CHAPTER 1

THE CHALLENGES OF LEARNING

We are surrounded by situations where we need to make a decision or solve a problem, but where we do not know some or all of the relevant information for the problem perfectly. Will the path recommended by my navigation system get me to my appointment on time? Am I charging the right price for my product, and do I have the best set of features? Will a new material make batteries last longer? Will a molecular compound help reduce a cancer tumor? If I turn my retirement fund over to this investment manager, will I be able to outperform the market? Sometimes the decisions have a simple structure (which investment advisor should I use), while others require complex planning (how do I deploy a team of security agents to assess the safety of a set of food processing plants). Sometimes we have to learn while we are doing (the sales of a book at a particular price), while in other cases we may have a budget to collect information before making a final decision.

There are some decision problems that are hard even if we have access to perfectly accurate information about our environment: planning routes for aircraft and pilots, optimizing the movements of vehicles to pick up and deliver goods, or scheduling machines to finish a set of jobs on time. This is known as deterministic optimization. Then there are other situations where we have to make decisions under uncertainty, but where we assume we know the probability distributions of the uncertain quantities:

Optimal Learning. By Warren B. Powell and Ilya O. Ryzhov

How do I allocate investments to minimize risk while maintaining a satisfactory return, or how do I optimize the storage of energy given uncertainties about demands from consumers? This is known as stochastic optimization.

In this book, we introduce problems where the probability distributions themselves are unknown, but where we have the opportunity to collect new information to improve our understanding of what they are. We are primarily interested in problems where the cost of the information is considered "significant," which is to say that we are willing to spend some time thinking about how to collect the information in an effective way. What this means, however, is highly problem-dependent. We are willing to spend quite a bit before we drill a $10 million hole hoping to find oil, but we may be willing to invest only a small effort before determining the next measurement inside a search algorithm running on a computer.

The modeling of learning problems, which might be described as "learning how to learn," can be fairly difficult. While expectations are at least well-defined for stochastic optimization problems, they take on subtle interpretations when we are actively changing the underlying probability distributions. For this reason, we tend to work on what might otherwise look like very simple problems. Fortunately, there are very many "simple problems" which would be trivial if we only knew the values of all the parameters, but which pose unexpected challenges when we lack information.

1.1 LEARNING THE BEST PATH

Consider the problem of finding the fastest way to get from your new apartment to your new job in Manhattan. We can find a set of routes from the Internet or from our GPS device, but we do not know anything about traffic congestion or subway delays. The only way we can get data to estimate actual delays on a path is to travel the path. We wish to devise a strategy that governs how we choose paths so that we strike a balance between experimenting with new paths and getting to work on time every day.

Assume that our network is as depicted in Figure 1.1. Let p be a specific path, and let $x_p = 1$ if we choose to take path p. After we traverse the path, we observe a cost $\hat{c}_p$. Let μ_p denote the true mean value of $\hat{c}_p$, which is of course unknown to us. After n trials, we can compute a sample mean $\bar{\theta}_p^n$ of the cost of traversing path p along with a sample variance $\hat{\sigma}_p^{2,n}$ using our observations of path p. Of course, we only observe path p if $x_p^n = 1$, so we might compute these statistics using

$$N_p^n = \sum_{k=1}^{n} x_p^k, \tag{1.1}$$

$$\bar{\theta}_p^n = \frac{1}{N_p^n} \sum_{k=1}^{n} x_p^k \hat{c}_p^k, \tag{1.2}$$

$$\hat{\sigma}_p^{2,n} = \frac{1}{N_p^n - 1} \sum_{k=1}^{n} x_p^k (\hat{c}_p^k - \bar{\theta}_p^n)^2. \tag{1.3}$$

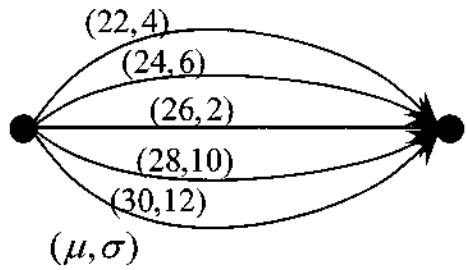


Figure 1.1 A simple shortest path problem, giving the current estimate of the mean and standard deviation (of the estimate) for each path.

Note that $\hat{\sigma}_p^{2,n}$ is our estimate of the variance of $\hat{c}_p$ by iteration n (assuming we have visited path p $N_p^n > 1$ times). The variance of our estimate of the mean, $\bar{\theta}_p^n$, is given by

$$\bar{\sigma}_p^{2,n} = \frac{1}{N_p^n}\hat{\sigma}_p^{2,n}.$$

Now we face the challenge: Which path should we try? Let's start by assuming that you just started a new job and you have been to the Internet to find different paths, but you have not tried any of them. If your job involves commuting from a New Jersey suburb into Manhattan, you have a mixture of options that include driving (various routes) and commuter train, with different transit options once you arrive in Manhattan. But you do have an idea of the length of each path, and you may have heard some stories about delays through the tunnel into Manhattan, as well as a few stories about delayed trains. From this, you construct a rough estimate of the travel time on each path, and we are going to assume that you have at least a rough idea of how far off these estimates may be. We denote these initial estimates using

$\bar{\theta}_p^0$ = initial estimate of the expected travel time on path p,

$\bar{\sigma}_p^0$ = initial estimate of the standard deviation of the difference between $\bar{\theta}_p^0$ and the truth.

If we believe that our estimation errors are normally distributed, then we think that the true mean, μ_p, is in the interval $(\mu_p - z_{\alpha/2}\bar{\sigma}_p^0, \mu_p + z_{\alpha/2}\bar{\sigma}_p^0)$ α percent of the time. If we assume that our errors are normally distributed, we would say that we have an estimate of μ_p that is normally distributed with parameters $(\bar{\theta}_p^0, (\bar{\sigma}_p^0)^2)$.

So which path do you try first? If our priors are as shown in Figure 1.1, presumably we would go with the first path, since it has a mean path time of 22 minutes, which is less than any of the other paths. But our standard deviation around this belief is 4, which means we believe this could possibly be as high as 30. At the same time, there are paths with times of 28 and 30 with standard deviations of 10 and 12. This means

that we believe that these paths could have times that are even smaller than 20. Do we always go with the path that we think is the shortest? Or do we try paths that we think are longer, but where we are just not sure, and there is a chance that these paths may actually be better?

If we choose a path we think is best, we say that we are *exploiting* the information we have. If we try a path because it might be better, which would help us make better decisions in the future, we say that we are *exploring*. Exploring a new path, we may find that it is an unexpectedly superior option, but it is also possible that we will simply confirm what we already believed. We may even obtain misleading results – it may be that this one route was experiencing unusual delays on the one day we happened to choose it. Nonetheless, it is often desirable to try something new to avoid becoming stuck on a suboptimal solution just because it "seems" good. Balancing the desire to explore versus exploit is referred to in some communities as the *exploration versus exploitation* problem. Another name is the *learn versus earn* problem. Regardless of the name, the point is the lack of information when we make a decision, along with the value of new information in improving future decisions.

1.2 AREAS OF APPLICATION

The diversity of problems where we have to address information acquisition and learning is tremendous. Below, we try to provide a hint of the diversity.

