$ is the sample estimate of the objective function, computed using the truth $\mu(\psi)$, the measurements $W^1(\omega), \ldots, W^N(\omega)$ and the policy $\pi$ that determines which measurements we actually observe.

Mathematically, we can lump the uncertainty about the truth and the uncertainty about measurements into a single space of outcomes. Table 4.2 illustrates how we might represent 3 sample realizations. Here, $\omega$ represents both a realization of the truth $\mu$ and a realization of the measurements $W^n_\omega$ drawn from this truth. We would
calculate an estimate of $\bar{F}^\pi$ using

$$\bar{F}^\pi = \frac{1}{K} \sum_{k=1}^{K} F^\pi(\mu(\omega_k), W(\omega_k)).$$

(4.6)

Note that calculating $\bar{F}^\pi$ using (4.6) or (4.4) is conceptually the same, although the actual numbers may be slightly different when we code up the calculation. Using (4.6), we calculate $F^\pi(\mu(\omega_k), W(\omega_k))$ $K$ times, while if we use (4.4), we calculate $F^\pi(\mu(\omega_k), W(\omega_k))$ $L \times K$ times.

Now let $\mathbb{E}$ be the expectation over the outcomes $\omega \in \Omega$ where now both $\mu$ and $W$ are changing with $\omega$. In this case, we would write our objective function as

$$\max_{\pi \in \Pi} \mathbb{E}F^\pi(\mu, W).$$

(4.7)

This is a perfectly legitimate way of writing our objective function. But it is important to realize that the expectation is summing (or integrating) over our uncertainty in our belief about the truth as well as the randomness in our measurements.

The preceding discussion becomes relevant when we simulate the performance of a policy. In (4.3), we first generate $L$ truth values, and then generate $K$ sets of observations for each one. We thus obtain a more precise estimate of the average performance of the policy on each truth. In (4.6), we generate a new truth for each set of observations. We can thus simulate a larger number of truth values, but we will sacrifice some precision in our estimate of average performance for any given truth.

### 4.2 MEASUREMENT POLICIES

Central to the concept of optimal learning is a measurement policy. This is a rule that tells us which action $x$ we should take next in order to observe something new. In addition, we may also be receiving rewards or incurring costs, which have to be balanced against the value of the information being gained.

In this section, we contrast deterministic versus sequential policies and then provide a mathematical framework for finding optimal sequential policies. Unfortunately, this framework does not provide us with practical policies that we can compute. The section closes with a presentation of a number of the more popular heuristic policies that have been used on this problem class.

#### 4.2.1 Deterministic Versus Sequential Policies

Before we begin our presentation, it is important to make a distinction between what we call deterministic policies and sequential policies. In a deterministic policy, we decide what we are going to measure before we begin making any measurements.

For example, a business may decide to perform four market research studies in different parts of the country before finalizing the pricing and advertising strategy in a full roll-out to the entire country. The decision to do four studies (and their locations) is made before we have any information from any of the studies.
There are problem classes where deterministic policies are optimal. For example, we might be interested in making measurements that minimize some function of the variance of the quantities that we are trying to estimate. If you take a close look at our formula for updating the variance (or equivalently the precision) in equation (4.2), we see that our estimate of the variance is a deterministic function of what we choose to measure. This means that any rule that depends purely on the variance can be solved deterministically.

Our interest is primarily in sequential policies, where the decision of what to measure next may depend on past measurements. For example, when we are trying to find the shortest path, we may decide to continue sampling a path if it remains competitive, or give up on a path if the observed travel times are simply too long. Our decisions of what to measure in this case depend on the outcomes of prior measurements.

### 4.2.2 Optimal Sequential Policies

It is possible to provide a mathematical characterization of an optimal measurement policy. Imagine that we are moving over a graph, where we gain information about the graph (including other links) every time we make a transition from one node to the next. For this system, let $S$ be the state variable which captures both our physical state (which might be the node we are sitting at) and the knowledge state (what we know about all the links in the graph as a result of our previous measurements). Further, let $S^n$ be the state after $n$ measurements. We note in passing that the ranking and selection problem does not have a physical state; $S^n$ consists purely of the state of knowledge. Whenever we discuss this problem, we can use $S^n$ interchangeably with our earlier notation $K^n$.

