12

A Rational-Ecological Approach to the Exploration/Exploitation Trade-offs

Bounded Rationality and Suboptimal Performance

Wai-Tat Fu

How do humans or animals adapt to a new environment? After years of research, it is embarrassing how little we understand the underlying processes of adaptation: just look at how difficult it is to build a robot that learns to navigate in a new environment or to teach someone to master a second language. It is amazing how seagulls and vultures have learned to be landfill scavengers in the last century and be able to sort through human garbage to dig out edible morsels. At the time this chapter is written, an alligator is found in the city park of Los Angeles, outwitting licensed hunters who tried to trap the alligator for over 2 months. The ability to adapt to new environments goes beyond hardwired processes and relies on the ability to acquire new knowledge of the environment. An important step in the adaptation process is to sample the effects of possible actions and world states so that the right set of actions can be chosen to attain important goals in the new environment.

The acquisition of new knowledge of the environment is often achieved through the dynamic interactions between an organism and the environment, in which actions are performed and their effects evaluated based on the outcomes of actions. In most cases, the organism has to deal with a *probabilistically textured* environment (Brunswik, 1952), in which the outcomes of actions are uncertain. The evaluation of different actions is therefore similar to the process of sampling from probability distributions of possible effects of the actions (e.g., see Fiedler & Juslin, 2005). The sampling process can therefore be considered an interface between the organism's cognitive representation of the environment and the probabilistically textured environment (see Figure 12.1). [**AQ1**]

A central problem in the adaptation process is how to balance *exploration* of new actions against *exploitation* of actions that are known to be good. The benefit of exploration is often measured as the *utility of information*—the expected improvement in performance that might arise from the information obtained from exploration. Exploring the environment allows the agent to observe the results of different actions,

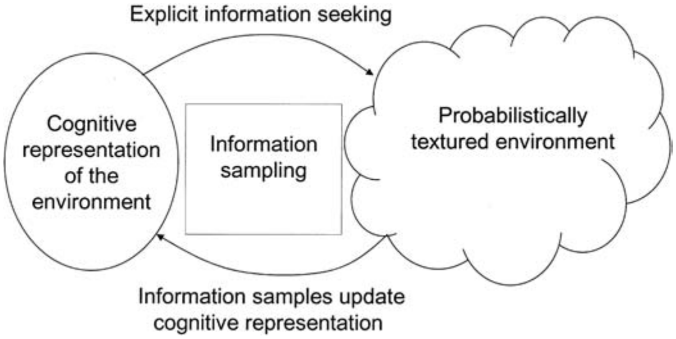
Explicit information seeking
Cognitive representation of the environment
Information sampling
Probabilistically textured environment
Information samples update cognitive representation

FIGURE 12.1 The information sampling process as an interface between the cognitive representation of the environment and the external environment.

from which the agent can learn to estimate the utility of information by some forms of reinforcement-learning [AQ2] algorithms (see Fu & Anderson, in press; Sutton & Barto, 1998; and Ballard & Sprague, chapter 20, this volume). The estimates allow "good" actions to be differentiated from "bad" actions, and that exploitation of good actions will improve performance in the future. On the one hand, the agent should keep on exploring, as exploiting the good actions too early may settle on suboptimal performance; on the other hand, the cost of exploring all possible actions may be too large to be justified. The balance between the expected cost and benefit of exploration and exploitation is therefore critical to performance in the adaptation process.

Reinforcement learning is one of the important techniques in artificial intelligence (AI) and control theory that allows an agent to adapt to complex environments. However, most reinforcement-learning techniques either require perfect knowledge of the environment or extensive exploration of the environment to reach the optimal solution. Because of these requirements, these computationally extensive techniques often fail to provide a good descriptive account of human adaptation. Instead, theories have been proposed that humans often adopt simple heuristics or *cognitive shortcuts* given the cognitive and knowledge constraints they face (see the chapters by Todd & Schooler, chapter 11, and Kirlik, chapter 14). These heuristics seem to work reasonably well, presumably because they were well adapted to the *invariants* of the environment (e.g., Anderson, 1990; Simon, 1996). The major assumption is that these invariants arise from the statistical structure of the environment that cognition has adapted to through the lengthy process of evolution. By exploiting these invariant properties, simple heuristics may perform reasonably well in most situations within the limits of knowledge and cognition.

The study of the constraints imposed by the environment to behavior is often referred to as the *ecological approach* that emphasizes the importance of the interactions between cognition and the environment and has shown considerable success in the past (e.g., chapters 11 and 14, this volume). A similar, but different, approach called the rational approach further assumes that cognition is adapted to the constraints imposed by the environment, thus allowing the construction of adaptive mechanisms that describe behavior (e.g., Anderson, 1990; Oaksford & Chater, 1998). The rational approach has been applied to explain a diverse set of cognitive functions such as memory (Anderson & Milson, 1989; Anderson & Schooler, 1991), categorization (Anderson, 1991), and problem solving (Anderson, 1990). The key assumption is that these cognitive functions optimize the adaptation of the behavior of the organism to the environment. In this chapter, I combine the ecological and rational approaches to perform a two-step procedure to construct a set of adaptive mechanisms that explain behavior. First, I perform an analysis to identify invariant properties of the environment; second, I construct adaptive mechanisms that exploit these invariant properties and show how they attain performance at a level comparable to that of computationally heavy AI algorithms. The major advantage of this rational–ecological approach is that, instead of constructing mechanisms based on complex mathematical tricks, one is able to provide answers to why these mechanisms exist in the first place, and how the mechanisms may interact with different environments.

To explain how cognition adapts to new environments, the rational–ecological approach assumes that, if cognition is well adapted to the invariant properties of the general environment, cognition should have a high tendency to use the same set of mechanisms that work well in the general environment when adapting to a new environment, assuming (implicitly) that the new environment is likely to have the same invariant properties. The implication is that, when the new environment has specific properties that are different from those in the general environment, the mechanisms that work well in the general environment may lead to suboptimal performance. Traditionally, the information samples collected from the environment are often considered unbiased and suboptimal performance or judgment biases are often explained by cognitive heuristics that fail to process the information samples according to some normative standards. In the current proposal, suboptimal performance can be explained by dynamic interactions between the cognitive processes that collect information samples and the cognitive representation of the environment that is updated by the information samples obtained. Indeed, as I will show later in two different tasks, suboptimal performance often emerges as a natural consequence of this kind of dynamic interaction among cognition, information samples, and the characteristics of the environment.

In this chapter, I will present a model of how humans adapt to complex environments based on the rational–ecological approach. In the next section, we will first cast the exploration/exploitation trade-off as a general sequential decision problem. We will then focus on the special case where alternatives are evaluated sequentially and each evaluation incurs a cost. We will then present a Bayesian satisficing model (BSM) that decides when exploration should stop. We will then show that the BSM provided good match to human performance in two different tasks. The first task is a simple map-navigation task, in which subjects had to figure out the best route between two cities. In the second task, subjects were asked to search for a wide range of information using the World Wide Web (WWW). In both tasks, the BSM matched human data well and provided good explanations of human performance, suggesting that the simple mechanisms in the BSM provide a good descriptive account of human adaptation.