Transportation

- Responding to disruptions - Imagine that there has been a disruption to a network (such as a bridge failure) forcing people to go through a process of discovering new travel routes. This problem is typically complicated by noisy observations and by travel delays that depend not just on the path but also on the time of departure. People have to evaluate paths by actually traveling them.

- Revenue management - Providers of transportation need to set a price that maximizes revenue (or profit), but since demand functions are unknown, it is often necessary to do a certain amount of trial and error.

- Evaluating airline passengers or cargo for dangerous items - Examining people or cargo to evaluate risk can be time-consuming. There are different policies that can be used to determine who/what should be subjected to varying degrees of examination. Finding the best policy requires testing them in field settings.

- Finding the best heuristic to solve a difficult integer program for routing and scheduling - We may want to find the best set of parameters to use our tabu search heuristic, or perhaps we want to compare tabu search, genetic algorithms, and integer programming for a particular problem. We have to loop over different algorithms (or variations of an algorithm) to find the one that works the best on a particular dataset.

Figure 1.2 The operations center for NetJets®, which manages over 750 aircraft[1]. NetJets® has to test different policies to strike the right balance of costs and service.

- Finding the best business rules - A transportation company needs to determine the best terms for serving customers, the best mix of aircraft, and the right pilots to hire[1] (see Figure 1.2). They may use a computer simulator to evaluate these options, requiring time-consuming simulations to be run to evaluate different strategies.
- Evaluating schedule disruptions - Some customers may unexpectedly ask us to deliver their cargo at a different time, or to a different location than what was originally agreed upon. Such disruptions come at a cost to us, because we may need to make significant changes to our routes and schedules. However, the customers may be willing to pay extra money for the disruption. We have a limited time to find the disruption or combination of disruptions where we can make the most profit.

Energy and the Environment

- Finding locations for wind farms - Wind conditions can depend on micro-geography - a cliff, a local valley, a body of water. It is necessary to send teams with sensors to find the best locations for locating wind turbines in a geographical area. The problem is complicated by variations in wind, making it necessary to visit a location multiple times.
- Finding the best material for a solar panel - It is necessary to test large numbers of molecular compounds to find new materials for converting sunlight to electricity. Testing and evaluating materials is time consuming and very expensive, and there are large numbers of molecular combinations that can be tested.

[1]Includes aircraft under management by Executive Jet® Management.

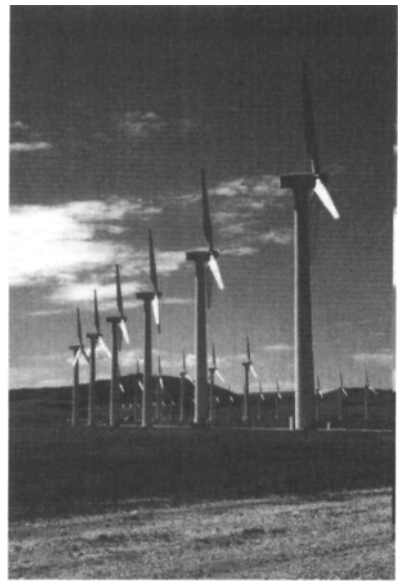


Figure 1.3 Wind turbines are one form of alternative energy resources (from http://www.nrel.gov/data/pix/searchpix.cgi).

- Tuning parameters for a fuel cell - There are a number of design parameters that have to be chosen to get the best results from a full cell: the power density of the anode or cathode, the conductivity of bipolar plates, and the stability of the seal.

- Finding the best energy-saving technologies for a building - Insulation, tinted windows, motion sensors and automated thermostats interact in a way that is unique to each building. It is necessary to test different combinations to determine the technologies that work the best.

- R&D strategies - There are a vast number of research efforts being devoted to competing technologies (materials for solar panels, biomass fuels, wind turbine designs) which represent projects to collect information about the potential for different designs for solving a particular problem. We have to solve these engineering problems as quickly as possible, but testing different engineering designs is time-consuming and expensive.

- Optimizing the best policy for storing energy in a battery - A policy is defined by one or more parameters that determine how much energy is stored and in what type of storage device. One example might be, "charge the battery when the spot price of energy drops below x." We can collect information in the field or a computer simulation that evaluates the performance of a policy over a period of time.

- Learning how lake pollution due to fertilizer run-off responds to farm policies - We can introduce new policies that encourage or discourage the use of fertilizer, but we do not fully understand the relationship between these policies and lake pollution, and these policies impose different costs on the farmers. We need to test different policies to learn their impact, but each test requires a year to run and there is some uncertainty in evaluating the results.

- On a larger scale, we need to identify the best policies for controlling CO^2 emissions, striking a balance between the cost of these policies (tax incentives on renewables, a carbon tax, research and development costs in new technologies) and the impact on global warming, but we do not know the exact relationship between atmospheric CO^2 and global temperatures.

Homeland Security

- You would like to minimize the time to respond to an emergency over a congested urban network. You can take measurements to improve your understanding of the time to traverse each region of the traffic network, but collecting these observations takes time. How should you structure your observations of links in the network to achieve the best time when you need to find the shortest path?

- You need to manage a group of inspectors to intercept potentially dangerous cargo being smuggled through ports and across borders. Since you do not know the frequency with which smugglers might try to use a port of entry, it is important to allocate inspectors not just to maximize the likelihood of an interception given current beliefs, but to also collect information so that we can improve our understanding of the truth. For example, we may believe that a particular entry point might have a low probability of being used, but we may be wrong.

- Radiation is detected in downtown Manhattan. Inspectors have to be managed around the city to find the source as quickly as possible. Where should we send them to maximize the likelihood of finding the source?

Science and Engineering

- The National Ignition Facility uses large crystals to focus lasers into a very small region to perform nuclear research. The crystals become damaged over time and have to be repaired or replaced, but the process of examining each crystal is time-consuming and reduces the productivity of the facility. NIF has to decide when to examine a crystal to determine its status.

- A company is trying to design an aerosol device whose performance is determined by a number of engineering parameters: the diameter of the tube that pulls liquid from a reservoir, the pressure, the angle of a plate used to direct the spray, and the size of the portal used to project the spray and the angle of

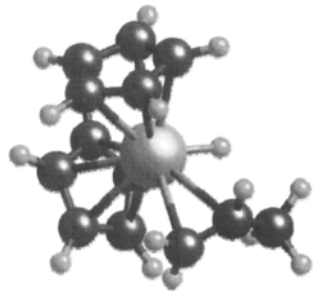


Figure 1.4 Drug discovery requires testing large numbers of molecules.

the departure portal. These have to be varied simultaneously to find the best design.

Health and Medicine

- Drug discovery - Curing a disease often involves first finding a small family of base molecules, and then testing a large number of variations of a base molecule. Each test of a molecular variation can take a day and consumes costly materials, and the performance can be uncertain.
- Drug dosage - Each person responds to medication in a different way. It is often necessary to test different dosages of a medication to find the level that produces the best mix of effectiveness against a condition with minimum side effects.
- How should a doctor test different medications to treat diabetes, given that he will not know in advance how a particular patient might respond to each possible course of treatment?
- What is the best way to test a population for an emerging disease so that we can plan a response strategy?