Assume that we are in state $S^n$ and we make a decision $x^n$ which might change both our physical state (we move from $i$ to $j$) and our knowledge state (which might be updated using equations (4.1) and (4.2)). Let $W^{n+1}$ be the next observation we observe which is used to determine $S^{n+1}$. When we do not want to get into the details of how the state changes, we introduce a transition function, often referred to as a state model (or sometimes simply “model”) $S^M(\cdot)$ which updates the state using

$$S^{n+1} = S^M(S^n, x^n, W^{n+1}).$$

When we have normally distributed beliefs, our transition function (or knowledge transition function) is given by equations (4.1) and (4.2).

Next let $V(S^n)$ be the value of being in state $S^n$. The quantity $C(S^n, x)$ captures our total contribution (or reward), minus any measurement costs, from being in state $S^n$ and taking action $x$. Assume we wish to maximize the total discounted reward, with discount factor $\gamma$. Bellman’s equation characterizes the optimal decision using

$$V(S^n) = \max_x \left( C(S^n, x) + \gamma \mathbb{E}\{ V(S^{n+1}) \mid S^n \} \right). \quad (4.8)$$

We let $x^n$ represent the optimal solution to (4.8). We let $X^*(S)$ be the complete mapping of states $S \in \mathcal{S}$ to actions $x \in \mathcal{X}$, where $\mathcal{X}$ describes the set of feasible
actions. We refer to the function $X^*(S)$ as the optimal policy if it describes the solution to (4.8) for all states $S^n \in S$.

It may be mathematically comforting to characterize the optimal policy, but equation (4.8) is virtually impossible to solve, even for very small problems. The simplest knowledge state uses at least one continuous variable for each action $x$. Calculating a value function with as few as two continuous dimensions can, in practice, be quite a challenge. Needless to say, we do not have very many problems of practical significance that meets this modest criterion. Not surprisingly, the field of optimal learning consists primarily of finding shortcuts or, failing this, good heuristics.

### 4.2.3 Heuristic Policies

Our goal, ultimately, is to find the best possible policies for learning. The reality, however, is that most of the time we are happy to find good policies. Below are some popular methods that have been suggested for problems that are typically associated with discrete selection problems, which is to say that the set of measurement decisions is discrete and “not too large.”

**Pure Exploration** A pure exploration strategy might sample a decision $x^n = x$ with probability $1/M$ (the probabilities do not have to be the same - they just have to be strictly positive). We would only use a pure exploration policy if we were focusing purely on estimating the value of each choice, as opposed to making a good economic decision. If we really are trying to find the best value of $\mu_x$, a pure exploration strategy means that we would spend a lot of time measuring suboptimal choices.

Pure exploration can be effective for offline learning problems, especially when the number of choices is extremely large (and especially if a measurement $x$ is multidimensional). This is often what has been used when we are given a dataset of observations from which we have to fit a model so that we can find the best choice or design. With offline learning, it does not hurt us to observe a poor choice, and extreme choices can give us the best estimates of a function. For example, consider

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure4.1.png}
\caption{Estimating a function where we sample closest to the point that might be best (a), versus sampling a wide range of points so that we get a better estimate of the function (b).}
\end{figure}
the problem of fitting a linear regression model of the form
\[ Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon. \]
Imagine that \( x \) is a scalar between 0 and 10, and we believe the highest values of \( Y \) are likely to be found close to the middle of the range. Figure 4.1(a) shows what happens when we draw most of our samples from a narrow range. We may get a lot of data to estimate the function around those points, but we do not get the kind of information we need to get an accurate estimate of the function, which would allow us to do the best job finding the maximum. In Figure 4.1(b), we explore a wider range of points, which allows us to do a better job of estimating the entire curve. For instance, we discover that \( Y \) decreases once \( x \) is large enough, whereas 4.1(a) leads us to believe that \( Y \) always increases with \( x \).

**Pure Exploitation**  
Exploitation means making the best decision given our current set of estimates (we are “exploiting” our knowledge). So, after iteration \( n \), we would next measure
\[
x^n = \arg \max_{x \in X} \theta^n_x.
\]
This strategy would seem to focus our energy on the options that appear to be the best. However, it is very easy to get stuck measuring choices that seem to be the best, especially when we simply had some bad luck measuring the better choices.

Pure exploitation is a common strategy in online problems, where we have to live with the results of each measurement. With pure exploitation, we can always defend our choice because we are doing what we believe is the best, but we may be ignoring errors in our own beliefs.

**Epsilon-Greedy Exploration**  
A simple strategy that avoids the limitations of pure exploration and pure exploitation is to use a mixed strategy, where we explore with probability \( \epsilon \) (known as the exploration rate) and we exploit with probability \( 1 - \epsilon \). The value of \( \epsilon \) has to be tuned for each application.