The Exploration/Exploitation Trade-off

A useful concept to study human activities in unfamiliar domains is the construct of a "problem space" (Newell & Simon, 1972). A problem space consists of four major components: (1) a set of states of knowledge, (2) operators for changing one state into another, (3) constraints on applying operators, and (4) control knowledge for deciding which operator to apply next. The concept of a problem space is useful in characterizing how a problem solver searches for (exploration) different operators in the connected states of knowledge and how the accumulation of experiences in the problem space allows the problem solver to accumulate search control knowledge for deciding which operator to apply next in the future (exploitation). The concept of the problem space is similar to a Markov decision process (MDP), which has been studied extensively in the domain of machine learning and AI in the last 20 years (e.g., see Puterman, 2005, for a review). A MDP is defined as a discrete time stochastic control process characterized by a set of states, actions, and transition probability matrices that depend on the actions chosen within a given state. Extensive sets of algorithms, usually in some forms of dynamic programming and reinforcement learning, have been derived by defining a wide range of optimization problems as MDPs. Although these algorithms are efficient, they often require extensive computations that make them psychologically implausible. However, the ideas of a MDP and the associated algorithms have provided a useful set of terminologies and methods for constructing a descriptive theory of human performance. Indeed, applying ideas from machine learning to psychological theories (or vice versa) has a long history in cognitive science. By relating the concepts of operator search in a problem space to that of MDPs, another goal of the current analyses is to bridge the gap between research in cognitive psychology and machine learning.

In this section, I will borrow the terminologies from MDPs to characterize the problem of balancing exploration and exploitation and apply the rational–ecological approach to replace the complex algorithms by a BSM. I will show that the BSM uses simple, psychologically plausible mechanisms that successfully describe human behavior as they adapt to new environments.

Sequential Decision Making

Finding the optimal exploration/exploitation trade-off in a complex environment can be cast as a sequential decision-making (SDM) problem in a MDP. In general, a SDM problem is characterized by an agent choosing among various actions at different points in time to fulfill a particular goal, usually at the same time trying to maximize some form of total reinforcement (or minimize the total costs) after executing the sequence of actions. The actions are often interdependent, so that later choice of actions depends on what actions have been executed. In complex environments, the agent has to choose from a vast number of combinations of actions that eventually may lead to the goal. In situations where the agent does not have complete knowledge of the environment, finding the optimal sequence of actions requires exploring the possible sequences of actions while learning which of them are better than the others. A good balance of exploration and exploitation is necessary when the utility of exploring does not justify the cost of exploring all possible sequences, as in the case of a complex environment such as the WWW.

Many cognitive activities, such as skill learning, problem solving, reasoning, or language processing, can be cast as SDM problems, and the exploration/exploitation trade-off is a central problem to these activities.[1] The optimal solution to the SDM problem is to find the sequence of actions so that the total reinforcement obtained is maximized. This can often be done by some forms of reinforcement learning, which allows learning of the values of the actions in each problem state so that the total reinforcement received is maximized after executing the sequence of actions that lead to the goal (see, e.g., Watkins, 1989). These algorithms, however, often require perfect knowledge of the environment; even if the knowledge is available, complex computations are required to derive the optimal solution. The goal of the rational–ecological approach is to show that the requirement of perfect knowledge of the environment and complex computations can be replaced by simple mechanisms with certain assumption of the properties of the environment.

When Search Costs Matters: A Rational Analysis

Algorithms for many SDM problems use the *softmax* method to select actions in each state. Interestingly, the softmax method by itself offers a simple way to tackle the problem of balancing exploration and exploitation. Specifically, the softmax equation is based on the Gibbs, or Boltzmann, distribution:

$$P(a_k|s) = \frac{\exp(v(a_k, s)/t)}{\sum_i \exp(v(a_i, s)/t)}, \tag{1}$$

in which $P(a_k|s)$ is the probability that the action a_k will be selected in state s, $v(a_k,s)$ is the value of action a in state s, t is positive parameter called the temperature, and the summation is over all possible actions in state s. The equation has the property that when the temperature is high, actions will be (almost) equally likely to be chosen (full exploration). As the temperature decreases, actions with high values will be more likely to be chosen (a mix of exploration and exploitation), and in the limiting case, where $t \rightarrow 0$, the action with the highest value will always be chosen (pure exploitation). The balance between exploration and exploitation can therefore be controlled by the temperature parameter in the softmax equation. In fact, the softmax method has been widely used in different architectures to handle the exploration/exploitation tradeoffs, including architecture of cognition theory–rational (ACT-R; see Anderson et al., 2004). Recently, it has also been shown that the use of the softmax equation in reinforcement learning is able produce a wide range of human and animal choice behavior (Fu & Anderson, in press). **[AQ3]**

Although the softmax method can lead to reasonable exploration/exploitation trade-offs, it assumes that the values of all possible actions are immediately available without cost. The method is therefore only useful in simple or laboratory situations where the alternatives are presented at the same time to the decision maker; in that case, the search cost is negligible. In realistic situations, the evaluation process itself may often be costly. For example, in a chess game, the number of possible moves is enormous, and it is unlikely that a person will exhaust the exploration of all possible moves in every step. A more plausible model is to assume that alternatives are considered sequentially, in that case a stopping rule is required to determine when evaluation should stop (e.g., Searle & Rapoport, 1997). The problem of deciding when to stop searching can be considered a special case of the exploration/ exploitation trade-offs discussed earlier: At the point where the agent decides to stop searching, the best item encountered so far will be selected (exploitation) and the search for potential better options (exploration) will stop.

Finding the optimal stopping rule (thus the optimal exploration/exploitation trade-off) is a computationally expensive procedure in SDM problems. The goal of the following analyses is to replace these complex computations by simple mechanisms that exploit certain characteristics of the environment. The basic idea is to use a local stopping rule based on some estimates of the environment so that when an alternative is believed to be good enough no further search may be necessary. This is the essence of *bounded rationality* (Simon, 1955), a concept that assumes that the agent does not exhaust all possible options to find the optimal solution. Instead, the agent makes choices based on the mechanism of *satisficing*, that is, the goodness of an option is compared to an aspiration level and the evaluation of options will stop once an option that reaches the aspiration level is found. There are a number of ways the aspiration level can be estimated. Here, I will show how the aspiration level can be estimated by an adaptation process to an environment based on the rational analysis framework.

Optimal Exploration in a Diminishing-Return Environment

One major assumption in the current analysis is that when the agent is searching sequentially for the right actions, the potential benefits of obtaining a better action tend to diminish as the search cost increases. This kind of diminishing-return environment is commonly found in the natural world, as well as in many artificial environments. For example, in his seminal
[AQ4] article, Stigler (1961) shows that most economic information in the market place has this diminish-return property. Research on animal foraging has found that food patches in the wild seem to have this characteristic of diminishing returns, as the more of the patch the animal consumes, the lower the rate of return will be for the remainder of the patch because the food supply is running out (Stephens & Krebs, 1986). Recently, Pirolli and Card (1999) also found that large information structures tend to have this diminishing-return characteristic. To further illustrate the generality of this diminishing-return property, I will give a real-world example of information-seeking task below.