Sports

- How do you find the best set of five basketball players to use as your starting lineup? Basketball players require complementary skills in defense, passing, and shooting, and it is necessary to try different combinations of players to see which group works the best.
- What is the best combination of rowers for a four person rowing shell? Rowers require a certain synergy to work well together, making it necessary to try different combinations of rowers to see who turns in the best time.

- Who are the best hitters that you should choose for your baseball team? It is necessary to see how a player hits in game situations, and of course these are very noisy observations.

- What plays work the best for your football team? Specific plays draw on different combinations of talents, and a coach has to find out what works best for his team.

Business

- What are the best labor rules or terms in a customer contract to maximize profits? These can be tested in a computer simulation program, but it may require several hours (in some cases, several days) to run. How do we sequence our experiments to find the best rules as quickly as possible?

- What is the best price to charge for a product being sold over the Internet? It is necessary to use a certain amount of trial and error to find the price that maximizes revenue.

- We would like to find the best supplier for a component part. We know the price of the component, but we do not know about the reliability of the service or the quality of the product. We can collect information on service and product quality by placing small orders.

- We need to identify the best set of features to include in a new laptop we are manufacturing. We can estimate consumer response by running market tests, but these are time-consuming and delay the product launch.

- A company needs to identify the best person to lead a division that is selling a new product. The company does not have time to interview all the candidates. How should a company identify a subset of potential candidates?

- Advertising for a new release of a movie - We can choose between TV ads, billboards, trailers on movies already showing, the Internet, and promotions through restaurant chains. What works best? Does it help to do TV ads if you are also doing Internet advertising? How do different outlets interact? You have to try different combinations, evaluate their performance, and use what you learn to guide future advertising strategies.

- Conference call or airline trip? Business people have to decide when to try to land a sale using teleconferencing, or when a personal visit is necessary. For companies that depend on numerous contacts, it is possible to experiment with different methods of landing a sale, but these experiments are potentially expensive, involving (a) the time and expense of a personal trip or (b) the risk of not landing a sale.

E-Commerce

- Which ads will produce the best consumer response when posted on a website? You need to test different ads, and then identify the ads that are the most promising based on the attributes of each ad.
- Netflix can display a small number of movies to you when you log into your account. The challenge is identifying the movies that are likely to be most interesting to a particular user. As new users sign up, Netflix has to learn as quickly as possible which types of movies are most likely to attract the attention of an individual user.
- You need to choose keywords to bid on to get Google to display your ad. What bid should you make for a particular keyword? You measure your performance by the number of clicks that you receive.
- YouTube has to decide which videos to feature on its website to maximize the number of times a video is viewed. The decision is the choice of video, and the information (and reward) is the number of times people click on the video.
- Amazon uses your past history of book purchases to make suggestions for potential new purchases. Which products should be suggested? How can Amazon use your response to past suggestions to guide new suggestions?

The Service Sector

- A university has to make specific offers of admission, after which it then observes which types of students actually matriculate. The university has to actually make an offer of admission to learn whether a student is willing to accept the offer. This information can be used to guide future offers in subsequent years. There is a hard constraint on total admissions.
- A political candidate has to decide in which states to invest his remaining time for campaigning. He decides which states would benefit the most through telephone polls, but has to allocate a fixed budget for polling. How should he allocate his polling budget?
- The Federal government would like to understand the risks associated with issuing small business loans based on the attributes of an applicant. A particular applicant might not look attractive, but it is possible that the government's estimate of risk is inflated. The only way to learn more is to try granting some higher risk loans.
- The Internal Revenue Service has to decide which companies to subject to a tax audit. Should it be smaller companies or larger ones? Are some industries more aggressive than others (for example, due to the presence of lucrative tax write-offs)? The government's estimates of the likelihood of tax cheating may be incorrect, and the only way to improve its estimates is to conduct audits.

Figure 1.5 The Air Force has to design new technologies and determine the best policies for operating them.

The Military

- The military has to collect information on risks faced in a region using UAVs (unmanned aerial vehicles). The UAV collects information about a section of road, and then command determines how to deploy troops and equipment. How should the UAVs be deployed to produce the best deployment strategy?
- A fighter has to decide at what range to launch a missile. After firing a missile, we learn whether the missile hit its target or not, which can be related to factors such as range, weather, altitude and angle-of-attack. With each firing, the fighter learns more about the probability of success.
- The Air Force has to deploy tankers for mid-air refueling. There are different policies for handling the tankers, which include options such as shuttling tankers back and forth between locations, using one tanker to refuel another tanker, and trying different locations for tankers. A deployment policy can be evaluated by measuring (a) how much time fighters spend waiting for refueling and (b) the number of times a fighter has to abort a mission from lack of fuel.
- The military has to decide how to equip a soldier. There is always a tradeoff between cost and the weight of the equipment, versus the likelihood that the soldier will survive. The military can experiment with different combinations of equipment to assess its effectiveness in terms of keeping a soldier alive.

Tuning Models and Algorithms

- There is a large community that models physical problems such as manufacturing systems using Monte Carlo simulation. For example, we may wish to simulate the manufacture of integrated circuits which have to progress through a series of stations. The progression from one station to another may be limited by the size of buffers which hold circuit boards waiting for a particular machine.

We wish to determine the best size of these buffers, but we have to do this by sequential simulations which are time-consuming and noisy.

- There are many problems in discrete optimization where we have to route people and equipment, or scheduling jobs to be served by a machine. These are exceptionally hard optimization problems that are typically solved using heuristic algorithms such as tabu search or genetic algorithms. These algorithms are controlled by a series of parameters which have to be tuned for specific problem classes. One run of an algorithm on a large problem can require several minutes to several hours (or more), and we have to find the best setting for perhaps five or ten parameters.

- Engineering models often have to be calibrated to replicate a physical process such as weather or the spread of a chemical through groundwater. These models can be especially expensive to run, often requiring the use of fast supercomputers to simulate the process in continuous space or time. At the same time, it is necessary to calibrate these models to produce the best possible prediction.

1.3 MAJOR PROBLEM CLASSES

Given the diversity of learning problems, it is useful to organize these problems into major problem classes. A brief summary of some of the major dimensions of learning problems is given below.

- Online versus offline - Online problems involve learning from experiences as they occur. For example, we might observe the time on a path through a network by traveling the path, or adjust the price of a product on the Internet and observe the revenue. We can try a decision that looks bad in the hopes of learning something, but we have to incur the cost of the decision, and balance this cost against future benefits. In offline problems, we might be working in a lab with a budget for making measurements, or we might set aside several weeks to run computer simulations. If we experiment with a chemical or process that does not appear promising, all we care about is the information learned from the experiment; we do not incur any cost from running an unsuccessful experiment. When our budget has been exhausted, we have to use our observations to choose a design or a process that will then be put into production.

- Objectives - Problems differ in terms of what we are trying to achieve. Most of the time we will focus on minimizing the expected cost or maximizing the expected reward from some system. However, we may be simply interested in finding the best design, or ensuring that we find a design that is within five percent of the best.

