Learning with Dynamic Programming

Peter I. Frazier

April 15, 2011

Abstract

We consider the role of dynamic programming in sequential learning problems. These problems require deciding which information to collect in order to best support later actions. Such problems are ubiquitous, appearing in simulation, global optimization, revenue management, and many other areas. Dynamic programming offers a coherent framework for understanding and solving Bayesian formulations of these problems. We present the dynamic programming formulation applied to a canonical problem, Bernoulli ranking and selection. We then review other sequential learning problems from the literature and the role of dynamic programming in their analysis.

Frequently in operations research and management science we face uncertainty, whether in the form of uncertain demand for goods, uncertain travel times in networks, or uncertainty about which set of parameters best calibrates a simulation model. Often when facing uncertainty we are also offered the opportunity to collect some information that will mitigate uncertainty's effects. On one hand, collecting information allows us to make better decisions and produce better outcomes. On the other hand, collecting information carries a cost, whether in time or money spent, or in the opportunity cost sacrificed collecting one piece of information when we could have collected another. How can we optimally balance these costs and benefits?

One way to formulate such problems is with Bayesian methods (see, e.g., Berger (1985)), in which we begin with a *prior* probability distribution that describes our subjective belief about how likely each of the many different possible truths are. Dynamic programming (DP) (Bellman, 1954) offers a general way to formulate and solve such Bayesian sequential learning problems. This DP formulation provides value in three ways. First, in some cases, the DP formulation allows computing the optimal solution. This is usually only true for smaller problems (e.g., the inventory problem considered in Lariviere and Porteus (1999)), because the curse of dimensionality (see, e.g., Powell (2007)) prevents computing the solution to larger problems in a practically feasible amount of time. Second, the DP formulation sometimes allows structural results to be shown theoretically, which either provides intuition about the problem and the behavior of the optimal policy as in, for example, Ding et al. (2002), or provides a characterization of the optimal policy that may be directly exploited to find an optimal policy, as in sequential hypothesis testing (Wald and Wolfowitz, 1948). Third and finally, the DP formulation sometimes suggests useful heuristics (e.g., knowledge-gradient methods (Frazier, 2009)) for complex and large-scale learning problems.

The sequential learning problems that we consider here have been studied in several separate communities. Within the operations research community, sequential learning problems, as well as the way in which DP can be used to confront them, were recognized in early work by Howard (Howard, 1960, 1966). Within the resulting literature, surveyed in Monahan (1982); Lovejoy (1991), such sequential learning problems were called Partially Observable Markov Decision Processes (POMDP). This work on POMDPs was continued in the artificial intelligence and reinforcement learning communities, (see, e.g., Cassandra et al. (1995); Kaelbling et al. (1998)), where the emphasis is on general purpose complexity results and algorithmic strategies that apply to the whole spectrum of POMDPs. Specific applications tend to come from robotics, game playing, and other problems that are thought to be exemplars for the problems that humans face when interacting in a general way with their environment. More recently, these problems have also been called Bayes

Adaptive Markov Decision Processes and optimal learning problems (Duff, 2002; Powell and Frazier, 2008). Within statistics, a closely related field is sequential design of experiments (see, e.g., Berry and Fristedt (1985); DeGroot (1970); Wetherill and Glazebrook (1986)). Here, the emphasis is on rigorous treatment of a class of problems that tend to appear in laboratory and other experimental settings. This field is also closely related to sequential analysis (Siegmund, 1985). While much of the literature is written for a specialized research audience, the tutorial Powell and Frazier (2008) introduces this class of problems in an accessible and comprehensive way for the general OR audience.

The problems that we consider here are all *sequential*, by which we mean that data is processed as it is collected, and decisions are made based upon all available data. Acting sequentially allows adaptation and provides the possibility of more efficient action. This is in contrast to non-sequential problems, where decisions are made before observing any data at all. The terms "online" and "offline" are sometimes used instead of "sequential" and "non-sequential," although these terms can also refer to differences in reward structure (see section 2.1). Sequential methods require a more sophisticated analysis than do non-sequential methods, but often provides much better performance.

We begin by describing in detail one sequential learning problem, the Bernoulli ranking and selection (R&S) problem. This problem is simple enough to be described in detail here and to be solved explicitly using DP, but it also contains essential elements observed in all sequential learning problems. Thus it serves as an excellent example of this class of problems and the way in which DP can be applied toward their solution. We then expand our scope to describe several other problems from operations research and management science to which DP has been fruitfully applied.

1 The Ranking and Selection Problem

In this section we describe an important learning problem in detail, and show how DP may be used to solve it. This problem is the ranking and selection (R&S) problem, which has a long history beginning with the work of Bechhofer (Bechhofer, 1954; Bechhofer et al., 1968). Historically, this problem was most often considered in a non-Bayesian framework (see Kim and Nelson (2006) for a review), but the Bayesian formulation has grown more prominent in recent decades (see Chick (2006) for a review).

In the R&S problem, we consider the efficient use of experimentation (either simulation-based or physical) to determine which of several "alternatives" is best. As an example, suppose that we have several different inventory polices, and we would like to use computer simulation to determine which of these has the highest average profit. We are limited in the number of simulations we can perform by the amount of simulation time that we have available, and we would like to allocate this time to simulations of the different inventory policies. This allocation should maximize our ability to determine which of the inventory policies is best.

One approach is to run a few simulations of each alternative to roughly estimate which alternatives are likely to be among the best, and then concentrate most of our later simulation effort on just these alternatives. When done intelligently, this allows us to find the best alternative with many fewer samples than we could have by spreading our simulation budget equally across the alternatives. The essential question in R&S is how this allocation may be done as efficiently as possible. DP offers an answer.