Mixing exploration and exploitation is appealing because it allows you to spend more time evaluating the choices that appear to be best (to make sure this is the case) while still doing a certain amount of exploration. As our measurement budget goes to infinity, we can still provide guarantees that we will find the best alternative because of the exploration component. But this policy still suffers from the significant limitation that when we do choose to explore, we sample from the entire population of alternatives, which may be extremely large, including choices that are clearly suboptimal.

The problem with a mixed exploration/exploitation strategy with fixed \( \epsilon \) is that the correct balancing of exploration and exploitation changes with the number of iterations. In the beginning, it is better to explore. As we build confidence in our estimates, we would prefer to exploit more. We can do this by using an exploration probability \( \epsilon^n \) at iteration \( n \) that declines with \( n \). We have to make sure it does not decline too quickly. We do this by setting
\[
\epsilon^n = \frac{c}{n}
\]
for $0 < c < 1$. If we explore, we would choose measurement $x$ with probability $1/|\mathcal{X}|$. This means that in the limit, the number of times we will measure $x$ is given by

$$\sum_{n=1}^{\infty} \frac{c}{n|\mathcal{X}|} = \infty.$$ 

This assures us that we will estimate each measurement $x$ perfectly, but as the measurements progress, we will spend more time measuring what we think are the best choices.

**Boltzmann Exploration** A different strategy for balancing exploration and exploitation is known as Boltzmann exploration. With this strategy, we sample measurement $x$ with probability $p^n_x$ given by

$$p^n_x = \frac{e^{\rho \hat{\theta}_n^x}}{\sum_{x' \in \mathcal{X}} e^{\rho \hat{\theta}_n^{x'}}}. \tag{4.9}$$

This policy is also known as the *soft max* policy. If $\rho = 0$, we are going to sample each measurement with equal probability (pure exploration). As $\rho \to \infty$, we will sample the measurement with the highest value of $\hat{\theta}^n$ with probability 1 (pure exploitation). In between, we explore the better options with higher probability. Furthermore, we can make $\rho$ increase with $n$ (which is typical), so that we explore more in the beginning, converging to a pure exploitation strategy.

Care should be used when computing the probabilities using a Boltzmann distribution, especially if you are increasing $\rho$ as you progress to focus attention on the best alternatives. The problem is that the exponent $\rho \hat{\theta}_n^x$ can become so large as to make it impossible to evaluate $e^{(\rho \hat{\theta}_n^x)}$. A better way is to first compute

$$\hat{\mu}_n = \max_{x \in \mathcal{X}} \theta^n_x,$$

and then compute the probabilities using

$$p^n_x = \frac{e^{\rho(\hat{\theta}_n^x - \hat{\mu}_n)}}{\sum_{x' \in \mathcal{X}} e^{\rho(\hat{\theta}_n^{x'} - \hat{\mu}_n)}}.$$ 

This calculation can be further streamlined by excluding any choices $x$ where $\theta^n_x$ is sufficiently far from $\hat{\mu}_n$ (for example, where $\rho|\hat{\mu}_n - \theta^n_x| > 10$).

A limitation of Boltzmann exploration is that computing these probabilities can be fairly expensive when there are large numbers of potential measurements. It is a popular policy in computer science where problems typically have fewer than 100 alternatives, but $x$ might be a vector, producing an exponentially large number of choices. However, Boltzmann exploration has the distinct advantage that our exploration is based on our belief about the quality of each alternative. As a result, we do not spend much time evaluating truly bad alternatives.
**Interval Estimation** Imagine that instead of choosing a measurement that we think is best, we will choose a measurement that we think might eventually be best if we were to take enough measurements. With this idea, we might construct a confidence interval and then value an option based on the upper limit of, say, a 95% confidence interval. Letting $\alpha$ be our confidence level and denoting by $z_\alpha$ the standard normal deviate leaving $\alpha$ in the upper tail, our upper limit would be

$$\nu^{LE,n}_x = \theta^n_x + z_\alpha \sigma^n_x,$$  

(4.10)

where $\sigma^n_x = \sqrt{\frac{1}{n} \beta^{nx}}$ is the standard deviation of the distribution of our belief at time $n$. When we use an interval exploration policy, we choose the measurement $x^n$ with the highest value of $\nu^{LE,n}_x$.