Consider a person looking for a plane ticket from Pittsburgh to Albany on the internet. Assume that the P value of each link is calculated by the following simple preference function:

$$P = \text{Time} + \text{Stopover} + \text{Layover}, \qquad (2)$$

in which Time, Stopover, and Layover are variables that take values from 1 (least preferable) to 5 (most preferable). For example, a flight that leaves at 11 a.m., makes one stopover and has a layover of 5 hours has Time = 5, Stopover = 3, and Layover = 1 (thus $P = 9$). By using a simple set of rules that transform each flight encounter on the Web to a P value, Figure 12.2 shows the P values of the flights in their order of encounter

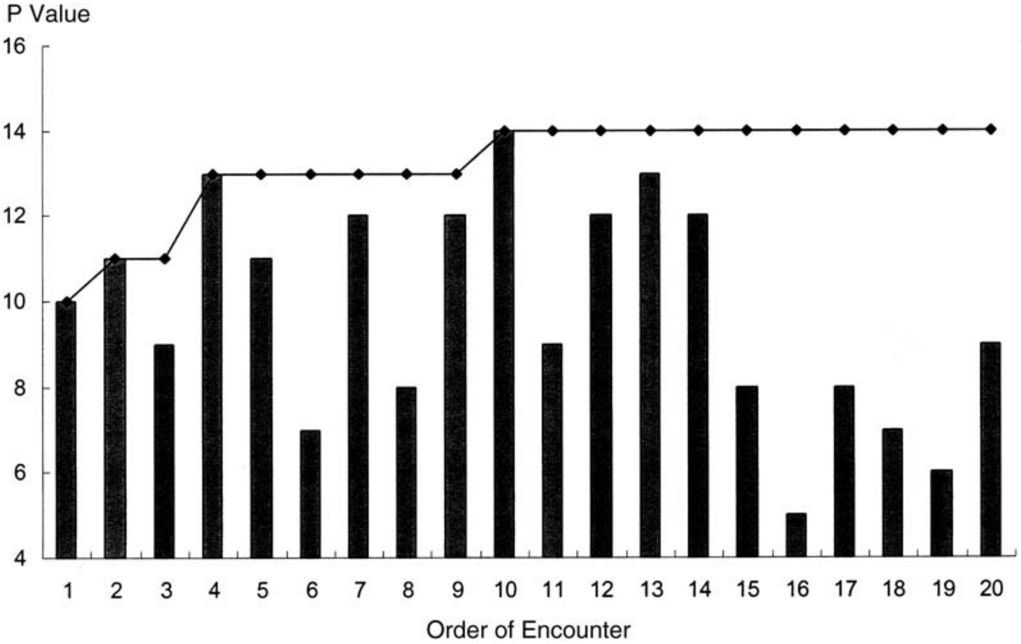

FIGURE 12.2 The P(reference) value of the links encountered on a Web page. The line represents the P value of the best link encountered so far.

from a popular Web site that sells plane tickets. We can see that a few desirable flights are found in the first few encounters, but the likelihood of finding a better flight is getting smaller and smaller, as shown by the line in the figure. It can be shown that this property of diminishing-return is robust with different preference functions or Web sites.[2]

If we assume the simplistic view that the evaluation [**AQ5**] of each action incurs a constant cost,[3] and the information obtained from each evaluation (i.e., exploration of a new action) reduces the expected execution cost required to finish a task, we can calculate the relationship among the number of evaluations (n), the evaluation [**AQ6**] costs (n^*C), the expected execution costs $f(n)$, and the total costs $f(n) + n^*C$). As shown in Figure 12.3, the positively sloped straight line represents the increase of evaluation costs with the number of evaluations. The curve $f(n)$ represents the expected execution costs as a function of the number of actions evaluated. The function $f(n)$ has the characteristic of diminishing return, so that more evaluations will lead to smaller savings in execution costs. The U-shape curve is the total costs, which equals the sum of evaluation costs and execution costs. The U-shape curve implies that optimal performance is associated with a moderate number of evaluations. In other words, too much or too little evaluations may lead to suboptimal performance (as measured by the total costs).

The Bayesian Satisficing Model (BSM)

With the assumption of the invariant property of diminishing return in the general information environment, the next step is to propose a set of adaptive mechanisms that exploit this property. I will show that the Bayesian satisficing model, which combines a Bayesian learning mechanism and a simple, local decision rule, produces good match to how humans adaptively balance exploration and exploitation in a general environment with this diminishing-return property (Fu & Gray, in press). The Bayesian learning mechanism in the BSM calculates the expected utility of information in terms of the expected improvement in performance resulting from the new information. The local decision rule then compares the evaluation cost with the utility of information and will stop evaluating actions when the cost exceeds the utility. The logic of the model is that if cognition is well adapted to the characteristic of diminishing returns in which a local decision rule performs well, then cognition should have

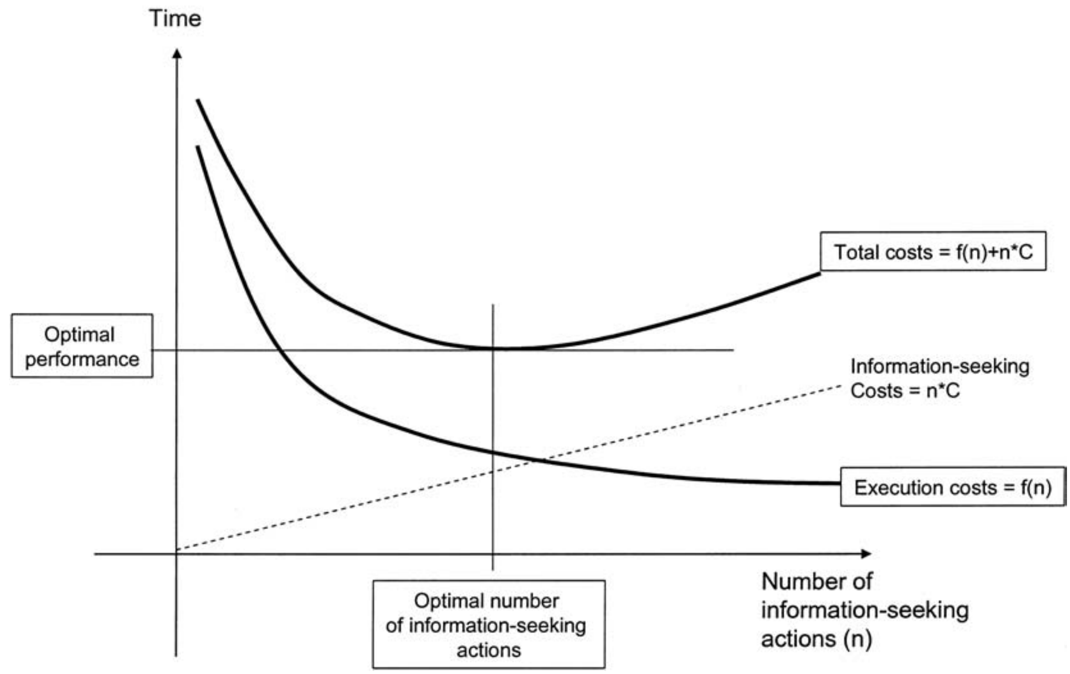

FIGURE 12.3 Optimal exploration in a diminishing-return environment.