We formulate the R&S problem formally by supposing that we have a limited budget N of samples to be allocated among k alternatives, and that associated with each alternative is a sampling distribution. The R&S literature often assumes that this sampling distribution is normal, but it will be easier to illustrate this problem if we assume that the sampling distribution is Bernoulli. Let θ_x be the parameter of the Bernoulli distribution for alternative x, so θ_x is the probability of observing a "success" from alternative x. This quantity θ_x is unknown, but we will be able to learn it through sampling. Our goal is to use sampling to find the alternative with the largest θ_x.

Time is indexed by n, from $n = 1$ up to $n = N$. At each time $n \leq N$ we choose an alternative $x_n \in \{1, \ldots, k\}$ to sample from among the full set of k alternatives. This choice may depend upon all

previously observed samples, $(x_m, y_m)_{m<n}$, where x_1 is chosen deterministically. We then observe a sample y_n whose distribution is

$$y_n \mid \theta, x_n \sim \text{Bernoulli}(\theta_{x_n}).$$

This sample is conditionally independent of all other x_m and y_m, $m < n$, given θ and x_n. At the final time N, we select an alternative x_*, and we receive a terminal reward θ_{x_*}. This is the reward that we receive when we implement the alternative x_* in the real world. The decision x_* may depend upon all previously observed samples, $(x_m, y_m)_{m\leq N}$. Given perfect information we would choose x_* to be equal to $\arg\max_x \theta_x$, but this information is unavailable. Instead, we must base our choice of x_* upon the sampling information acquired. The crux of the R&S problem is choosing the sampling decisions $x_1, \ldots, x_N$ in order to best support our final implementation choice x_*.

Formally, we define a policy π to be a collection of random variables $(x_n)_{1\leq n\leq N}$ and x_*. We define $\mathcal{F}_n$ to be the sigma-algebra generated by the random variables x_m, y_m, $m \leq n$. We enforce the way in which decisions may depend upon previous observations by requiring that $x_n \in \mathcal{F}_{n-1}$ for $n \geq 1$ and $x_* \in \mathcal{F}_N$. We define Π to be the set of all such policies π.

Momentarily fixing θ, the expected reward obtained by a policy π is

$$\mathbb{E}^\pi \left[\theta_{x_*} \mid \theta\right], \tag{1}$$

where $\mathbb{E}^\pi$ indicates the expectation taken with respect to policy π. When an expectation depends upon the policy we write $\mathbb{E}^\pi$, and when it does not we simply write $\mathbb{E}$. Although the only decision that appears in this expression is x_*, the information upon which x_* is based depends critically upon the sampling decisions x_n. Indeed, we will see that choosing x_* given the available information is relatively straightforward, and the difficulty in the R&S problem is choosing x_n, $n \leq N$, i.e., choosing which information to collect.

We would like to choose the policy π that maximizes this quantity (1). The difficulty is that (1) depends upon θ, which is unknown. If we knew θ, we could simply choose $x_* \in \arg\max_x \theta_x$ without sampling at all. Furthermore, different policies may be better for different values of θ. We need to combine the criteria (1) across multiple values of θ and find a policy π that does well according to this combined criterion. One method for combining (1) across multiple values of θ is the Bayesian method. In this method we suppose that we have a prior probability measure $\mathbb{P}_0$ on θ, and that we wish to maximize the expected reward received when θ is drawn at random according to $\mathbb{P}_0$. This expected reward is $\mathbb{E}\left[\mathbb{E}^\pi \left[\theta_{x_*} \mid \theta\right] \mid \theta \sim \mathbb{P}_0\right] = \mathbb{E}^\pi \left[\theta_{x_*}\right]$, and our goal in Bayesian R&S is to find a policy π that solves

$$\sup_{\pi\in\Pi} \mathbb{E}^\pi \left[\theta_{x_*}\right]. \tag{2}$$

The prior probability measure $\mathbb{P}_0$ may be interpreted in several ways. One may think of it as expressing our subjective belief about the relative importance of various possible configurations. This first interpretation is the one most commonly understood in Bayesian statistics. Secondly, the prior may be interpreted literally in certain limited cases where θ really has been drawn at random from a known probability distribution. This is the case when one is solving R&S problems on a repeated basis, as one might, for example, when calibrating a single simulation model each day to a new set of data. Third and finally, one may interpret the objective function $\mathbb{E}^\pi \left[\theta_{x_*}\right] = \int_{\mathbb{R}^k} \mathbb{E}^\pi \left[\theta_{x_*} \mid \theta\right] \mathbb{P}_0(d\theta)$ as a linear combination of the individual objectives $(\mathbb{E}^\pi \left[\theta_{x_*} \mid \theta\right])_{\theta\in\mathbb{R}^k}$, just as we often treat multi-objective optimization problems by optimizing a linear combination of the objectives. In this interpretation, the choice of $\mathbb{P}_0$ is simply the choice of which linear combination to optimize.

Given a prior distribution $\mathbb{P}_0$ and the information collected up to a time n, we have a posterior distribution $\mathbb{P}_n$. In particular, given a sequence of measurements $x_{1:n} = x_1, \ldots, x_n$ and observations $y_{1:n} = y_1, \ldots, y_n$, the posterior $\mathbb{P}_n$ is defined as the measure on $\mathbb{R}^k$, $\mathbb{P}_n \left\{\theta \in du\right\} = \mathbb{P}_0 \left\{\theta \in du \mid x_{1:n}, y_{1:n}\right\}$. We write $\mathbb{E}_n$ to indicate the expectation taken with respect to this posterior distribution. In general, the posterior may be calculated using Bayes' rule,

$$\mathbb{P}_n \left\{\theta \in du\right\} \propto \mathbb{P}\left\{y_{1:n} \mid x_{1:n}, \theta = u\right\} \mathbb{P}_0 \left\{\theta \in du\right\}. \tag{3}$$

We assume that the decisions x_n depend only upon the posterior. One can show that the supremum (2) is unchanged if Π is restricted to this class of policies.