- The measurement decision - In some settings, we have a small number of choices such as drilling test wells to learn about the potential for oil or natural gas. The number of choices may be small, but each test can cost millions of

dollars. Alternatively, we might have to find the best set of 30 proposals out of 100 that have been submitted, which means that we have to choose from 3×10^{25} possible portfolios. Or we may have to choose the best price, temperature, or pressure (a scalar, continuous parameter). We might have to set a combination of 16 parameters to produce the best results for a business simulator. Each of these problems introduce different computational challenges because of the size of the search space.

- The implementation decision - Collecting the best information depends on what you are going to do with the information once you have it. Often, the choices of what to observe (the measurement decision) are the same as what you are going to implement (finding the choice with the best value). But you might measure a link in a graph in order to choose the best path. Or we might want to learn something about a new material to make a decision about new solar panels or batteries. In these problems, the implementation decision (the choice of path or technology) is different from the choice of what to measure.

- What we believe - We may start by knowing nothing about the best system. Typically, we know something (or at least we will know something after we make our first measurement). What assumptions can we reasonably make about different choices? Can we put a normal distribution of belief on an unknown quantity? Are the beliefs correlated (if a laptop with one set of features has higher sales than we expected, does this change our belief about other sets of features)? Are the beliefs stored as a lookup table (that is, a belief for each design), or are the beliefs expressed as some sort of statistical model?

- The nature of a measurement - Closely related to what we believe is what we learn when we make a measurement. Is the observation normally distributed? Is it a binary random variable (success/failure)? Are measurements made with perfect accuracy? If not, do we know the distribution of the error in a measurement?

- Belief states and physical states - All learning problems include a "belief state" (or knowledge state) which captures what we believe about the system. Some problems also include a physical state. For example, to measure the presence of disease at city i, we have to visit city i. After making this measurement, the cost of visiting city j now depends on city i. Our physical location is a physical state.

We are not going to be able to solve all these problems in this book, but we can at least recognize the diversity of problems.

1.4 THE DIFFERENT TYPES OF LEARNING

It is useful to contrast learning problems with other types of optimization problems. Figure 1.1 depicts two optimization problem. The problem in Figure 1.1(a) shows

Table 1.1 (a) A problem involving five known alternatives, and (b) a problem where the value of each alternative is normally distributed with known mean and standard deviation.

Alternative	Value
1	759
2	722
3	698
4	653
5	616

(a) The Best of Five Known Alternatives

Alternative	Mean	Std. Dev.
1	759	120
2	722	142
3	698	133
4	653	90
5	616	102

(b) The Best of Five Uncertain Alternatives

five choices, each of which has a known value. The best choice is obviously the first one, with a value of 759. Of course, deterministic optimization problems can be quite hard, but this happens to be a trivial one.

A harder class of optimization problems arise when there is uncertainty in the parameters. Figure 1.1(b) depicts a problem with five choices where the reward we receive from a choice is normally distributed with known mean and standard deviation. Assume that we have to make a choice before the reward is received, and we want to make a choice that gives us the highest expected return. Again, we would select the first alternative, because it has the highest expected value.

The problems illustrated in Table 1.1 use either known values, or known distributions. This problem is fairly trivial (picking the best out of a list of five), but there are many problems in stochastic optimization that are quite hard. In all of these problems, there are uncertain quantities but we assume that we know the probability distribution describing the likelihood of different outcomes. Since the distributions are assumed known, when we observe an outcome we view it simply as a realization from a known probability distribution. We do not use the observation to update our belief about the probability distribution.

Now consider what happens when you are not only uncertain about the reward, you are uncertain about the probability distribution for the reward. The situation is illustrated in Table 1.2, where after choosing to measure the first alternative, we observe an outcome of 702 and then use this outcome to update our belief about the first alternative. Before our measurement, we thought the reward was normally distributed with mean 759 and standard deviation 102. After the measurement, we now believe the mean is 712 with standard deviation of 92. As a result, alternative 2 now seems to be the best.

Since we are willing to change our belief about an alternative, is it necessarily the case that we should try to evaluate what appears to be the best alternative? Later in this volume, we are going to refer to this as an *exploitation policy*. This means that we exploit our current state of knowledge and choose the alternative that appears to be best. But it might be the case that if we observe an alternative that does not appear

Table 1.2 Learning where we update our beliefs based on observations, which changes our distribution of belief for future measurements.

	Initial Mean and Std. Dev.		First	Updated Mean and Std. Dev.		Second
Alternative	Mean	Std. Dev.	Obs.	Mean	Std. Dev.	Obs.
1	759	102	702	712	92	
2	722	133		722	133	734
3	698	78		698	78	
4	653	90		653	90	
5	616	102		616	102	

to be the best to use right now, we may collect information that allows us to make better decisions in the future. The central idea of optimal learning is to incorporate the value of information in the future to make better decisions now.

Now consider another popular optimization problem known as the newsvendor problem. In this problem, we wish to order a quantity (of newspapers, oil, money, energy) x to satisfy a random demand D (that is, D is not known when we have to choose x). We earn p dollars per unit of satisfied demand, which is to say $\min(x, D)$, and we have to pay c dollars per unit of x that we order. The total profit is given by

$$F(x, D) = p \min(x, D) - cx.$$

The optimization problem is to solve

$$\max_x \mathbb{E}F(x, D).$$

There are a number of ways to solve stochastic optimization problems such as this. If the distribution of D is known, we can characterize the optimal solution using

$$P_D[x^* \leq D] = \frac{c}{p},$$

where $P_D()$ is the cumulative distribution function for D. So, as the purchase cost c is decreased, we should increase our order quantity so that the probability that the order quantity is less than demand also decreases.

In many applications, we do not know the distribution of D, but we are able to make observations of D (or we can observe if we have ordered too much or too little). Let x^{n-1} be the order quantity we chose after observing D^{n-1}, which was our best guess of the right order quantity to meet the demand on day n, and let D^n be resulting demand. Now let g^n be the derivative of $F(x, D)$, given that we ordered x^{n-1} and then observed D^n. This derivative is given by

$$g^n = \begin{cases} p - c & \text{if } x \leq D, \\ -c & \text{if } x > D. \end{cases}$$

A simple method for choosing x^n is a stochastic gradient algorithm which looks like

$$x^n = x^{n-1} + \alpha_{n-1} g^n. \tag{1.4}$$

Here, α_{n-1} is a stepsize that has to satisfy certain conditions that are not important here. If the stepsize is chosen appropriately, it is possible to show that in the limit, x^n approaches the optimal solution, even without knowing the distribution of D in advance.

What our algorithm in equation (1.4) ignores is that our choice of x^n allows us to learn something about the distribution of D. For example, it might be that the purchase cost c is fairly high compared to the sales price p, which would encourage us to choose smaller values of x, where we frequently do not satisfy demand. But we might benefit from making some larger orders just to learn more about the rest of the demand distribution. By ignoring our ability to learn, the algorithm may not converge to the right solution, or it may eventually find the right solution, but very slowly. When we use optimal learning, we explicitly capture the value of the information we learn now on future decisions.