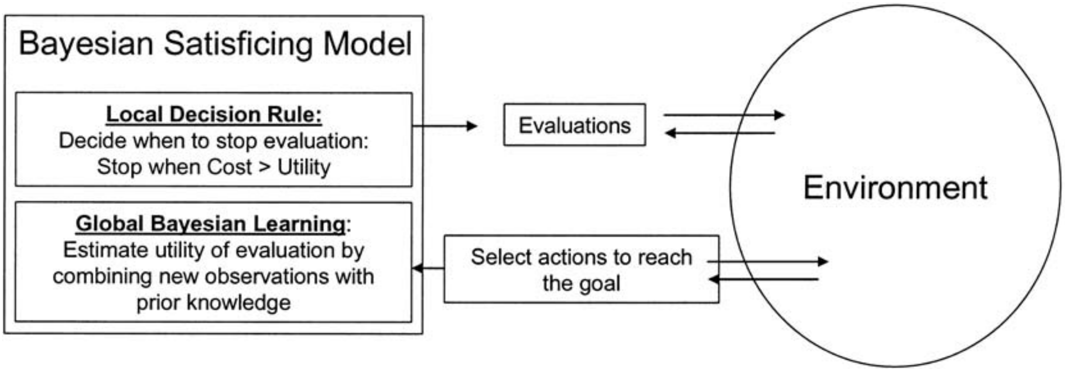
Bayesian Satisficing Model
Local Decision Rule:
Decide when to stop evaluation:
Stop when Cost > Utility
Global Bayesian Learning:
Estimate utility of evaluation by combining new observations with prior knowledge
Evaluations
Select actions to reach the goal
Environment

FIGURE 12.4 The structure of the Bayesian satisficing model.

a high tendency to use the same rule when adapting to a new environment, assuming that the new environment is likely to have the same diminishing-return characteristic.

The details of the BSM are illustrated in Figure 12.4, which shows the two major processes that allow the model to adapt to the optimal level of evaluations in a diminishing-return environment that maps to the variables in Figure 12.3: (1) the estimation of the function $f(n)$ and (2) the decision on when to stop evaluating further options. The first process requires the understanding of how people estimate the utility of additional evaluations based on experience. The second process requires the understanding of how the decision to stop evaluating further options is sensitive to the cost structure of the environment. In the global learning process, the model assumes that execution costs can be described by a diminishing-return function of the number of evaluations [i.e., $f(n)$]. A local decision rule is used to decide when to stop evaluating the next option (see Figure 12.5) based on the existing estimation of $f(n)$. Specifically, when the estimated utility of the next evaluation [i.e., $f(N) - f(N + 1)$] is lower than its cost, the model will stop evaluating the next option. This local decision rule decides how many evaluations are performed. The time spent to finish the task given the particular number of evaluations is then used to update the existing knowledge of $f(n)$ based on Bayes's theorem.

Fu and Gray (in press) ran a number of simulations of the BSM using a variety of diminish-return environments, and showed that the BSM made a number of

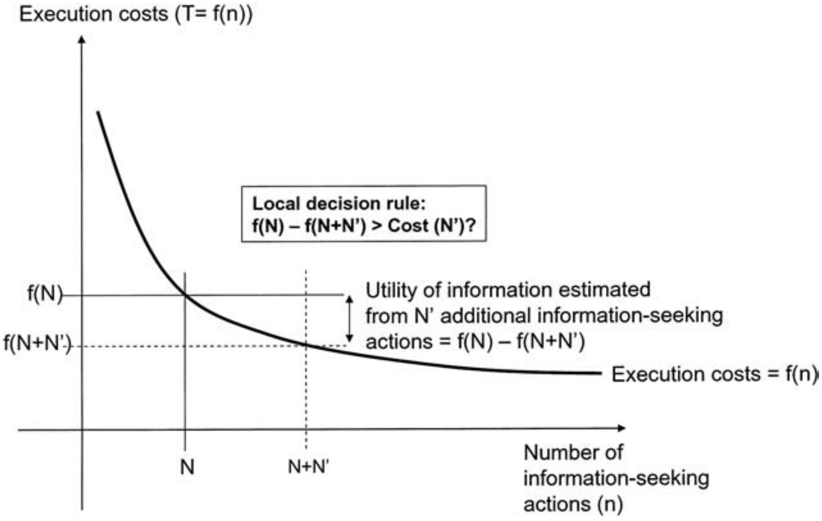

FIGURE 12.5 The local decision rule in the Bayesian satisficing model.

interesting predictions on behavior. In summary, the simulation results show that (1) with sufficient experience, people make good trade-offs between exploration and exploitation and converge to a reasonably good level of performance in a number of diminishing-return environments, (2) people respond to changes in costs faster than changes in utility of evaluation, and (3) in a local-minimum environment, high cost may lead to premature termination of exploration of the problem space, thus suboptimal performance.

Figure 12.6 illustrates the third prediction of the BSM. The flat portion of $f(n)$ (i.e., region B) represents what we refer to as a *local-minimum* environment, in which the marginal utility of exploration [i.e., the slope of $f(n)$] varies with the number of evaluations. The marginal utility is high during initial exploration, becomes flat with intermediate number of evaluations, but then becomes high again with greater number of evaluations. Using the local decision rule, exploration is likely to stop at the flat region (i.e., when the marginal utility of evaluation is lower than the cost), especially when the cost is high. We therefore predict that in a local-minimum environment the use of a local decision rule will predict poor exploration of the task space, especially when the cost is high.