To facilitate computation of the posterior, it is convenient to suppose that the prior has a certain parametric form. If θ_x has a $\text{Beta}(a_{0x}, b_{0x})$ distribution under the prior $\mathbb{P}_0$ and is independent of all θ_j, $j \neq x$, then a calculation beginning with (3) shows that θ_x is also beta-distributed under the posterior $\mathbb{P}_n$, and it remains independent of other θ_j. In particular, $\theta_x \sim \text{Beta}(a_{nx}, b_{nx})$ where $a_{nx} = a_{0x} + \sum_{m \leq n} \mathbf{1}_{\{x_m = x\}}(1 - y_m)$ is the sum of a_{0x} and the number of failures seen from alternative x, and $b_{nx} = b_{0x} + \sum_{m \leq n} \mathbf{1}_{\{x_m = x\}} y_m$ is the sum of b_{0x} and the number of successes seen from alternative x. We define

$$\mu_{nx} = \mathbb{E}_n [\theta_x] = b_{nx}/(a_{nx} + b_{nx})$$

to be the expectation of θ_x under the posterior at time n. The explicit expression $b_{nx}/(a_{nx}+b_{nx})$ follows from a standard computation of the expectation of a beta-distributed random variable (Gelman et al. (2004)).

Given this beta prior and its associated beta posteriors, we can simplify the objective function $\mathbb{E}^\pi [\theta_{x_*}]$. Using the tower property of conditional expectation and the $\mathcal{F}_N$-measurability of x_*, we write $\mathbb{E}^\pi [\theta_{x_*}] = \mathbb{E}^\pi [\mathbb{E}_N^\pi [\theta_{x_*}]] = \mathbb{E}^\pi [\mu_{Nx_*}]$. Thus, the stochastic control problem (2) may be rewritten

$$\sup_{\pi \in \Pi} \mathbb{E}^\pi [\theta_{x_*}] = \sup_{\pi \in \Pi} \mathbb{E}^\pi [\mu_{Nx_*}].$$

This problem can be solved using DP. Because $\mu_{Nx} = b_{Nx}/(a_{Nx} + b_{Nx})$ is a function of a_N and b_N, and because $(n, a_n, b_n)_n$ is a Markov process (we will see this below in greater detail), the state of the DP may be taken to be (n, a_n, b_n). We thus define a value function parameterized by n, a_n and b_n,

$$V(n, a_n, b_n) = \sup_{\pi} \mathbb{E}^\pi [\mu_{Nx_*} \mid a_n, b_n]. \tag{4}$$

At the final time $n = N$, the only decision left to make is x_*. This decision may depend upon all the information available at this time, and in particular it may depend upon μ_{Nx}, $x = 1, \ldots, k$. Thus, the optimal x_* is any satisfying $x_* \in \arg\max_x \mu_{Nx}$, and the value function at the final time is

$$V(N, a_N, b_N) = \max_x \mu_{Nx}. \tag{5}$$

Given this final value function, we may compute the value function at earlier times $n < N$ using Bellman's recursion,

$$V(n, a_n, b_n) = \max_x \mathbb{E} [V(n+1, a_{n+1}, b_{n+1}) \mid x_{n+1} = x, a_n, b_n] \quad \text{for } n < N. \tag{6}$$

To write this expectation explicitly we must consider the distribution of a_{n+1} and b_{n+1} given $x_n = x$, a_n and b_n. If we measure alternative x and observe a success ($y_{n+1} = 1$) then we increment b_{nx}. If we instead observe a failure ($y_{n+1} = 0$) then we increment a_{nx}. Thus, letting e_x be the vector of 0s with a single 1 at component x,

$$b_{n+1} = b_n + y_{n+1} e_x, \qquad a_{n+1} = a_n + (1 - y_{n+1}) e_x,$$

and the distribution of a_{n+1} and b_{n+1} is determined by the probability of success, $\mathbb{P}_n \{y_{n+1} = 1\}$. Given θ_x, we have $y_{n+1} = 1$ with probability θ_x, but θ_x is unknown. We do have a posterior distribution on θ_x, however, which allows us to write

$$\mathbb{P}_n \{y_{n+1} = 1\} = \mathbb{E}_n [y_{n+1}] = \mathbb{E}_n [\mathbb{E}_n [y_{n+1} \mid \theta]] = \mathbb{E}_n \left[\theta_{x_{n+1}}\right] = \mu_{n x_{n+1}}.$$

We then immediately have the following expression for the distribution of a_{n+1}, b_{n+1},

$$\mathbb{P}_n \{b_{n+1} = b_n + e_x, a_{n+1} = a_n \mid x_{n+1} = x\} = \mathbb{P}_n \{y_{n+1} = 1 \mid x_{n+1} = x\} = \mu_{nx}$$
$$\mathbb{P}_n \{a_{n+1} = a_n + e_x, b_{n+1} = b_n \mid x_{n+1} = x\} = \mathbb{P}_n \{y_{n+1} = 0 \mid x_{n+1} = x\} = 1 - \mu_{nx}.$$

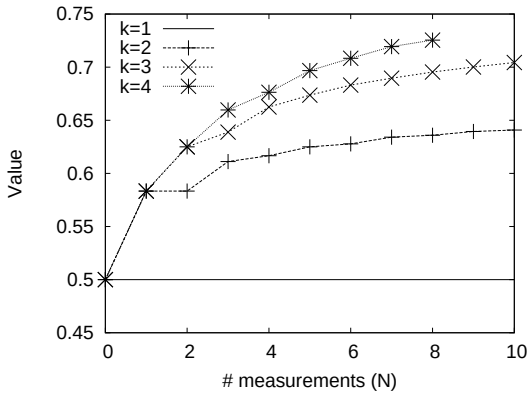

Figure 1: Value function at $n = 0$, with $a_{0x} = b_{0x} = 1$ for Bernoulli R&S as a function of the number of alternatives (k) and the measurement budget (N).