1.5 LEARNING FROM DIFFERENT COMMUNITIES

The challenge of efficiently collecting information is one that arises in a number of communities. The result is a lot of parallel discovery, although the questions and computational challenges posed by different communities can be quite different, and this has produced diversity in the strategies proposed for solving these problems. Below we provide a rough list of some of the communities that have become involved in this area.

- Simulation optimization - The simulation community often faces the problem of tuning parameters that influence the performance of a system that we are analyzing using Monte Carlo simulation. These parameters might be the size of a buffer for a manufacturing simulator, the location of ambulances and fire trucks, or the number of advance bookings for a fare class for an airline. Simulations can be time-consuming, so the challenge is deciding how long to analyze a particular configuration or policy before switching to another one.

- The ranking and selection problem - This is a statistical problem that arises in many settings, including the simulation optimization community. It is most often approached using the language of classical frequentist statistics (but not always) and tends to be very practical in its orientation. In ranking and selection, we assume that for each measurement, we can choose equally from a set of alternatives (there is no cost for switching from one alternative to another). Although the ranking and selection framework is widely used in simulation optimization, the simulation community recognizes that it is easier to run the simulation for one configuration a little longer than it is to switch to the simulation of a new configuration.

- The bandit community - There is a subcommunity that has evolved within applied probability and machine learning that studies what has long been referred to as bandit problems. This is the online (pay as you go) version of ranking and selection. A major breakthrough for this problem class was the discovery that a simple index policy (a quantity computed for each alternative that guides which alternative should be tested next) is optimal, producing a line of research (primarily in applied probability) aimed at discovering optimal index policies for more general problems. A separate subcommunity (primarily in computer science) has focused on a simple heuristic known as *upper confidence bounding* which has the property that the number of times we test the wrong alternative is bounded by a logarithmic function, which has then been shown to be the best possible bound. Upper confidence bounding has also been popular in the control theory community.

- Global optimization of expensive functions - The engineering community often finds a need to optimize complex functions of continuous variables. The function is sometimes a complex piece of computer software that takes a long time to run, but the roots of the field come from geospatial applications. The function might be deterministic (but not always), and a single evaluation can take an hour to a week or more.

- Learning in economics - Economists have long studied the value of information in a variety of idealized settings. This community tends to focus on insights into the economic value of information, rather than the derivation of specific procedures for solving information collection problems.

- Active learning in computer science - The machine learning community typically assumes that a dataset is given. When there is an opportunity to choose what to measure, this is known as active learning. This community tends to focus on statistical measures of fit rather than economic measures of performance.

- Statistical design of experiments - A classical problem in statistics is deciding what experiments to run. For certain objective functions, it has long been known that experiments can be designed deterministically, in advance, rather than sequentially. Our focus is primarily on sequential information collection, but there are important problem classes where this is not necessary.

- Frequentist versus Bayesian communities - It is difficult to discuss research in optimal learning without addressing the sometimes contentious differences in styles and attitudes between frequentist and Bayesian statisticians. Frequentists look for the truth using nothing more than the data that we collect, while Bayesians would like to allow us to integrate expert judgment.

- Optimal stopping - There is a special problem class where we have the ability to observe a single stream of information such as the price of an asset. As long as we hold the asset, we get to observe the price. At some point, we have to make a

decision whether we should sell the asset or continue to observe prices (a form of learning). Another variant is famously known as the "secretary problem" where we interview candidates for a position (or offers for an asset); after each candidate (or offer) we have to decide if we should accept and stop or reject and continue observing.

- Approximate dynamic programming/reinforcement learning - Approximate dynamic programming, widely known as reinforcement learning in the computer science community, addresses the problem of choosing an action given a state which generates a reward and takes us to a new state. We do not know the exact value of the downstream state, but we might decide to visit a state just to learn more about it. This is generally known as the "exploration versus exploitation" problem, and this setting has motivated a considerable amount of research in optimal learning.

- Psychology - Not surprisingly, the tradeoff between exploration and exploitation is a problem that has to be solved by people (as well as other animals ranging from chimpanzees to ants) for problems ranging from finding foot to finding mates. This has recently attracted attention in the psychology community (Cohen et al. 2007).

Readers who wish to study this field seriously will encounter the contributions of these (and perhaps other) communities. It is not possible to cover all the issues and perspectives of these communities in this volume, but we do provide a foundation that should make it possible for students and researchers to understand the issues and, in some cases, challenge conventional wisdom within specific communities.

1.6 INFORMATION COLLECTION USING DECISION TREES

The simplest types of information collection problems arise when there is a small number of choices to collect information. Should you check the weather report before scheduling a baseball game? Should you purchase an analysis of geologic formulations before drilling for oil? Should you do a statistical analysis of a stock price before investing in the stock?

These are fairly simple problems that can be analyzed using a decision tree, which is a device that works well when the number of decisions, as well as the number of possible outcomes, is small and discrete. We begin by first presenting a small decision tree where collecting information is not an issue.

1.6.1 A Basic Decision Tree

Decision trees are a popular device for solving problems that involve making decisions under uncertainty, because they illustrate the sequencing of decisions and information so clearly. Figure 1.6 illustrates the decision tree that we might construct to help with the decision of whether to hold or sell a stock. In this figure, square nodes are decision

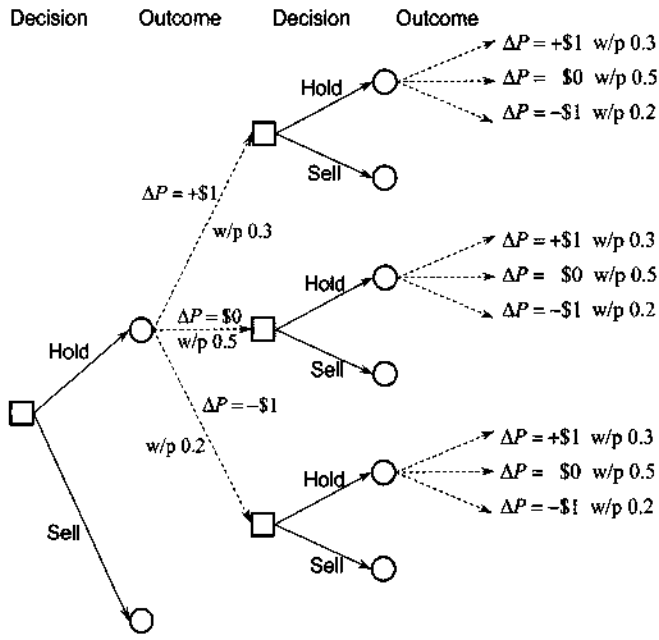


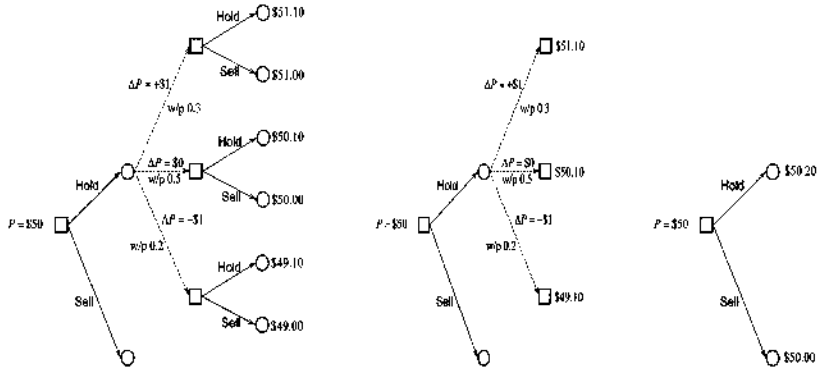
Figure 1.6 Illustration of a decision tree to determine if we should hold or sell a stock, where the stock might go up or down $1 or stay the same in each time period.

nodes, where we get to choose from a set of actions (where the number of actions cannot be too large). Circle nodes are outcome nodes, where something random happens, such as the change in the price of the stock. The solid lines represent decisions, while dashed lines are random outcomes. In our example, there is no cost for holding a stock, and the random outcome represents a change in the price of the stock.