Although the interpretation as the upper limit of a confidence interval is appealing, the confidence level $\alpha$ carries no particular meaning. Instead, $z_\alpha$ is simply a tunable parameter. It has been reported in the literature that values around 2 or 3 work best for many applications, but it is possible to construct problems where the best value may be anywhere from 0.1 to 100 (the high values arise when the priors are really poor). Furthermore, it has been found that the algorithm can be very sensitive to the choice of $z_\alpha$. However, if properly tuned, this policy can work extremely well in many settings.

Interval estimation introduces two important dimensions relative to policies based on exploration and exploitation. First, like Boltzmann estimation, the likelihood of measuring a choice depends on how well we think the choice may work. This means that we will avoid exploring options that seem to be genuinely bad. Second, we are going to be more willing to explore options which we are more uncertain about. The term $z_\alpha \sigma^n_x$ has been called the “uncertainty bonus.” As we explore an option with a high value of $\nu^{LE,n}_x$, $\sigma^n_x$ will decrease, which will often push us to try other options.

IE is not guaranteed to find the best option, even with an infinitely large measurement budget. It is possible for us to get some poor initial estimates of what might be the best option. If $\theta^n_x$ is low enough for some choice $x$, we may never revisit this choice again. Our experimental work suggests that this can happen, but rarely.

**Chernoff Interval Estimation** A variant of interval estimation is a method where the confidence interval is derived using Chernoff’s inequality. This is computed using

$$\nu^{Ch,n}_x = \theta^n_x + \frac{\alpha + \sqrt{2n \sigma^n_x^2 \alpha + \alpha^2}}{n},$$  

(4.11)

where

$$\alpha = \ln \frac{2nM}{\delta}.$$  

Here, $M$ is the number of alternatives and $0 < \delta < 2$ is a tolerance parameter that controls the likelihood that $\nu^{Ch,n}_x$ is greater than the true value $\mu_x$. The parameter $\delta$ comes from the result that if (4.11) is calculated $n$ times for all $M$ alternatives, then $\nu^{Ch,n}_x \geq \mu_x$ with probability $1 - \frac{\delta}{2}$ for all $n$ and $x$. Alternatively, it is possible to show that the probability that $\nu^{Ch,n}_x < \mu_x$ for a particular iteration $n$ and alternative $x$ is less than $\frac{\delta}{2nM}$. 
4.3 EVALUATING POLICIES

Now that we have an initial sense of some possible measurement policies, we have to start thinking about how to compare them so we can choose the best one. It turns out that comparing one measurement policy to another is a relatively subtle exercise.

To start, consider a simple deterministic optimization problem, where we are trying to find the best value of a function $f(x)$. If we have two solutions $x^1$ and $x^2$, all we have to do is compare $f(x^1)$ and $f(x^2)$ to see which is bigger. Although finding a good solution can be hard for certain classes of deterministic problems, deciding when one answer is better than another is relatively simple.

When we have a stochastic problem, the comparison is a bit harder. Imagine that we have a function $F(x, W)$ that involves a random variable $W$. We would like to find the value $x$ that maximizes $\mathbb{E}F(x, W)$. This could be a shortest path problem, where the random variable $W$ is the vector of costs on all the links. It could be a problem involving the allocation of vaccines to different parts of the country; $x$ could be a vector determining how many vaccine doses are sent to each region, after which we observe the random demands $W$ for the vaccine. Assume now that we cannot compute the expectation. Instead, we might use Monte Carlo methods to observe $n$ sample realizations $\omega_1, \omega_2, \ldots, \omega_n$ (see Section 2.4 for a brief review of Monte Carlo simulation methods). We would then evaluate a solution $x^1$ using

$$\bar{F}(x^1) = \frac{1}{n} \sum_{k=1}^{n} F(x^1, W(\omega_k)).$$

This is an estimate of how well $x^1$ performs. We would similarly calculate $\bar{F}(x^2)$, and we could compare $\bar{F}(x^1)$ and $\bar{F}(x^2)$. Of course, even if one is better than the other, we should compute confidence intervals around each to see if the difference is statistically different. So, this is a bit harder, but conceptually it is the same comparison that we would undertake with the deterministic case.

Now think about what happens when we compare two measurement policies. Consider our ranking and selection problem where we use a measurement policy $\pi$ to determine how often we sample the value of a decision $x$. Perhaps we want to compare Boltzmann exploration to interval estimation, or we want to compare Boltzmann exploration with parameter $\rho = \frac{1}{2}$ to the results when we use $\rho = .2$.