Testing the BSM Against Human Data

In this section, I will summarize how the BSM matched human performance in two tasks. In the first task, subjects were given a simple map navigation task in which they were asked to find the best route between two points on the map (Fu & Gray, in press). Subjects were given the option to obtain information on the speeds of different routes before they started to navigate on the map. The cost and utility of information was manipulated to study how these factors influenced the decision on when subjects would stop seeking information. To directly test whether subjects were using the local decision rule, a local-minimum environment was constructed. The local-minimum environment had an uneven diminish-return characteristic, so that the use of a local decision rule would be more likely to prematurely stop seeking information, leaving the problem space underexplored and, as a result, performance would be suboptimal. Indeed, the human data confirmed the prediction, providing strong support for the use of a local decision rule. The second task was a real-world task in which subjects were asked to search for information using the WWW. We combined the BSM with the measure of information scent (Pirolli & Card,

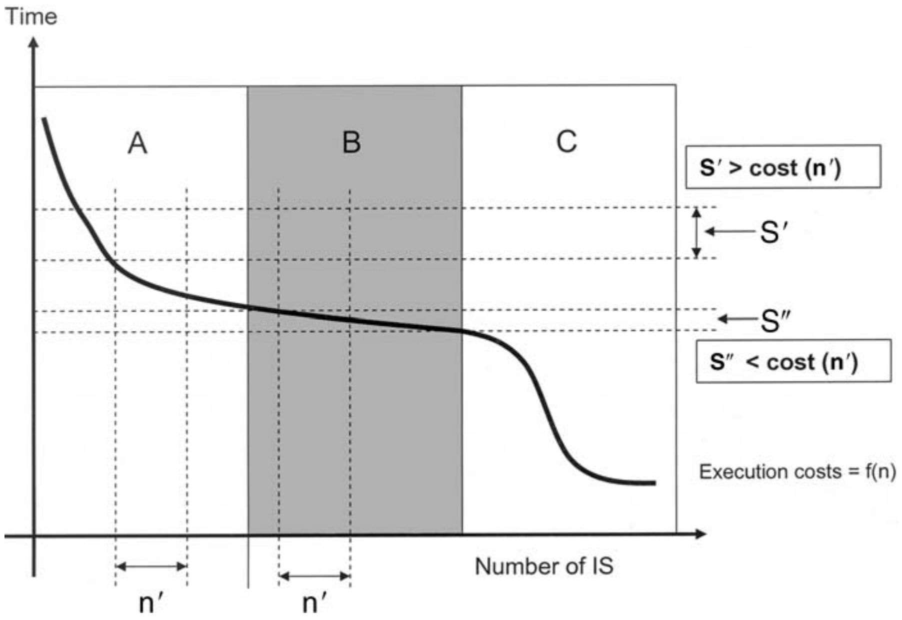

FIGURE 12.6 How a local decision rule may stop exploration prematurely in a local-minimum environment. In the figure, when the saving in execution costs is smaller than the cost of the exploration costs $[S'' < \text{cost}\ (n')]$, exploration will stop, leaving a large portion of the task space unexplored (i.e., task space C).

1999) to predict link selections and the decision on when to leave a particular Web page in two real-world Web sites. In both task, we found that the model fit the data well, suggesting that the adaptive exploration/exploitation trade-offs produced by the simple mechanisms of the BSM matched human performance well in different types of information environment. To preview our conclusions, the results from both tasks provided strong support for the BSM. The success of the BSM in explaining performance in the local-minimum environment also suggests that stable suboptimal performance is likely a result of the dynamic interactions between bounded rationality and specific properties of the environment.

The Map-Navigation Task

In the map-navigation task, subjects were presented with different maps on a computer screen and were asked to go from a start point to an endpoint on the map. A simple hill-climbing strategy (usually the shortest route) was always applicable and sufficient to accomplish the task (and any path can eventually lead to the goal), but the hill-climbing strategy was not guaranteed to lead to the best (i.e., fastest) path. With sufficient experience, subjects learned the speeds of different routes and turns and improved performance by a better choice of solution paths. The problem of finding the best path in a map could therefore be considered a SDM problem, in which each junction in the map was a discrete state, each of the possible routes passing through the junction defined a possible action in the state, and finding the fastest path defined a standard optimization problem.

The speed of the path chosen was experienced in real time (a red line went from one point to another on the map, at an average rate of approximately 1 cm/ s), but the speed of a path could also be checked beforehand by a simple mouse click (i.e., an information-seeking action). The number of information-seeking actions therefore served as a direct measure of how much exploration subjects were willing to do in the task, and the corresponding execution costs could be measured by the actual time spent to go from the start to the endpoint. We manipulated the exploration cost by introducing a lock-out time to the information-seeking action. Specifically, in the high-cost condition, after subjects clicked on the map to obtain the speed information of a path, they had to wait 1 s before they could see the information. The utility of information was manipulated by varying the difference between the fastest and slowest paths in the map. For example, when the difference was large, the potential saving in execution costs (i.e., the utility) per information-seeking action would be higher [i.e., the curve $f(n)$ in Figure 12.3 or Figure 12.5 is steeper).

To match behavior of the model to human data, the model was implemented in the ACT-R architecture (Anderson & Lebiere, 1998). The decision rule is implemented by having two competing productions, one abstractly representing exploration, and the other representing exploitation.[4] In ACT-R, each production has a utility value, which determines how likely that it will fire in a given cycle by the softmax equation stated above. The utility value of each production is updated after each cycle according to a Bayesian learning mechanism (see Anderson & Lebiere, 1998, for details) as in the BSM (see Figure 12.4). In general, when the utility of the exploration production is higher than that of the exploitation production, the model is likely to continue to explore. However, when the utility of the exploration production falls below that of the exploitation production, the model will likely stop exploring. The competition between the two productions through the softmax equation therefore serves as a stochastic version of the local decision rule in the BSM.

Because of the space limitation, only a briefly summary of the major findings of the three experiments was presented here (for details, see Fu & Gray, in press). First, in diminishing-return environments with different costs and utilities of information, subjects were able to adapt to the optimal levels of exploration. The BSM model provided good fits to the data, suggesting that the local decision rule in the BSM was sufficient to lead to optimal performance. Second, in environments where the costs or utilities of exploration were changed, subjects responded to changes in costs faster than changes in utilities of information. Finally, when the cost was high in a local-minimum environment, subjects prematurely stopped seeking information and stabilized at suboptimal performance.

The empirical and simulation results suggest that subjects used a local decision rule to decide when to stop seeking information. Perhaps the strongest evidence for the use of a local decision rule was the finding that in the local-minimum environment, high cost of exploration led to “premature” stopping of information seeking, and as a result, performance stabilized at a suboptimal level. Although the BSM was effective in finding the right level of information seeking in most situations, the nature of local processing inherently limits the exploration of the environment. Indeed, we found that the same model, when interacting with environments with different properties, exhibited very different behavior. In particular, in a local-minimum environment, the local decision rule often results in "insufficient" information seeking when high information-seeking costs discourage exploration of the environment. On the basis of this result, it is concluded that suboptimal performance may emerge as a natural consequence of the dynamic interactions between bounded rationality and the specific properties of the environment.

A Real-World Information-Seeking Task: Searching on the WWW

To further test the behavior of the BSM, a real-world task was chosen and human performance on this task was compared with that of the BSM. Similar to the map-navigation task, searching on the World Wide Web is a good example of a SDM problem: Each Web page defines a state in the problem space, and clicking on any of the links on the Web page defines a subset of all possible actions in that state (other major actions include going back to the previous pages or going to a different Web site). The activities on the WWW can therefore be analyzed as a standard MDP. Because the number of Web pages on the Internet is enormous, exhaustive search of Web pages is impossible. Before I present how to model the exploration/exploitation trade-off in this task, I need to digress to discuss a measure that captures the user's estimation of how likely a link will lead to the target information. One such measure is called *information scent*, which will be described next.