If we hold the stock (which currently can be sold at a price of $50), it might go up or down by a dollar, with probabilities of 0.3 and 0.2, respectively, or hold at the same price with a probability of 0.5. After observing the change in the price, we again have a chance to hold or sell the stock. If we continue holding, the stock might go up or down by a dollar, or stay the same.

We can solve the problem of whether to hold or sell the stock initially by doing what is called "rolling back the decision tree." Figure 1.7(a) shows the tree after the first rollback. Here, we have taken the final random outcome and replaced it with the expected value, which gives us the result that we expect the price to go up by $0.10 if we hold the stock. We now have the option of holding or selling, which is a decision that we control. Since the price is likely to go up if we hold, we make this choice.

In Figure 1.7(b), we now use the expected value of the second decision to give us what we will earn at the end of each random outcome resulting from the first decision.



1.7(a) 1.7(b) 1.7(c)

Figure 1.7 (a) Decision tree with second outcome replaced by expected value. (b) Decision tree with second decision replaced by best expected value. (c) Decision tree with first outcome replaced by expected value, producing a deterministic decision.

We have the same three outcomes (up or down $1, or stay the same), but each outcome produces an expected return of $51.10, $50.10, or $49.10. Now we again have to find the expectation over these returns, which gives us the expected value of $50.20. Finally, we have to evaluate our original decision to hold or sell, and of course we are willing to hold for $50.20 rather than sell now for $50.

1.6.2 Decision Tree for Offline Learning

The previous example provided a quick illustration of a basic decision tree. A common decision problem is whether or not we should collect information to make a decision. For example, consider a bank that is considering whether it should grant a short-term credit loan of $100,000. The bank expects to make $10,000 if the loan is paid off on time. If the loan is defaulted, the bank loses the amount of the loan.

From history, the bank knows that 95 percent of loans are repaid in full, while 5 percent default. If the bank purchases the credit report, this information will allow the bank to classify the customer into one of three groups: 52 percent fall into the top A rating, 30 percent fall into the middle B rating, while 18 percent fall into the lower C rating with the highest risk of default. The company selling the credit report provides the joint distribution $P(\text{Credit}, \text{Default})$ that a customer will receive each credit rating, and whether it defaulted or not. This data are summarized in Table 1.3.

We need to understand how the information from the credit report changes our belief about the probability of a default. For this, we use a simple application of

Table 1.3 The marginal probability of each credit rating, the joint probability of a credit rating and whether someone defaults on a loan, and the conditional probability of a default given a credit rating.

		P(Credit,Default)		P(Default | Credit)	
Credit Rating	P(Credit)	No	Yes	No	Yes
A	0.52	0.51	0.01	0.981	0.019
B	0.30	0.28	0.02	0.933	0.067
C	0.18	0.16	0.02	0.889	0.111
	P(Default) =	0.95	0.05		

Bayes' theorem, which states

$$P(\text{Default} \mid \text{Credit}) = \frac{P(\text{Credit} \mid \text{Default})P(\text{Default})}{P(\text{Credit})} = \frac{P(\text{Credit}, \text{Default})}{P(\text{Credit})}.$$

Bayes' theorem allows us to start with our initial estimate of the probability of a default, $P(\text{Default})$, and then use the information "Credit" from the credit history and turn it into a *posterior* distribution $P(\text{Default} \mid \text{Credit})$. The results of this calculation are shown in the final two columns of Table 1.3.

Using this information, we can construct a new decision tree, shown in Figure 1.8. Unlike our first decision tree in Figure 1.7, we now see that the decision to collect information changes the downstream probabilities.

We repeat the exercise of rolling back the decision tree in Figure 1.9. Figure 1.9(a) shows the expected value of the decision to grant the loan given the information about the credit history. We see that if the grantee has an A or B credit rating, it makes sense to grant the loan, but not if the rating is C. Thus, the information from the credit report has the effect of changing the decision of whether or not to grant the loan. After we roll the tree back to the original decision of whether to purchase the credit report, we find that the credit report produces an expected value of \$4,900, compared to \$4,500 that we would expect to receive without the credit report. This means that we would be willing to pay up to \$400 for the credit report.

1.6.3 Decision Tree for Online Learning

Now consider a problem where we learn as we go. We use the setting of trying to identify the best hitter on a baseball team. The only way to collect information is to put the hitter into the lineup. Assume that we are evaluating hitters for the fourth position in the lineup, typically reserved for power hitters. Part of what we are trying to learn is how a hitter actually performs in game situations.

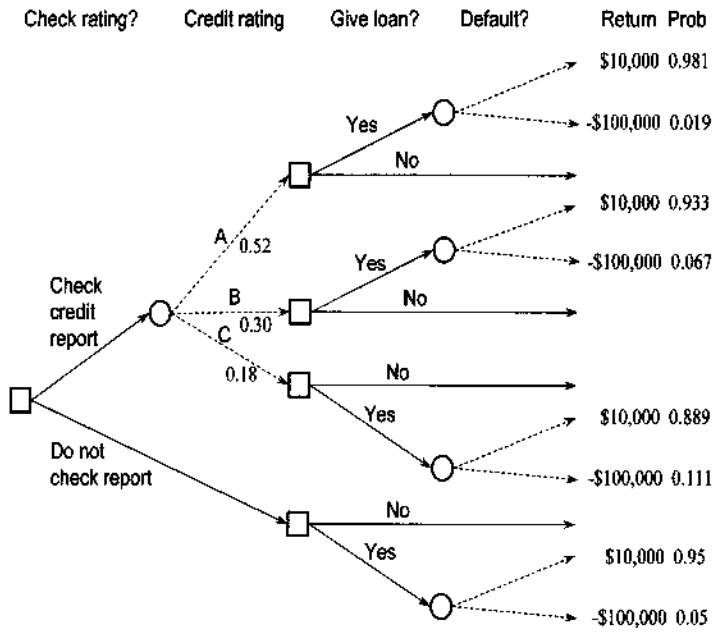


Figure 1.8 The decision tree for the question of whether we should purchase a credit risk report.

Assume that we have three candidates for the position. The information we have on each hitter from previous games is given in Table 1.4. If we choose player A, we have to balance the likelihood of getting a hit, and the value of the information we gain about his true hitting ability, since we will use the event of whether or not he gets a hit to update our assessment of his probability of getting a hit. We are going to again use Bayes' theorem to update our belief about the probability of getting a hit. To do this, we have to make some probabilistic assumptions that are not relevant to our discussion here; we defer until Chapter 2, Section 2.3.4, the discussion of the model that we use to calculate the updated probabilities. Fortunately, this model produces

Table 1.4 History of hitting performance for three candidates.