After $N$ measurements, $\theta^N_x$ is our belief about the value of alternative $x$. We can use these beliefs to find the best choice, which we can write using

$$x^\pi = \arg \max_{x \in X} \theta^N_x.$$  \hfill (4.12)

How well did we do with our measurements? We might evaluate our solution using

$$F^\pi = \max_{x \in X} \theta^N_x = \bar{\theta}^N_{x^\pi}.$$  \hfill (4.12)

Of course, $\bar{\theta}_{x^\pi}$ is random. The crucial point to remember is the same issue that we raised in Section 4.1: The value of $\bar{\theta}_{x^\pi}$ that we observe in our simulations depends
RANKING AND SELECTION

on the generated outcomes of the observations, as well as the generated value of the truth. In Section 4.1, we distinguished these two types of generated outcomes by using \( \omega \) to denote a sample realization of an observation and using \( \psi \) to denote a sample realization of the truth.

We proceed as before. For \( \ell = 1, \ldots, L \), let \( \mu(\psi_\ell) \) be sampled truth values, while \( W(\omega_k) \) for \( k = 1, \ldots, K \) are sampled observation values. We let \( x^\pi(\mu(\psi_\ell), W(\omega_k)) \) be our choice of the best alternative, which depends on the truth \( \mu(\psi) \), the sampled observations \( W(\omega) \), and the measurement policy \( \pi \). The estimates after \( N \) observations \( \theta^N_\pi(\mu, W) \) also depend on the truth \( \mu(\psi) \), the sampled observations \( W(\omega) \) and the policy \( \pi \) (which we suppress for notational compactness). We can create an estimate

\[
\bar{F}^\pi = \frac{1}{L} \sum_{\ell=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} \theta^N_\pi(\mu(\psi_\ell), W(\omega_k)) (\mu(\psi_\ell), W(\omega_k)) \right),
\]

representing average performance across \( L \) truths and \( K \) sets of observations per truth. We can then compare policies \( \pi_1 \) and \( \pi_2 \) by computing \( \bar{F}^\pi_1 \) and \( \bar{F}^\pi_2 \). If \( \bar{F}^\pi_1 > \bar{F}^\pi_2 \), we might conclude that \( \pi_1 \) is better. Of course, we should again find confidence intervals for each to see if the difference is significant.

There is a neat trick that can make policy evaluation much more efficient. Since we are generating the truth values \( \mu(\psi_\ell) \), they are now known to us, meaning that we can use them directly to evaluate performance. Define \( F^\pi(\mu(\psi), W(\omega)) = \mu x^\pi(\mu(\psi), W(\omega))(\psi) \), to be the true value of the alternative selected by policy \( \pi \). The policy does not get to see the true values \( \mu(\psi) \). The alternative \( x^\pi = \arg \max \mu x(\psi, W(\omega)) \) is chosen purely based on the time-\( N \) beliefs. However, since we have generated a sample realization of \( \mu \), we can use the true value of this alternative to evaluate the policy. Thus, (4.13) becomes

\[
\bar{F}^\pi = \frac{1}{L} \sum_{\ell=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} \mu x^\pi(\mu(\psi_\ell), W(\omega_k)) (\psi) \right),
\]

On average over many sample realizations, \( \theta^N_\pi \) is just a noisy estimate of \( \mu x \), so these two approaches are equivalent. However, (4.14) provides a more precise estimate with less variance. Essentially we are eliminating the noise due to the observations. Alternatively, we could define the regret \( \bar{R}^\pi \) as

\[
\bar{R}^\pi = \frac{1}{L} \sum_{\ell=1}^{L} \left( \frac{1}{K} \sum_{k=1}^{K} (\mu x^*(\psi) - \mu x^\pi(\mu(\psi_\ell), W(\omega_k)) (\psi)) \right),
\]

where \( \mu x^*(\psi) = \max \mu x(\psi) \) is the best we can do for a particular truth. The regret has a lower bound of zero, which provides a nice reference point.

We can use either equation (4.13) (which uses estimates of the value of each alternative) or (4.14) (which uses the simulated truths) to evaluate the value of a
policy. In Section 4.4.2, we show that (in expectation) \( \bar{F}^\pi \), given by equation (4.13) using the estimates \( \theta^N_x \), provides an unbiased estimate of \( \bar{F}^\pi \) as given by equation (4.14), where we use the assumed truth \( \mu(\psi) \). If you are running simulations to find the best policy, it is better to use equation (4.14), since (4.13) introduces additional statistical errors.

We suggest that the best environment for identifying good learning policies is inside the computer, where you can assume a truth and then try to discover the truth. Once you have decided on a good learning policy, you can go to the field where the truth is unknown, and you only have access to the estimates \( \theta^N_x \).