Thus, $(n, a_n, b_n)_{0 \le n \le N}$ is a Markov process whose distribution is determined by the measurement policy π. Note that we do not need to include θ into this stochastic process in order for its dynamics to be well-defined. This fact is mathematically quite natural but may seem counterintuitive. The Bayesian assumption of a prior on θ has allowed us to define the way in which our posterior belief on θ changes with measurements, without the need to know θ. This is what allows us to formulate the problem as a DP.

Given the conditional distribution of a_{n+1}, b_{n+1} we can simplify Bellman's recursion (6) to

$$V(n, a_n, b_n) = \max_x V(n+1, a_n + e_x, b_n)(1 - \mu_{nx}) + V(n+1, a_n, b_n + e_x)\mu_{nx}, \quad \text{for } n < N. \tag{7}$$

Using the terminal condition (5) and Bellman's recursion (7), one can calculate the value function for all possible states n, a_n and b_n. There are finitely many such states since $a_n - a_0$ and $b_n - b_0$ have non-negative integer components whose sum must equal n. The value function for $n = 0$ and $a_{0x} = b_{0x} = 1$ for all x is pictured in Figure 1 for problems with various values of k and N.

Given the value function as computed in this manner, we may compute an optimal policy as one that chooses its decisions to achieve the maximum in (7). That is, an optimal policy π^* is defined by

$$x_n \in \arg\max_x V(n+1, a_n + e_x, b_n)(1 - \mu_{nx}) + V(n+1, a_n, b_n + e_x)\mu_{nx}.$$

Although computing the value function using Bellman's recursion is straightforward in theory, it is quite difficult practically when the number of states is very large. In our problem, the number of states can be reduced by removing n from the state, since it is determined by summing the components of a_n and b_n. It can be further reduced by noting the invariance of the value function to permutations of the alternatives. Even with both space-reducing methods, however, the number of states is very large for even moderate values of k and N. This is the so-called "curse of dimensionality," and presents a great challenge to computation. Several approximation techniques have been proposed for addressing this difficulty (see, e.g., Powell (2007)), but the curse of dimensionality still presents a fundamental challenge to DP.

Despite the challenge presented by the curse of dimensionality, the DP solution to the Bernoulli R&S problem provides a number of benefits. First, it provides explicit solutions for small and moderately sized problems. Second, it provides a testing ground on which heuristic algorithms for larger problems may be benchmarked against optimal solutions. Third, it provides a solid theoretical understanding of the problem that may be extended to provide a number of insights into the nature of the problem and the behavior of the value function and associated optimal solution and (see Frazier et al. (2008) for similar insights in a sequential normal R&S problem). Software for computing several of the leading algorithms for Bayesian R&S of normal populations is available in the InfoCollection library at Frazier (2010).

2 Further Applications

In the previous section we saw how DP can be used to solve one particular sequential learning problem, the fully sequential Bernoulli R&S problem. This DP approach can be applied to nearly any learning problem in which there is some unknown truth (θ in the R&S problem), about which we are learning through our actions while simultaneously collecting rewards whose distribution depends on our actions and the truth. To formulate such problems as DPs, we create a state space consisting of the space of posterior beliefs, and the value function is defined using Bellman's recursion and the predictive distribution for the observation that follow from the current posterior distribution.

In this section we briefly describe other problems from the literature in which the DP formulation has been useful. In some of these problems, the DP has been solved explicitly. In others, the DP has suggested useful heuristics or provided insights into the structure of the problem. In addition to the problems detailed below, many other sequential learning problems have been approached using DP. These include sequential change detection (Dayanik et al., 2008; Hadjiliadis and Poor, 2008), dynamic pricing (Araman and Caldenty, 2010), medical diagnosis (Kapoor and Greiner, 2005), and many others (see Frazier (2009) Section 1.1 for a list).

2.1 Multi-armed Bandits

The multi-armed bandit problem is similar to the R&S problem, except that we receive each y_n as a *reward* rather than merely as an observation. The goal is to maximize the expected discounted total reward received over time,

$$\sup_{\pi\in\Pi} \mathbb{E}^{\pi}\left[\sum_{n=1}^{N}\gamma^{n-1}y_n\right],$$

where $\gamma \leq 1$ is a discount factor and $N \leq \infty$.

The name "multi-armed bandit" originates in a colorful example in which each alternative represents one arm of a multi-armed slot-machine (slot machines are sometimes called "one-armed bandits"). Pulling arm $x_n \in \{1,\ldots,k\}$ results in a payoff whose distribution is fixed but unknown. These distributions can only be learned through experimentation. The goal in this problem is to maximize the expected total value of our earnings over time.

This playful example is a surrogate for a large number of problems in many different fields in which the only way to learn the quality of some action is to actually perform it in the real world and suffer its corresponding real loss or reward. The optimal behavior in such problems is necessarily different and more cautious than in R&S, in which one has an experimental testbed available and does not suffer extreme consequences for testing poor alternatives. This difference in behavior due to reward structure is sometimes discussed by calling "offline" those reward structures similar to R&S, and calling "online" those similar to multi-armed bandits. This difference should not be conflated with the difference discussed in the introduction between sequential policies, which make their decisions x_n based upon all of the available data, and non-sequential policies, which make their decisions before seeing any data.