Information Scent

Pirolli and Card (1999) developed the information foraging theory (IFT) to explain information-seeking behavior in different user interfaces and WWW navi-
[**AQ7**] gation (Pirolli & Fu, 2003; Fu & Pirolli, in press). The IFT assumes that information-seeking behavior is adaptive within the task environment and that the goal of the information-seeker is to maximize information gain per unit cost. The concept of information scent measures the mutual relevance of text snippets (such as the link text on a Web page) and the information goal. The measure of information scent is based on a Bayesian estimate of the relevance of a distal source of information conditional on the proximal cues. Specifically, the degree to which proximal cues predict the occurrence of some unobserved target information is reflected by the strength of association between cues and the target information. For each word *i* involved in the user's information goal, the accumulated activation received from all associated information scents for word *j* is calculated by

$$IS(Link_k) = \sum_i \sum_j W_j \log\left(\frac{\Pr(i \mid j)}{\Pr(i)}\right), \tag{3}$$

where Pr(*i*|*j*) is the probability (based on past experience) that word *i* has occurred when word *j* has occurred in the environment; W_j represents the amount of attention devoted to word *j*; and Pr(*i*) is the base rate probability of word *i* occurring in the environment. Equation 1 is also known as *pointwise mutual informa-* [**AQ8**]
tion (Manning & Schuetze, 1999) or PMI.[5] The actual probabilities are often estimated by calculating the co-occurrence of word *i* and *j* and the base frequencies of word *i* from some large text corpora (see Pirolli & Card, 1999). The measure of information scent therefore provides a way to measure how subjects evaluate the utility of information contained in a link on a Web page.

The SNIF-ACT Model

On the basis of the IFT, Fu and Pirolli (in press) developed a computational model called SNIF-ACT (scent-based navigation and information foraging in the ACT architecture) that models user–WWW interactions. The newest version of the model, SNIF-ACT 2.0, is based on a rational analysis of the information environment. I will focus on the part where the model is facing a single Web page and has to decide when to stop evaluating links on the Web page. In fact, the basic idea of this part of the SNIF-ACT model was identical to that of the BSM, which was composed of a Bayesian learning mechanism and a local decision rule (Figure 12.4). Specifically, the model assumed that when users evaluated each link on a Web page, they incrementally updated their perceived relevance of the Web page to the target information according to a Bayesian learning process. A local decision rule then decided when to stop evaluating link: the evaluation of the next link continued until the perceived relevance was lower than the cost of evaluating for the next link. At that point, the best link encountered so far will be selected. Details of the model will be presented below.

When the model is facing a single Web page, it had the same exploration/exploitation trade-off problem: to balance between the utility of evaluating the next link and the cost of doing so. However, in contrast to the map-navigation task, the utility of information was not measured by time. Instead, the utility of information (from evaluating the next link) is measured by the likelihood that the next link will lead to the target information. Details of the analysis can be found in Fu & Pirolli (in press). The probability that the current Web page will eventually lead to the target information after the evaluation of a set of links L_n is

$$P(\text{Target}|L_n) = K\sum_{j=0}^{n}\frac{\Gamma(\alpha + J)}{\Gamma(\alpha + n)}x(\text{O}_j), \tag{4}$$

where X is a variable that measures the closeness to the target; O_j is the observation of link j on the current Web page; K, α, and n are parameters to be estimated. The link likelihood equation is derived from the Bayes's theorem and thus is identical to the Bayesian learning mechanism in BSM (see Figure 12.4). As explained, $X(\text{O}_j)$ can be substituted by the measure of information scent (i.e., the information scent equation) of each link j. The link likelihood equation provides a way to incrementally update the probability that a given Web page will eventually lead to the target information after each link is evaluated [i.e., $f(n)$ in Figures 12.3 and 12.5]).

The model is again implemented in the ACT-R architecture. To illustrate the behavior of the model, we will focus on the case where the model is facing a single Web page with multiple links. There are three possible actions, each represented by a separate production: *attend-to-link*, *click-link*, and *backup-a-page*. Similar to the BSM model in the map-navigation task, these productions compete against each other according to the softmax equation (which implements the local decision rule in the BSM; see Figure 12.3). In other words, at any time, the model will attend to the next link on the page (exploration), click on a link on a page (exploitation), or decide to leave the current page and return to the previous page. The utilities of the three productions are derived from the link likelihood equation, and they can be shown as:

$$\text{Attend-to-Link}: U(n+1) = \frac{U(n) + IS(link)}{1 + \text{N}(n)} \tag{5}$$

$$\text{Click-Link}: \qquad U(n+1) = \frac{U(n) + IS(bestlink)}{1 + k + \text{N}(n)}$$

$$\text{Backup-a-Page}: U(n+1) = MIS(Previous\ Pages) - MIS(links\ 1\ to\ n) - \text{GoBackCost}.$$

In the equations above, $U(n)$ represents the utility of the production at cycle n, $IS(link)$ represents the information scent of the currently attended link, $N(n)$ represents the number of links already attended on the Web page after cycle n (one link is attended per cycle), $IS(bestlink)$ is the link with the highest information scent on the Web page, k is a scaling parameter, $MIS(page)$ is the mean information scent of the links on the Web page, and *GoBackCost* is the cost of going back to the previous page. The values of k and GoBackCost are estimated to fit the data. The equation for backup-a-page assumes that the model is keeping a moving average of the information scent encountered in previous pages. It can be easily shown that the utility of backup-a-page will increase as the information scent of the links encountered on the current Web page declines.

Figure 12.7 shows a hypothetical situation when the model is processing a Web page in which the information scent decreases from 10 to 2 as the model attends and evaluates Links 1 to 5. The information scent of the links from 6 onward stays at 2. The mean information scent of the previous page was 10, and the noise parameter t in the softmax equation was set to 1.0. The value of k and GoBackCost were both set to 5. The initial utilities of all productions were set to 0. We can see that initially, the probability of choosing attend-to-link is high. This is based on the assumption that when a Web page is first processed, there is a bias in learning the utility of links on the page before a decision is made. However, as more links are evaluated, the utilities of the productions decreases [as the denominator gets larger as $N(n)$ increases]. Since the utility of attend-to-link decreases faster than that of click-link [since $IS(Best) = 10$, but $IS(link)$ decreases from 10 to 2], the probability of choosing attend-to-link decreases but that of click-link increases. The implicit assumption of the model is that since evaluation of links takes time, the more links that are evaluated, the more likely that the best link evaluated so far will be selected (otherwise the time cost may outweigh the benefits of finding a better link). As shown in Figure 13.7, after four links on the hypothetical Web page have been evaluated, the probability of choosing click-link is larger than that of attend-to-link. At this point, if click-link is selected, the model will choose the best (in this case the first) link and the model will continue to process the next page. However, as the selection