4.4 MORE ADVANCED TOPICS*

Ranking and selection is a relatively simple problem, as are the basic policies that are introduced in this chapter. But this simple problem hides some fairly advanced probabilistic concepts. In this section we introduce readers with an interest in a more advanced treatment to some of these concepts.

4.4.1 An Alternative Representation of the Probability Space

In Section 4.1, we represented a sample realization of observations \( W \) in Table 4.1 (or Table 4.2) as if specifying \( \omega \) had the effect of telling us what all possible realizations of \( W \) might be. That is, if we fix \( \omega \), we create a table of what \( W \) would be for every alternative \( x \) and every measurement \( n \). Let \( P(\omega) \) be the probability of outcome \( \omega \) (assuming these are discrete). We write our objective function \( F^\pi(\mu, \Omega) \) as a function of \( \pi \) because the policy for choosing which alternative to measure is embedded in the objection function. If we wanted to take its expectation, we would write

\[
F^\pi = \mathbb{E}F^\pi(\mu, W) = \sum_{\omega \in \Omega} P(\omega) F^\pi(\mu(\omega), W(\omega)).
\]

For some (and possibly many), this is not the most natural way of thinking about a sample realization. Another way of thinking about a sample path is to assume that for iteration \( n \), we first choose \( x^n \) and then observe \( W_x^n(\omega_\pi^n) \). We essentially “separate” the distributions of our observations from one another, so \( \omega_\pi^n \) now refers to the outcome of the \( n \)th observation only. That is, we only get to see the outcome after we choose the alternative we are going to observe according to the policy \( \pi \).

With this construction, a sample realization would consist of

\[
(x^0, W^1_x(\omega_\pi^1), x^1(\omega_\pi^1), W^2_x(\omega_\pi^2), \ldots, x^n(\omega_\pi^n), \ldots, x^{N-1}(\omega_\pi^{N-1}), W^N_x(\omega_\pi^N)).
\]

That is, a sample path is now a sequence of decisions and observations. In this interpretation, our decisions are also random variables. For instance, our decision at time \( n \) is based on the knowledge state \( K^n \), which has just been updated with the most recent observation \( W^n(\omega_\pi^n) \). Thus, \( x_n \) implicitly depends on \( \omega_\pi^n \).

Let \( \omega \in \Omega \) be a sample realization using our original construction from Section 4.1, as illustrated in Table 4.1. For each \( \omega \in \Omega \), and given a measurement policy \( \pi \).
that determines the decision, we can construct a new sample realization of decisions $x^n(\omega^n)$ and observations $W_{n+1}^{n+1}(\omega^n_{n+1})$. Let $\omega^n = (\omega^n_1, \omega^n_2, ..., \omega^n_N)$ be a sample realization of both decisions and the measurements that go with the decision, as depicted in the sample path above. Given a policy $\pi$, we can construct a set of outcomes $\Omega^x$ where every outcome $\omega \in \Omega$ can be mapped to an outcome $\omega^x \in \Omega^x$. We would then write the probability of $\omega^x$ as $P^x(\omega^x)$.

If we want to find our objective function, we no longer have to embed the measurement policy $\pi$ in the objective function, since $\omega^x$ already contains the decisions that we want to make. In this case, we can write $F(\mu(\omega^n), W(\omega^n))$ since we no longer need to index $F(\cdot)$ by the policy $\pi$. Now if we wanted to take an expectation of $F(\mu, W)$ we would write

$$F^x = E^x F(\mu, W) = \sum_{\omega^x \in \Omega^x} P^x(\omega) F(\mu(\omega), \Omega(\omega^x)).$$

In the research literature, it is possible to see authors writing $E F^x(\mu, W)$ and $E^x F(\mu, W)$. The ultimate meaning is the same, but these two ways of writing the expectation implicitly represent different constructions of the underlying probability space.

4.4.2 Equivalence of Using True Means and Sample Estimates

We formally state the equivalence of the two approaches to policy evaluation from Section 4.3, and present an argument as to why it holds.