Player	No. Hits	No. At-Bats	Average
A	36	100	0.360
B	1	3	0.333
C	7	22	0.318

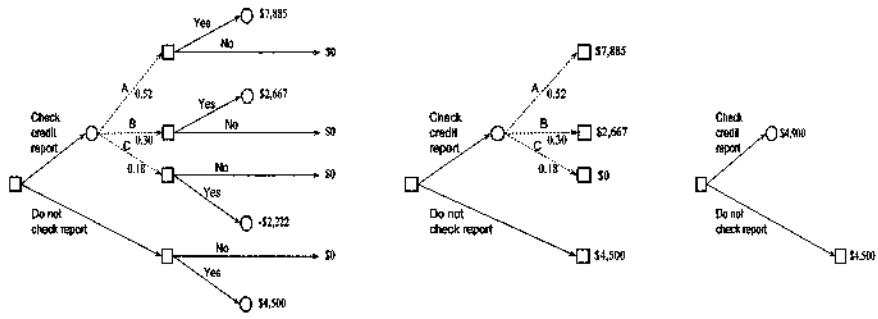


Figure 1.9 (a) Decision tree with final default replaced with expected value. (b) Decision tree with second decision replaced by best expected value. (c) Decision tree with the uncertainty of the credit risk report replaced with its expected value.

some very intuitive updating equations. Let H^n be number of hits a player has made in n at-bats. Let $\hat{H}^{n+1} = 1$ if a hitter gets a hit in his $(n+1)$st at-bat. Our prior probability of getting a hit after n at-bats is

$$\mathbb{P}[\hat{H}^{n+1} = 1 | H^n, n] = \frac{H^n}{n}.$$

Once we observe $\hat{H}^{n+1}$, it is possible to show that the posterior probability is

$$\mathbb{P}[\hat{H}^{n+2} = 1 | H^n, n, \hat{H}^{n+1}] = \frac{H^n + \hat{H}^{n+1}}{n+1}.$$

In other words, all we are doing is computing the batting average (hits over at-bats). In Chapter 2 we are going to put a lot more rigor behind this, but for now, we are going to take advantage of the simple, intuitive updating equations that this theory provides.

Our challenge is to determine whether we should try player A, B, or C right now. At the moment, A has the best batting average of .360, based on a history of 36 hits out of 100 at-bats. Why would we try player B, whose average is only .333? We easily see that this statistic is based on only three at-bats, which would suggest that we have a lot of uncertainty in this average.

We can study this formally by setting up the decision tree shown in Figure 1.10. For practical reasons, we can only study a problem that spans two at-bats. We show the current prior probability of a hit, or no hit, in the first at-bat. For the second at-bat, we show only the probability of getting a hit, to keep the figure from becoming too cluttered.

Figure 1.11 shows the calculations as we roll back the tree. Figure 1.11(c) shows the expected value of playing each hitter for exactly one more at-bat using the information

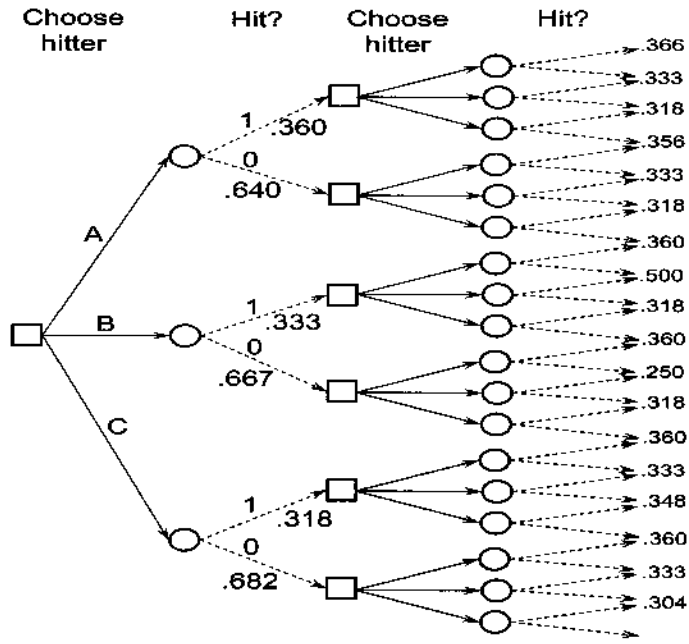


Figure 1.10 The decision tree for finding the best hitter.

obtained from our first decision. It is important to emphasize that after the first decision, only one hitter has had an at-bat, so the batting averages only change for that hitter. Figure 1.11(b) reflects our ability to choose what we think is the best hitter, and Figure 1.11(a) shows the expected value of each hitter before any at-bats have occurred. We use as our reward function the expected number of total hits over the two at-bats. So, if we choose batter A, the expected value is

$$.720 = .360(1 + .366) + .640(0 + .356),$$

where .360 is our prior belief about his probability of getting a hit; .366 is the expected number of hits in his second at-bat (the same as the probability of getting a hit) given that he got a hit in his first at-bat. If player A did not get a hit in his first at-bat, his updated probability of getting a hit, .356, is still higher than any other player. This means that if we have only one more at-bat, we would still pick player A even if he did not get a hit in his first at-bat.

Although player A initially has the highest batting average, our analysis says that we should try player B for the first at-bat. Why is this? On further examination, we realize that it has a lot to do with the fact that player B has had only three at-bats. If this player gets a hit, our estimate of his probability of getting a hit jumps to 0.500, although it drops to .250 if he does not get a hit. If player A gets a hit, his batting average moves from .360 to .366, reflecting the weight of his much longer record.

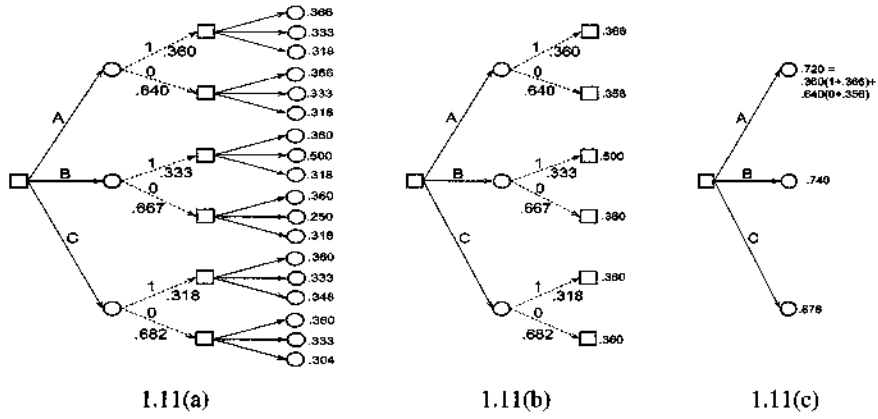


Figure 1.11 (a) Expected value of a hit in the second at-bat. (b) Value of best hitter after one at-bat. (c) Expected value of each hitter before first at-bat.