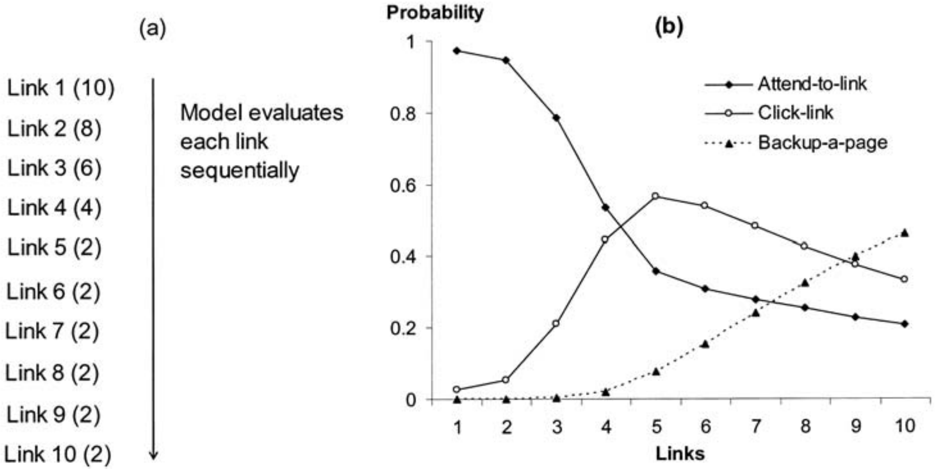

FIGURE 12.7 (a) A hypothetical Web page in which the information scent of links decreases linearly from 10 to 2 as the model evaluated links 1 to 5. The information scent of the links from 6 onward stays at 2. The number in parenthesis represents the value of information scent. (b) The probability of choosing each of the competing productions when the model processes each of the link in (a) sequentially. The mean information scent of the previous pages was 10. The noise parameter *t* was set to 1.0. The initial utilities of all productions were set to 0. *k* and *GoBackCost* were both set to 5.

process is stochastic (because of the softmax equation), attend-to-link may still be selected. If this is the case, as more links are evaluated [i.e., as $N(n)$ increases], the probability of choosing attend-to-link and click-link decreases. However, the probability of choosing backup-a-page is low initially because of the high *GoBackCost*. However, as the mean information scent of the links evaluated [i.e., *MIS*(*links 1 to n*)] on the page decreases, the probability of choosing backup- a-page increases. This happens because the mean information scent of the current page is *perceived* to be dropping relative to the mean information scent of the previous page. In fact, after eight links are evaluated, the probability of choosing backup-a-page becomes higher than that of attend-to-link and click-link, and the probability of choosing backup-a-page keeps on increasing as more links are evaluated (as the mean information scent of the current page decreases). We can see how the competition between the productions can serve as a local decision rule that decides when to stop exploration.

The Tasks

Data from tasks performed at two Web sites in the Chi et al. (2003) data set were selected: (1) help.yahoo.com (the help system section of Yahoo!) and (2) parcweb.parc.com (an intranet of company internal information). We will refer to these sites as *Yahoo* and *ParcWeb*, respectively, for the rest of the article. Each of these Web sites (Yahoo and ParcWeb) had been tested with a set of eight tasks, for a total of $8 \times 2 = 16$ tasks. For each site, the eight tasks were grouped into four categories of similar types. For each task, the user was given a specific information goal in the form of a question (e.g., “Find the 2002 Holiday Schedule”). The Yahoo and ParcWeb data sets come from a total of $N = 74$ adult users (30 users in the Yahoo data set and 44 users in the ParcWeb data set). Of all the user sessions collected, the data were cleaned to throw out any sessions that employed the site’s search engine as well as any sessions that did not go beyond the starting home page.

In general, we found that in both sites, there were only a few (<10) “attractor” pages visited by most of the users, but there were also many pages visited by fewer than 10 users. In fact, many Web pages in both sites were visited only once. To set our priorities, we decided that it was more important to test whether the model was able to identify these attractor pages. In fact, Web pages that were visited fewer than five times among all users seemed more random than systematic, and thus were excluded from our analyses. These Web pages amounted to approximately 30% of all the Web pages visited by the users.

To test the predictions of the model on its selection of links, we first started the model on the same pages as the participants in each task. The model was then run the same number of times as the number of participants in each task and the selection of links were recorded. After the recordings, in case the model did not pick the same Web page as participants did, we forced the model to follow the same paths as participants. This process repeated until the model had made selections on all Web pages visited by the participants. The selections of links by the model were then aligned with those made by participants. The model provided good fits to the data ($R^2 = 0.90$ for Yahoo and $R^2 = 0.72$ for ParcWeb).

One unique feature of the WWW in the exploration of actions was the ability to go back to a previous state. Indeed, the decision to go back to the previous state indicated that the user believed further search along the same path might not be justified. It was therefore important that the model was able to match when users decided to go back to the previous state. In the model, when the information scent of a page dropped below the mean information scent of previous pages, the probability of going back increased. Indeed, the model's decisions to go back a page were highly correlated with human decisions to go back for both the Yahoo ($R^2 = 0.80$) and ParcWeb ($R^2 = 0.73$) sites. These results provided further support for the adaptive trade-offs between exploration and exploitation implemented by the model.

When searching for information on the WWW, the large number of Web pages makes exhaustive search impossible. When faced with a Web page with a list of links, the decision on which link to follow can be considered a balance between exploration and exploitation. I showed that the BSM matched the behavior of the users well. Since the study was not a controlled experiment, it was hard to manipulate the information environment to test directly whether suboptimal performance would result from the use of a local decision rule. However, it was promising that the BSM, combining a Bayesian learning mechanism and the use of a local decision rule, was able to match the human data well when users interact with a large information structure such as the WWW.

Summary and Conclusions

When an organism adapts to a new environment, the central problem is how to balance exploration and exploitation of actions. The idea of exploration is similar to the traditional concept of *search in a problem space* (Newell & Simon, 1972), in which the problem solver needs to know when to stop searching and choose actions based on limited search control knowledge. Recently, the idea has also been studied extensively in the area of machine learning in the form of an SDM problem, and complex algorithms have been derived for finding the optimal trade-offs between exploration and exploitation in different environments.

A rational–ecological approach to the problem of balancing exploration and exploitation was described. The approach adopts a two-step procedure: (1) identify invariant properties of the general environment and (2) construct adaptive mechanisms that exploit these properties. The underlying assumption is that cognition is well adapted to the invariant properties of the general environment; when faced with a new environment, cognition tends to apply the same set of mechanisms that work well in the general environment to perform in the new environment.

It is assumed that the general information environment has an invariant property of diminishing-return. A BSM was then derived to exploit this property. The BSM dynamically obtains information samples from the new environment to update its internal representation of the new environment according to the Bayesian learning mechanism. A local decision rule is then applied to decide when to stop exploration of actions. The model matched human data well in two very different tasks that involved different information environments, showing that the simple mechanisms in the BSM can account for the adaptive trade-offs between exploration and exploitation when adapting to a new environment, a problem that usually requires complex algorithms and computations.