**Theorem 4.4.1** Let $\pi$ be a policy, and let $x^\pi$ be the alternative selected by the policy. Then,

$$E_{\mu^\pi} = E \max_x \theta^N_x.$$ 

**Proof:** Recall that $E^N_{\mu^\pi} = \theta^N_x$ for any fixed $x$. By the tower property of conditional expectations,

$$E_{\mu^\pi} = E E^N_{\mu^\pi} = E \theta^N_x,$$

because $x^\pi = \arg \max \theta^N_x$ is known at time $N$ (that is, it is fixed from the point of view of our time-$N$ beliefs). However, $\theta^N_x = \max \theta^N_x$ by definition of $x^\pi$. □

Theorem 4.4.1 has an interesting corollary. Denote by $\chi$ an “implementation policy” for selecting an alternative at time $N$. We can think of $\chi$ as a function mapping the knowledge state $K^N$ to an alternative $\chi(K^N) \in \{1, ..., M\}$. Then,

$$\max_{\chi} E_{\chi(K^N)} = \max_x \theta^N_x.$$ 

In other words, the optimal decision at time $N$ is always to select $\theta^N_x$. If we have no more opportunities to learn, the best possible decision we can make is to go with our final set of beliefs. This result addresses the issue of why most of our policies seek to maximize $\max_x \theta^N_x$ in some way, even though what we really want to learn is the unknown true value $\max_x \mu_x$. 


4.5 BIBLIOGRAPHIC NOTES

Section 4.1 - We present here a standard Bayesian framework for ranking and selection; see, e.g., Gupta & Miescke (1996) or Chick (2006). Our presentation is based on the measure-theoretic idea of random variables as functions on a space of outcomes or sample paths; although measure theory is far outside the scope of this book, interested readers are directed to Cinlar (2011), a definitive rigorous exposition of measure-theoretic probability.

Section 4.2 - The design of policies for taking observations (or measurements) of noisy functions has its roots in the 1950s and 1960s. It evolved originally under the umbrella of stochastic optimization over a continuous domain from the seminal paper of Robbins and Monro (Robbins & Monro 1951), but this literature did not focus on the issue of maximizing the information gained from each observation (there was more attention on asymptotic convergence than rate of convergence). The ranking and selection community evolved with a focus on the problem of finding the best out of a set of discrete alternatives; see Barr & Rizvi (1966) for an early review, and see Fu (2002) for a more current review. The challenge of collecting information in an optimal way has its roots in DeGroot (1970), which appears to give the first presentation of optimal learning as a dynamic program (but without an algorithm). Interval estimation was introduced by Kaelbling (1993). The adaptation of interval estimation using Chernoff bounds was done by Streeter & Smith (2006). Epsilon-greedy is described in Sutton & Barto (1998), with an analysis of convergence properties given in Singh et al. (2000). There is an emerging area of research which uses the concept of upper confidence bounding (UCB), originally developed for online problems, in an offline setting. We introduce the idea of upper confidence bounding in Chapter 6 for online ("bandit") problems. Drawing on this framework, Audibert et al. (2010) presents a frequentist approach to ranking and selection with finite alternatives, with provable guarantees.

PROBLEMS

4.1 We wish to find a good learning policy to solve the problem in Table 6.6.

a) Briefly describe the epsilon-greedy policy, the Boltzmann policy, and the interval-estimation policy. Evaluate each policy in terms of its ability to capture important characteristics of a good learning policy.

b) Define the expected opportunity cost (EOC). Describe in words how you would approximate the EOC (since computing it exactly is impossible) using Monte Carlo simulation.

c) Let $\omega^n$ be the index for the $nth$ sample realization of the random observations. Give an expression for the EOC for some policy $\pi$ and give a precise formula for the confidence interval for the true performance of a policy $\pi$. 
4.2 This exercise requires that you test an exploration policy in MATLAB. You will need to download two files from the course website:

http://optimallearning.princeton.edu/exercises/exploration.m
http://optimallearning.princeton.edu/exercises/explorationRun.m

The MATLAB file exploration.m executes a pure exploration policy for a general ranking and selection problem. The file explorationRun.m creates the data for a problem with 10 alternatives, where it simulates 1000 truths and 50 samples per truth. The program exploration.m computes the average opportunity cost “o_cost” and the standard deviation “se_result.”

a) Write out the meaning of “o_cost” mathematically using the notation we have been using in the course.

b) The standard deviation “se_result” is the standard deviation of what variable?

c) Construct a 95 percent confidence interval for the value of the exploration policy, and modify the code to produce this confidence interval.