The multi-armed bandit problem can be solved using DP if we assume rewards are conditionally independent given the alternative measured, and if we take the infinite horizon discounted case ($\gamma < 1$ and $N = \infty$). This solution is known as the Gittins index policy, and was introduced in Gittins and Jones (1974). This policy is derived by considering a number of smaller problems, one for each alternative. In each smaller problem, we may pull only a single arm x, and we decide adaptively when to stop. Upon stopping, we are given a fixed retirement reward M. This can be modeled by requiring the policy to satisfy $x_n \in \{0, x\}$ for all n, and defining $\tau = \inf\{n \geq 0 : x_n = 0\}$ as the time at which the policy decides to retire. Let Π_x be the space of all such policies. The smaller problem then has a value function V parameterized by the arm x

being pulled and the retirement reward M,

$$V(P; M, x) = \sup_{\pi \in \Pi_x} \mathbb{E}_n^\pi \left[\sum_{n=1}^{\tau-1} \gamma^{n-1} y_n + \gamma^{\tau-1} M \mid \mathbb{P}_0 = P \right],$$

where P is a probability distribution on the sampling distribution from arm x. In many important cases, it can be parameterized with a few (often one or two) real numbers. We then define a *Gittins index*, $m(P, x)$, to be the smallest retirement reward M such that the optimal policy restricted to pull arm x retires immediately. This is written as

$$m(P, x) = \inf \{M : V(P; M, x) = M\}.$$

One can then show using DP that the optimal policy for the original multi-armed problem is to pull the arm with the largest Gittins index. One can refer to the original proof in Gittins and Jones (1974), but more accessible and equally rigorous proofs may be found in Whittle (1980) or Tsitsiklis (1994). The resulting optimal policy is

$$x_n \in \arg\max_x m(\mathbb{P}_n, x).$$

To use this optimal policy, we must be able to compute $m(\mathbb{P}_n, x)$, which requires solving and calculating value functions for the smaller problems. This is possible because the state space of this smaller problem is *much* smaller than for the full multi-armed bandit problem, consisting only of the posterior distribution on a single alternative. In many cases, including Bernoulli rewards with a beta prior on the probability of reward, and normal rewards with known variance and a normal prior on the sampling mean, this state space has only two dimensions. Such small dynamic programs can be solved numerically using standard techniques. Tables of Gittins indices for the beta-Bernoulli and normal-normal cases can be found in Gittins (1989), and an easy-to-use analytic approximation for the normal case may be found in Yao (2006). This is in contrast to the full problem under these priors, whose state space has $2k$ dimensions. Computation scales exponentially in the dimension, which makes the smaller problem eminently solvable while a brute-force numerical solution of the full problem is intractable.

If we relax the assumptions of the Gittins index policy, e.g., by taking a finite horizon or correlated y_n, then no tractable optimal solution is known. The proof of the Gittins index policy relies on a decomposition of the problem into single-arm subproblems, which is quite fragile to such changes in assumptions. To solve the problem optimally one must solve the full DP, which is computationally intractable because of its large state space. Although the optimal policy is unknown, knowledge of the underlying DP formulation and the Gittins index policy has supported the creation and analysis of a number of heuristics. Many of these heuristics are index-based solutions that are not optimal, but have pleasing empirical and theoretical properties. See, e.g., heuristics for restless bandit problems (Whittle, 1988) and other variants discussed in Section 3 of Mahajan and Teneketzis (2007). DP-motivated heuristics for bandits with finite-horizons and/or correlated rewards are discussed in Ryzhov et al. (2009); Ryzhov and Frazier (2010).

For complete treatments of the classic Bayesian bandit results, see Berry and Fristedt (1985), Wetherill and Glazebrook (1986), or Mahajan and Teneketzis (2007). See also the surveys (Bergemann and Valimaki, 2006) from the economics literature, (Szepesvári, 2009) from the computer science literature, or Chapter 10 of Powell (2007) from the operations research literature. There is also a large literature on non-Bayesian formulations of the multi-armed bandit problem in which worst-case regret is minimized. For an introduction, see Chapter 6 of Cesa-Bianchi and Lugosi (2006).

2.2 Bayesian Global Optimization

Suppose that we have a continuous function f that we would like to optimize over some compact domain $\mathcal{X} \subset \mathbb{R}^d$. The only access to the function is through a black box that will tell us the function's value f at points x_n of our choosing, but possibly obscured by noise. We suppose that we have no access to

derivative or other information. We further suppose that function evaluation takes a long time (minutes or hours), and so we wish to minimize the number of function evaluations we perform in order to find a "good" solution. Such problems occur quite frequently, for example when optimizing a complex simulation model, or optimizing temperature and pressure in an industrial chemical manufacturing process. Time-consuming function evaluation differentiates the problem from other derivative free global optimization problems for which the goal may be to minimize algorithmic computation time rather than the number of function evaluations, for which algorithms requiring less computation may be desirable.

We may formally pose the problem of finding an algorithm that finds as good a solution as possible in a bounded number of function evaluations as a Bayesian learning problem. This formulation shares quite a bit with the aforementioned ranking and selection problem, with the critical difference being that we are now searching on a continuous domain for the best point rather than among a finite set of alternatives.

As in the R&S problem we begin with a prior distribution on the truth, but here the prior is on the function f. Despite the large and complex nature of the space of all possible continuous functions on a compact domain, the class of Gaussian process priors form an analytically convenient class of prior distributions on this space. They have been successfully applied to problems in spatial statistics (Cressie, 1993), machine learning (Rasmussen and Williams, 2006), and computer experiments (Santner et al., 2003). A Gaussian process prior is parameterized by its mean function $\mu : \mathbb{R}^d \mapsto \mathbb{R}$ and its covariance function $\Sigma : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ (Σ is required to be a positive semidefinite function), and may be written $f \sim \mathrm{GP}(\mu, \Sigma)$. With this prior, the problem is stated as,

$$\sup_{\pi} \mathbb{E}^{\pi} \left[f(x_*) \right], \tag{8}$$

where as in R&S, x_* is chosen as well as possible based on the measurements (x_n, y_n), $n = 1, \ldots, N$.