This is our first hint that it can be useful to collect information about choices where there is the greatest uncertainty.

1.6.4 Discussion

These simple examples illustrate some of the issues that are fundamental to information collection. First, we see that collecting information changes our beliefs about uncertain quantities. Second, when we change our beliefs, we change our decisions which produces an economic impact. From this analysis, we can compute the expected value of the information.

In the credit risk example, the *measurement decision* (purchasing the credit risk report) was completely separate from the *implementation decision* (whether or not to give a loan). There are many problems where we learn from our actions. For example, we might grant a loan and then learn from this experience. This means giving a loan allows us to observe whether or not someone defaults on the loan, which might then change our behavior in the future. We can ignore the value of this information when making a decision, but one of the goals of this book is to use this value of information in our decisions.

Most of the problems that we address in this book can be visualized as a decision tree. Decisions trees, however, grow very quickly with the number of decisions being made, the number of random outcomes that might occur, and the number of time periods. For example, the very simple decision tree that we used in Section 1.6.3 to analyze which baseball player we should use grows quickly if we look more than two at-bats into the future. With three players to choose from, and only two possible outcomes, the tree grows by a factor of six for every at-bat we add to our planning

horizon. If we want to plan over the next 10 at-bats, our decision tree would have $6^{10} = 60,466,176$ end points. And this is a tiny problem.

Although decision trees are often impractical as a computational device, they are extremely useful as a conceptual mechanism for understanding sequential decision problems.

1.7 WEBSITE AND DOWNLOADABLE SOFTWARE

The book is supported by additional material at the website

http://optimallearning.princeton.edu/

The website will provide additional readings, chapter notes and comments, sample applications (many drawn from the projects generated by students in the course *Optimal Learning* taught at Princeton University), and downloadable software. Most of the software is in the form of MATLAB modules that offer implementations of some of the algorithms. One module, the Optimal Learning Calculator, uses a spreadsheet front-end which talks to a series of Java-based modules which implement some of the algorithms. The spreadsheet offers an interactive environment which makes it possible to experiment with a variety of learning policies. You can solve a learning problem manually, or simulate any of a series of policies on problems. These problems can be randomly generated, or entered manually by hand.

1.8 GOALS OF THIS BOOK

There are a number of communities that address the general problem of collecting information. These communities span economics, computer science, statistics, and operations research (which in turn includes subcommunities such as decision analysis, applied probability, and simulation). One of our goals is to bring these communities together under a common vocabulary and notation.

Given the diversity of problems where learning arises, it is not surprising that a multitude of techniques have evolved to address these problems. Our presentation reviews a number of these techniques, but considerable emphasis is put on a concept called the *knowledge gradient* which guides measurements based on the marginal value of a single measurement. The knowledge gradient is comparable to the gradient in classical optimization, but focuses on the value of information. The power of the knowledge gradient is that it is a simple concept that can be applied to a wide range of problems. In fact, this technique opens up new problems, and it allows us to consider learning in settings where it has previously not been considered. Our empirical work to date suggests that it is a surprisingly robust strategy, in that it is usually competitive with competing techniques, while often outperforming other methods without the need for tunable parameters.

The book is aimed at students and professionals with a basic course in statistics and a full course in probability. The presentation emphasizes concepts and practical tools over heavy mathematical development.

PROBLEMS

1.1 Pick a problem that involves sequential information and decision processes. Ideally, the problem is one of special interest to you. Give a short description of the problem in plain English. Then describe the following components listed below. In each case, a candidate answer is provided using the setting of finding a new compound for storing energy in a battery (this is a fairly complicated example - there are many settings which are much simpler).

a) What decision are you making that determines the information you collect? [Example: Testing a particular molecule.]

b) Precisely what information is being observed as a result of your choice? [Example: The amount of energy stored per pound.]

c) What decision are you going to make with this information? [Example: We will use this information to decide which type of battery is the most economical, which in turn will impact our decision of whether to use batteries, flywheels, or hydroelectric reservoirs as our major form of storage.]

d) What is the economic impact of the decision? [Example: The information will allow the Department of Energy to determine if it should invest in rooftop solar panels, where batteries are needed to smooth out variations in solar energy, or more distant wind turbines, which can be used in conjunction with water reservoirs.]

1.2 For each of the situations below, identify whether it is an online or offline learning problem. Then, identify the measurement decision (what is being measured), and how you evaluate the quality of the measurements (that is, what is the value of the information you are collecting).

a) The adventurer Steve Fossett was lost somewhere in Nevada. The problem is to design a search process that might identify the location of the plane he was flying.

b) You would like to find the best price for a downloadable song by adjusting the price.

c) A bank evaluates loans based on a set of attributes determined from a loan application. Some loans are turned down, others are approved. For the loans that are approved, the bank can observe if the loan is repaid or defaults. The bank can then later use this data to correlate the default rate to the attributes of the loan. Since the bank is not able to observe defaults on loans which are not

approved, it occasionally may decide to grant a loan which its policy suggests should be turned down, just to observe the information.

d) A doctor administers drugs to control the blood pressure in a patient. The doctor will adjust both the type of medication and the dosage, observing the effect on the patient's blood pressure.

1.3 In Section 1.6.2, we addressed the problem of whether a bank should request a credit report. Repeat the exercise of finding out whether the bank should purchase the credit report, and determine how much the bank would be willing to pay for the report. As before, assume a loan is for \$100,000. If the loan is paid off, the bank makes \$10,000. If the loan defaults, assume that the bank loses on average \$30,000 (since a portion of the loan would have been repaid). Assume that 85 percent of the loans are repaid. The bank has collected past statistics on credit reports which are expressed as the conditional probabilities in Table 1.5 below. So, 72 percent of the loans that did not default had a credit rating of A.

Table 1.5 Data for exercise 1.3.

		P(Credit rating | Default)		
Default	P(Credit)	A	B	C
No	0.85	0.72	0.24	0.04
Yes	0.15	0.25	0.57	0.18

1.4 In Table 1.3 for the credit risk problem, the probability $P(Credit = C, Default = Yes) = 0.16$ and $P(Credit = C, Default = No) = 0.01$. Change these probabilities to 0.17 and 0.01, respectively, and solve the decision tree in Figure 1.9 again. How did this change in the data change the behavior of the system after reviewing the credit risk report? What is the new value of the information in the credit risk report? Given an intuitive explanation.

1.5 Return to the decision tree in Section 1.6.3 where we are trying to decide which hitter to use. This decision tree has been implemented in the spreadsheet available on the book website at

http://optimallearning.princeton.edu/exercises/BaseballTree3levels.xlsx

Note that the decision tree considers three successive at-bats.

- The decision tree takes into account the information we gain from observing the outcome of the first at-bat (whether it be player A, B, or C). How would your answer have changed if we formulated the decision tree without taking advantage of the information gained from the first at-bat?

- Use the spreadsheet to compute the value of using each of the three batters, while changing the batting statistics from player B from 1 for 3 to 2 for 6 to 3 for 9 to 4 for 12 and finally 5 for 15. Draw a graph giving the value of choosing each of the three hitters (in the first at-bat) for all five sets of batting statistics. Explain intuitively why your choice of who to use for the first at-bat changes.