4.3 In this exercise you have to implement the Boltzmann learning policy, which you should model after the exploration.m routine in the previous exercise.

a) Create a new file boltzmann.m which implements the Boltzmann learning policy. Keep in mind that you will have to introduce a tunable parameter \( \rho \). The Boltzmann policy gives you the probability that you should sample an alternative. Imagine you have three alternatives, and the Boltzmann distribution gives you sampling probabilities of .2, .5 and .3. To sample from this distribution, generate a random number between 0 and 1. If this number is less than or equal to .2, you choose the first alternative; if it is between .2 and .7, choose the second alternative; otherwise choose the third. You will need to generalize this for an arbitrary number of alternatives.

b) Using \( N = 5000 \), vary \( \rho \) over .001, .01, .1, 1.0, until it appears that you have found the range which bounds the best value of \( \rho \). For each value of \( \rho \) that you try, record it in a table and report the value of the policy and the standard error. Narrow your range until you begin getting results that are indistinguishable from each other.

c) Compare the performance of the best Boltzmann policy that you can find to a pure exploration policy that was developed in exercise 4.2.

4.4 Implement the interval estimation policy using the exploration.m routine provided in exercise 4.2.

a) Create a file called ieo.m which implements the interval estimation policy. Again, you will need a tunable parameter \( z_\alpha \) which you might call zalpha.
b) The optimal value of \(z_0\) will be in the range from 0 to 5. Search over this range, first in increments of 1.0, and then in smaller increments, until you again are unable to distinguish between values (continue using \(N = 5000\)).

c) Compare the performance of the best Boltzmann policy that you can find to a pure exploration policy that was developed in exercise 4.2.

4.5 This exercise builds on the exploration policy, Boltzmann policy and interval estimation policy that was implemented in exercises 4.2, 4.3, and 4.4.

a) Run each policy using \(N = 5000, M = 50\) and report the confidence intervals, using your best estimate of the tunable parameters. Can you conclude that one policy is better than the other two? If so, which one? If not, use the fact that the standard deviation declines inversely with \(\sqrt{N}\) (it also decreases inversely with \(\sqrt{M}\), but we are going to hold \(M\) constant for our study). Use this to determine how small you need the standard error to be, and then how large \(N\) needs to be to produce a confidence interval that is small enough to conclude that one policy is best. You have may to repeat this exercise a few times for the numbers to settle down.

b) Now set \(N = 1, M = 1\) and run each policy 10 times, reporting the actual outcome (noticed that you will not get a standard error in this case). This simulates the application of a learning policy in a specific situation, where you put it into practice (but we are going to pretend that you can replicate this 10 times). Record how often each policy discovers the best alternative (\(o_{\text{cost}} = 0\)). Comment on the likelihood that your best policy would outperform the other two policies on a single sample path.

4.6 You have to choose the best of three medications to treat a disease. The performance of each medication depends on the genetic characteristics of the individual. From the perspective of this medication, people can be divided into five genetic subgroups. If a doctor knew a patient’s subgroup, he would know the medication he should use, but this information is not available. Lacking this information, the doctor has to resort to trial and error. Complicating the process is that measuring the performance of a medication involves a certain level of noise.

Table 4.3 gives the average performance of each type of medication on the five patient types, based on extensive prior records. The five patient types occur with equal probability in the population. The doctor will typically test a medication on a patient for one month, after which a blood test provides a measure of the performance of the medication. But these measurements are not precise; the error between the measurement and the actual impact of the medication is normally distributed with a standard deviation of 2.2 (we assume this is constant for all medications and patient types).

a) What is the prior vector \(\theta^0\) that you would use for this problem, given the data in the table?
Table 4.3  The performance of each type of medication for each type of patient.

<table>
<thead>
<tr>
<th>Patient Type</th>
<th>Medication</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>M1</td>
<td>6.2</td>
<td>7.3</td>
<td>5.4</td>
<td>7.2</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>8.4</td>
<td>6.9</td>
<td>6.8</td>
<td>6.6</td>
<td>4.2</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>5.2</td>
<td>5.8</td>
<td>6.3</td>
<td>5.5</td>
<td>3.7</td>
<td></td>
</tr>
</tbody>
</table>

b) What is your prior probability distribution for the performance of medication M1? Note: It is not normally distributed.

c) Test each of the following policies below. You may use the MATLAB routines developed in the previous exercises, or perform the exercise manually with a budget of \( N = 10 \) observations. If you are using the MATLAB routines, use a budget of \( N = 50 \) observations.

1) Pure exploitation.
2) Boltzmann exploration, using scaling factor \( \rho = 1 \).
3) Epsilon-greedy exploration, where the exploration probability is given by \( 1/n \).
4) Interval estimation. Test the performance of \( z_\alpha = 1.0, 2.0, 3.0 \) and 4.0 and select the one that performs the best.

For each policy, report the average performance based on 100 trials, and compute a 95 percent confidence interval.