The DP formulation of this problem then proceeds as it did in the R&S problem, but where we now have a posterior on the space of functions. One finite-dimensional parameterization of this posterior is simply the set of points and functions evaluations observed so far $(x_m, y_m)_{m \le n}$. The state space is then quite large, being continuous with a number of dimensions that grows with N. One may also discretize the domain into k points and work with the posterior only on these points. The state space then may be chosen to have a fixed size, but its dimension is still on the order of k^2. Under either parameterization, the high-dimensionality of the state space prevents computation in all but the simplest cases.

Despite the difficulty of solving the DP, several heuristics that work well empirically may be understood as DP-based approximations to the optimal policy. One such heuristic is the knowledge-gradient (KG) method, which defines

$$\mathrm{KG}_n(x) = \mathbb{E}_n \left[\sup_{x' \in \mathcal{X}} \mu_{n+1}(x') \mid x_{n+1} = x \right],$$

where $\mu_{n+1}(x') = \mathbb{E}_{n+1} \left[f(x') \right]$. The KG method then selects $x_n \in \arg\max_x \mathrm{KG}_n(x)$ as the next point to measure. This is the decision that would be optimal under the DP if we had only a single measurement left to make, i.e., if $n = N - 1$. In this sense, the KG method is a one-step lookahead policy. In general, computing the KG policy is a difficult computational problem. One approach is to discretize the domain, as in Frazier et al. (2009). For truly continuous problems, discretization is an approximation, while for problems with naturally integral decision variables interpolated by a continuous mean function, e.g., choosing the reorder point for an inventory policy, this is a natural modification of the BGO problem.

Another heuristic is the expected improvement (EI) method, which is most commonly formulated in the noise-free case. Although this method is often presented without describing the underlying stochastic optimization problem (8) and its DP framework, it can be understood as motivated by the DP. Consider the noise-free case and restrict our implementation decision to be from among those points we have already measured. To obtain the policy that is now optimal with one measurement remaining under this restriction, we must replace the supremum over the whole domain $\mathcal{X}$ in the definition of $\mathrm{KG}(x)$ by a maximum over only the points we have measured, $\{x_m : m \le n + 1\}$. This maximum may be written $\max(f(x_m) : m \le n+1) = \max(y_{n+1}, f_n^*)$ where we have used that $\mu_{n+1}(x_m) = f(x_m) = y_m$ for $m \le n + 1$ in the noise-free case, and

then defined $f_n^* = \max_{m \leq n} f(x_m)$ to be the value of the best point observed so far. Subtracting f_n^*, which does not depend on x_{n+1}, provides the expected improvement function

$$\mathrm{EI}(x) = \mathbb{E}_n \left[(y_{n+1} - f_n^*)^+ \mid x_{n+1} = x \right].$$

The EI method then recommends that we sample the alternative with the largest $\mathrm{EI}(x)$. This method was introduced in (Schonlau, 1997; Schonlau et al., 1998; Jones et al., 1998), building on previous work (Žilinskas, 1975; Mockus et al., 1978; Mockus, 1994). It was modified to allow measurements with stochastic noise in Huang et al. (2006), and to numerical noise in Forrester et al. (2006). For a complete general introduction, see Forrester et al. (2008), or for a brief overview see Section 6.3 of Santner et al. (2003).

Thus, both KG and EI methods may be seen as motivated by the DP. Both are one-step lookahead policies, but under different assumptions on the set of allowed implementation decisions. The KG method is more difficult to compute, but generalizes more naturally to noisy measurements, and was shown to perform better than EI-based methods in an empirical study Frazier et al. (2009).

A number of software packages have been written that can calculate Bayesian estimates of the function f using Gaussian process priors. One such software package is DACE (Design and Analysis of Computer Experiments) (Lophaven et al., 2002a,b). An up-to-date collection of other software packages is maintained at the website Rasmussen (2009), which also lists texts, articles, and other websites relevant to estimation with Gaussian processes.

There are also software packages for calculating the measurement decisions of DP-based optimization policies. Publicly available code associated with Forrester et al. (2008) implements a number of prediction methods and sampling algorithms. The SPACE software package (Schonlau, 2001) implements the Efficient Global Optimization (EGO) algorithm Jones et al. (1998) for noise-free optimization, and the commercial software package TOMLAB includes an implementation of the Sequential Kriging Optimization (SKO) algorithm of Huang et al. (2006) for the noisy measurement case. An implementation of the KG algorithm (Frazier et al., 2009) for discretized spaces is available in the MatlabKG package (Frazier, 2010).

2.3 Sequential Hypothesis Testing

In the sequential hypothesis testing problem, we learn which of two hypotheses is correct by observing a sequence of samples $y_1, y_2, \ldots$. Under one hypothesis, the samples are coming from a density f_0, and under the other hypothesis they are coming from a different density f_1. Let $\theta \in \{0, 1\}$ be the correct hypothesis, so f_θ is the true sampling density. We pay a cost c for each sample observed and we may stop observing samples at any time. Let τ be the number of samples observed, and we require that it be a stopping time of the filtration generated by $y_1, y_2, \ldots$. Upon stopping, we choose a hypothesis x_* and suffer a loss which is 0 if $\theta = x_*$ and equal to d_θ otherwise. In the Bayesian version of the problem, we begin with a prior probability that $\theta = 1$, and we seek a stopping time τ that minimizes the expected loss $\mathbb{E}\left[c\tau + d_\theta \mathbf{1}_{\{\theta \neq x_*\}}\right]$.

This problem was introduced and solved in Wald and Wolfowitz (1948), where the solution was shown to be of the form

$$\begin{aligned} \tau &= \inf \{t \geq 0 : L_t \notin [A, B]\}, \\ x_* &= \mathbf{1}_{\{L_t \geq 0\}}, \end{aligned}$$

for some levels A and B, where $L_t = \sum_{s \leq t} \log(f_1(y_s)) - \log(f_0(y_s))$ is the log-likelihood at time t. A policy of this form is called a sequential probability ratio test (SPRT).