One major advantage of the current approach is that one is able to provide an explanation for why certain mechanisms compute the way they do. In the BSM, the local decision is effective based on the assumption of the assumed invariant property of diminishing return. Another major advantage is that complex computations can be replaced by simple heuristics that exploit the statistical properties of the environment. Indeed, finding the optimal solution in each new environment has been a tough problem for research in the area of AI and machine learning that focuses on various kinds of optimization problems in SDM. It is promising that the single set of simple mechanisms in BSM seems to be sufficient to replace complex computational algorithms by providing good match to human performance in two diversely different task environments.

Insufficient exploration often leads to suboptimal performance, as better actions are unexplored and thus not used. The model demonstrates nicely how a simple mechanism that exploits the invariant properties of the general environment may fail to provide an unbiased representation of the new environment. In fact, elsewhere we argued that this is the major reason for why inefficient procedures persist even after years of experience with the various artificial tools in the modern world, such as the many computer applications that people use everyday (Fu & Gray, 2004). We found that many of these artificial tools have the characteristics of a local-minimum environment as shown in Figure 12.6. Since the cost of exploring new (and often more efficient) procedures is often high in these computer applications, users tend to stop exploring more efficient procedures and stabilize at suboptimal procedures even after years of experience.

Notes

1. In the machine learning literature, the SDM problem is often solved as a Markov decision problem over the set of information states S, and the agent has to choose one of the possible actions in the set A. After taking action $a \in A$ from state $s \in S$, the agent's state becomes some state s' with the probability given by the transition probability $P(s'|s,a)$. However, the agent is often not aware of the current state (because of lack of complete knowledge of the environment). Instead, the agent only knows the information state i, which is a probability distribution over possible states. We can then define $i(s)$ as the probability that the person is in state s. After each transition, the agent makes an observation o of its current state from the set of possible observations O. We can define $P(o|s',a)$ as the probability that observation o is made after action a is taken and state s′ is reached. We can then calculate the next information state as:

$$i(s'|o,a) = \frac{P(o|s',a) \sum_{s \in S} P(s'|s,a)x(s)}{\sum_{s' \in S} P(o|s',a) \sum_{s \in S} P(s'|s,a)x(s)}$$

2. In fact, if one considers the value of P as a normally distributed variable, then the likelihood of finding a better alternative will naturally decrease as the sampling process continues, as one gets more to the tail of the distribution.

3. One may argue that the cost of exploration is likely to be an increasing function, which is probably true. However, the actual function does not play a crucial role in the current analyses (one still gets a U-shaped curve for the total costs in). For the sake of simplicity, a linear relationship is assumed in this analysis.

4. The productions were called hill-climbing (exploitation) and information-seeking (exploration) in Fu and Gray (in press).

5. The PMI calculations can also be found at http://glsa.parc.com.

References

Ainslie, G., & Haslam, N. (1992). Hyperbolic discounting. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 57–92). New York: Russell Sage Foundation. **[AQ9]**

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, NJ: Erlbaum. **[AQ10]**

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

———. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

———, Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 11*(4), 1036–1060.

———, & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

———, & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703–719.

———, & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2*, 396–408.

Barto, A., Sutton, R., & Watkins, C. (1990). Learning and sequential decision making. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 539–602). Cambridge, MA: MIT Press.

Brunswik, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Card, S. K., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R. W., Schraedley, P., et al. (2001). Information Scent as a Driver of Web Behavior Graphs: Results of a Protocol Analysis Method for Web Usability. *CHI 2001, ACM Conference on Human Factors in Computing Systems.* **[AQ11]**

Chi, E. H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., et al. (2003). The Bloodhound Project: Automating discovery of Web usability issues using the InfoScent simulator. *CHI 2003, ACM Conference on Human Factors in Computing Systems, CHI Letters*, 5(1), 505–512.

Fiedler, K., & Juslin, P. (2000). *Information sampling and adaptive cognition*. Cambridge: Cambridge University Press.

[AQ12] Fu, W. T., & Anderson, J. R. (in press). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*.

———, & Gray, W. D. (2004). Resolving the paradox of the active user: Stable suboptimal performance in interactive tasks. *Cognitive Science*, *28*(6).

[AQ13] ———, & Gray, W. D. (in press). Suboptimal tradeoffs in information-seeking. *Cognitive Psychology*.

[AQ14] ———, & Pirolli, P. SNIF-ACT: A model of information-seeking on the World Wide Web. *Human-Computer Interaction*. Accepted for publication.

[AQ15] Loewenstein, G. & Prelec, D. (1991). Negative time preference. *The American Economic Review*, *81*(2), 347–352.

Manning, C. D., & Schuetze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

[AQ16] Marr, D. (1982). *Vision*. San Francisco: W. H. Freedman.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford: Oxford University Press.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review 106*(4), 643–675.

Pirolli, P. L., & Fu, W.-T. (2003). *SNIF-ACT: A model of information foraging on the World Wide Web*. Ninth International Conference on User Modeling, Johnstown, Pennsylvania.

Puterman, M. L. (2005). *Markov decision processes*. Hoboken, NJ: Wiley.

Seale, D. A., & Rapoport, A. (1997). Sequential decision making with relative ranks: An experimental investigation of the "secretary problem." *Organizational Behavior and Human Decision Processes*, *69*(3), 221–236. [AQ17]

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

———. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory*. Princeton, NJ: Princeton University Press.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Watkins, C. (1989). *Learning from delayed rewards*. Unpublished doctoral dissertation, King's College, Oxford.

Author Query Sheet

Manuscript No.: Chapter-12

Dear Author,

The following queries have arisen during the setting of your manuscript. Please answer the queries using the form below.

Query no.	Query	Reply
AQ1	Fiedler & Juslin (2005) is not cited in the references. Do you mean 2000? Please verify publication date.	
AQ2	Does "chapter by Ballard et al." refer to Ballard and Spragues's chapter 20?	
AQ3	Update publication information for Fu and Anderson (in press)?	
AQ4	Stigler (1961) is not cited in the references. Please add a complete reference citation.	
AQ5	Please add missing information to the end of note 3.	
AQ6	Is the "*" the correct operational sign, or should it be a multiplication cross?	
AQ7	Please update publication status of Fu & Pirolli.	
AQ8	Display equations have been numbered sequentially. Therefore, revised "Equation 1" reference to match the correct equation.	
AQ9	Ainslie and Haslam is not cited in the text. Please add text citation; otherwise, delete reference citation.	
AQ10	Anderson 1976 is not cited in the text. Please add text citation; otherwise, delete reference citation.	
AQ11	Card et al. 2001 is not cited in the text. Please add text citation; otherwise, delete reference citation.	
AQ12	Please update publication status for Fu & Anderson (in press).	
AQ13	Update publication status for Fu & Gray (in press).	
AQ14	Please update publication status for Fu & Pirolli.	
AQ15	Loewenstein & Prelec (1991) is not cited in the text. Please add text citation; otherwise, delete reference citation.	
AQ16	Marr (1982) is not cited in the text. Please add text citation; otherwise, delete reference citation.	
AQ17	Seale & Rapoport (1997) is not cited in the text. Please add text citation; otherwise, delete reference citation.	