Wald and Wolfowitz (1948) shows that the SPRT is also optimal in another, non-Bayesian, sense. For any given $\alpha_0 \in (0, 1)$ and $\alpha_1 \in (0, 1)$, the expected number of samples under both hypotheses, $\mathbb{E}[\tau \mid \theta = 0]$ and $\mathbb{E}[\tau \mid \theta = 1]$, is minimal among all policies satisfying $\mathbb{P}^\pi \{x_* \neq \theta \mid \theta = 0\} \leq \alpha_0$ and $\mathbb{P}^\pi \{x_* \neq \theta \mid \theta = 1\} \leq \alpha_1$. This form of optimality is surprisingly strong because it simultaneously minimizes the expected number of samples taken under both hypotheses. It is known as Neyman-Pearson optimality because of its similarity to Neyman-Pearson optimality for non-sequential statistical tests (see, e.g., Bickel and Doksum (2007)). A

modern and accessible treatment of the results of Wald and Wolfowitz (1948) may be found in Poor (1994). Also see Lai (2001) for a detailed survey of this and many related problems is sequential analysis.

The original motivation for this problem was in reducing the number of rounds of ammunition required to test anti-aircraft guns and ordnance during World War II (Wallis, 1980), but it or problems similar to it appear whenever we wish to take samples in order to make a decision. One class of similar problems of particular importance appear in the design of clinical trials (Jennison and Turnbull, 2000).

2.4 Inventory Control

In the classic newsvendor problem (see, e.g., Porteus (2002)), we have daily demand for a perishable product, e.g., newspapers, and we wish to maximize our expected revenue. Each day we choose a number of units x_n of inventory to buy at a per-unit cost c, and we sell units at a per-unit price of p until we either run out of inventory or meet the days' demand D_n.

If the demand distribution is known, then the classic analysis easily characterizes the optimal solution. In many problems, however, we do not know the true demand distribution but are instead learning it from sales data. This is particularly true with a new product, or with a product for which the demand distribution is changing over time. When the demand distribution is unknown but fixed, the Bayesian version of this problem may be written

$$\sup_{\pi} \mathbb{E}^{\pi} \left[\sum_{n=0}^{N} -cx_n + p \min(D_n, x_n) \right]$$

where the demand distribution F is unknown and $D_n \mid F$ come independently and identically distributed from F. This problem may be solved using DP, where the state of the DP is a parameterization of the posterior belief on F at the current time.

If we observe all of the demand, even if we are unable to meet it, then the DP simplifies and the optimal stocking decision is the same as the greedy stocking decision, $\arg\max_x \mathbb{E}_n \left[-cx_n + p \min(D_n, x_n) \right]$. If instead observation of demand is censored when we stock out of product, as it often is in practice, then the solution to the DP is more complicated. Ding et al. (2002) use the DP formulation to show that the optimal policy in the presence of censoring orders at least as much as the greedy policy, and possibly more. An easily computed solution is provided in Lariviere and Porteus (1999) for cases with either exponential demand with unknown rate, or Weibull demand with known shape but unknown mean. In these cases, a DP with a two-dimensional state space may be solved using a one-dimensional calculation.

3 Conclusion

We have shown how DP can be used to analyze a number of different learning problems from operations research and management science: ranking and selection, multi-armed bandits, global optimization, sequential hypothesis testing, and inventory control. These are just a few of the interesting and important problems in sequential learning that can be approached using DP. In some cases, the DP formulation provides an explicit solution, while in others it provides useful heuristics and insight into the problem and the behavior of the optimal policy. Many other important problems remain that we believe can be profitably approached with DP, and we hope that this article inspires readers to apply DP methods to these problems in the future.

References

Araman, V. and Caldenty, R. (2010). Dynamic pricing for perishable products with demand learning. *Operations Research*. to appear. 6

Bechhofer, R. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *The Annals of Mathematical Statistics*, 25(1):16–39. 2

Bechhofer, R., Kiefer, J., and Sobel, M. (1968). *Sequential Identification and Ranking Procedures*. University of Chicago Press, Chicago. 2

Bellman, R. (1954). The theory of dynamic programming. *Bull. Amer. Math. Soc.*, 60:503–516. 1

Bergemann, D. and Valimaki, J. (2006). Bandit problems. Cowles Foundation Discussion Paper 1551, Yale University. 7

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer-Verlag, New York, second edition. 1

Berry, D. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. Routledge. 2, 7

Bickel, P. and Doksum, K. (2007). *Mathematical statistics: basic ideas and selected topics*. Prentice Hall, 2nd ed. updated printing edition. 10

Cassandra, A., Kaelbling, L., and Littman, M. (1995). Acting optimally in partially observable stochastic domains. In *Proceedings of the National Conference on Artificial Intelligence*, pages 1023–1023. Citeseer. 1

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press. 7

Chick, S. (2006). Bayesian ideas and discrete event simulation: why, what and how. In *Proceedings of the 37th conference on Winter simulation*, pages 96–105. Winter Simulation Conference. 2

Cressie, N. (1993). *Statistics for Spatial Data, revised edition*. Wiley Interscience. 8

Dayanik, S., Goulding, C., and Poor, H. (2008). Bayesian sequential change diagnosis. *Mathematics of Operations Research*, 33(2):475–496. 6

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. John Wiley and Sons. 2

Ding, X., Puterman, M., and Bisi, A. (2002). The Censored Newsvendor and the Optimal Acquisition of Information. *Operations Research*, 50(3):517–527. 1, 10

Duff, M. (2002). *Optimal Learning: Computational Procedures for Bayes-Adaptive Markov Decision Processes*. PhD thesis, University of Massachusetts Amherst. 2

Forrester, A., Keane, A., and Bressloff, N. (2006). Design and Analysis of" Noisy" Computer Experiments. *AIAA Journal*, 44(10):2331–2339. 9

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. Wiley. 9

Frazier, P. (2009). *Knowledge-Gradient Methods for Statistical Learning*. PhD thesis, Princeton University. 1, 6

Frazier, P. (2009–2010). http://people.orie.cornell.edu/pfrazier/src.html. 6, 9

Frazier, P., Powell, W. B., and Dayanik, S. (2008). A knowledge gradient policy for sequential information collection. *SIAM J. on Control and Optimization*, 47(5):2410–2439. 6

Frazier, P., Powell, W. B., and Dayanik, S. (2009). The knowledge gradient policy for correlated normal beliefs. *INFORMS J. on Computing*, 21(4):599–613. 8, 9

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian data analysis*. CRC Press, second edition. 4

Gittins, J. (1989). *Multi-Armed Bandit Allocation Indices*. John Wiley and Sons, New York. 7

Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In Gani, J., editor, *Progress in Statistics*, pages 241–266. 7

Hadjiliadis, O. and Poor, H. (2008). *Quickest Detection*. Cambridge Univ Press. 6

Howard, R. (1960). *Dynamic programming and Markov process*. MIT press. 1

Howard, R. (1966). Information Value Theory. *Systems Science and Cybernetics, IEEE Transactions on*, 2(1):22–26. 1

Huang, D., Allen, T., Notz, W., and Miller, R. (2006). Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32(5):369–382. 9

Jennison, C. and Turnbull, B. (2000). *Group sequential methods with applications to clinical trials*. CRC Press. 10

Jones, D., Schonlau, M., and Welch, W. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492. 9

Kaelbling, L., Littman, M., and Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134. 1

Kapoor, A. and Greiner, R. (2005). Learning and Classifying Under Hard Budgets. In *16th European Conference on Machine Learning, Porto, Portugal, October 3-7, 2005*. Springer-Verlag New York Inc. 6

Kim, S. and Nelson, B. (2006). *Handbook in Operations Research and Management Science: Simulation*, chapter Selecting the best system. Elsevier. 2

Lai, T. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica*, 11(2):303–408. 10

Lariviere, M. and Porteus, E. (1999). Stalking Information: Bayesian Inventory Management with Unobserved Lost Sales. *Management Science*, 45(3):346–363. 1, 10

Lophaven, S., Nielsen, H., and Søndergaard, J. (2002a). Aspects of the Matlab toolbox DACE. Technical Report IMM-REP-2002-13, Technical University of Denmark. 9

Lophaven, S., Nielsen, H., and Søndergaard, J. (2002b). Dace-a matlab kriging toolbox. Technical Report IMM-TR-2002-13, Technical University of Denmark. 9

Lovejoy, W. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28(1):47–65. 1

Mahajan, A. and Teneketzis, D. (2007). Multi-armed bandit problems. In A. O. Hero III, D. A. Castanon, D. C. and Kastella, K., editors, *Foundations and Applications of Sensor Management*. Springer-Verlag. 7

Mockus, J. (1994). Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365. 9

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In Dixon, L. and Szego, G., editors, *Towards Global Optimisation*, volume 2, pages 117–129. Elsevier Science Ltd., North Holland, Amsterdam. 9

Monahan, G. (1982). A survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, pages 1–16. 1

Poor, H. (1994). *An Introduction to Signal Detection and Estimation*. Springer Verlag. 10

Porteus, E. (2002). *Foundations of stochastic inventory theory*. Stanford Business Books. 10

Powell, W. and Frazier, P. (2008). Optimal Learning. *TutORials in Operations Research: State-of-the-Art Decision-Making Tools in the Information-Intensive Age*. 2

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the curses of dimensionality.* John Wiley and Sons, New York. 1, 5, 7

Rasmussen, C. (2009). The gaussian process website. http://www.gaussianprocess.org/. accessed Apr 16, 2010. 9

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning.* MIT Press. 8

Ryzhov, I. and Frazier, P. (2010). On the robustness of a one-period look-ahead policy in multi-armed bandit problems. submitted. 7

Ryzhov, I., Powell, W., and Frazier, P. (2009). The knowledge gradient algorithm for a general class of online learning problems. submitted. 7

Santner, T., Willians, B. W., and Notz, W. (2003). *The Design and Analysis of Computer Experiments.* Springer. 8, 9

Schonlau, M. (1997). *Computer experiments and global optimization.* PhD thesis, University of Waterloo. 9

Schonlau, M. (2001). *Manual: Space (Stochastic Process Analysis of Computer Experiments).* http://www.schonlau.net/space.html. 9

Schonlau, M., Welch, W., and Jones, D. (1998). *New Developments and Applications in Experimental Design*, volume 34, chapter Global versus local search in constrained optimization of computer models, pages 11–25. Institute of Mathematical Statistics. 9

Siegmund, D. (1985). *Sequential analysis: tests and confidence intervals.* Springer Verlag. 2

Szepesvári, C. (2009). Reinforcement learning algorithms for mdps. http://webdocs.cs.ualberta.ca/ szepesva/. 7

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202. 7

Žilinskas, A. (1975). Single-step Bayesian search method for an extremum of functions of a single variable. *Cybernetics and Systems Analysis*, 11(1):160–166. 9

Wald, A. and Wolfowitz, J. (1948). Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics*, 19(3):326–339. 1, 9, 10

Wallis, W. (1980). The statistical research group, 1942-1945. *Journal of the American Statistical Association*, pages 320–330. 10

Wetherill, G. and Glazebrook, K. (1986). *Sequential Methods in Statistics.* CRC Press. 2, 7

Whittle, P. (1980). Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149. 7

Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, pages 287–298. 7

Yao, Y. (2006). Some results on the Gittins index for a normal reward process. *Lecture Notes-Monograph Series*, pages 284–294. 7