Proceedings of the 2010 Winter Simulation Conference
B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, eds.

OPTIMAL LEARNING OF TRANSITION PROBABILITIES IN THE TWO-AGENT NEWSVENDOR PROBLEM

Ilya O. Ryzhov

Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540 USA

Martin R. Valdez-Vivas

Management Science and Engineering
Stanford University
Stanford, CA 94305 USA

Warren B. Powell

Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08540 USA

ABSTRACT

We examine a newsvendor problem with two agents: a requesting agent that observes private demand information, and an oversight agent that must determine how to allocate resources upon receiving a bid from the requesting agent. Because the two agents have different cost structures, the requesting agent tends to bid higher than the amount that is actually needed. As a result, the allocating agent needs to adaptively learn how to interpret the bids and estimate the requesting agent's biases. Learning must occur as quickly as possible, because each suboptimal resource allocation incurs an economic cost. We present a mathematical model that casts the problem as a Markov decision process with unknown transition probabilities. We then perform a simulation study comparing four different techniques for optimal learning of transition probabilities. The best technique is shown to be a knowledge gradient algorithm, based on a one-period look-ahead approach.

1 INTRODUCTION

Consider a game in which two players with newsvendor payoffs, a requesting agent and an oversight agent, have to coordinate to meet an uncertain demand. This arrangement occurs in collaborative settings, where members of a common organization have to make a joint resource allocation decision to satisfy consumer demand. More often than not, however, the adverse effects of over- or underestimating this demand differ for the two parties, a fact that translates into different underage and overage cost structures between them. Consider the following scenarios:

- The marketing branch of a business requests a budget for an upcoming advertising campaign. Underfunding the project may result in an ineffective campaign, whereas overfunding draws monetary resources away from other important projects within the organization.
- A project manager within a consulting firm coordinates with other managers to assemble a team for a given project. An understaffed project may be subject to unnecessary delays and missed deadlines, while overstaffing projects can create coordination problems and force the company to take on fewer projects.
- The IT department of a company requests a timeframe for completing a programming assignment. The IT department faces a penalty for not finishing on time, but an estimate that is too long will unduly stall the company's objectives.

The requesting agent is usually endowed with access to better information about the demand, and has an incentive to engage in opportunistic behaviour, submitting misleading allocation requests that exaggerate the difference between the optimal allocation quantities for the two agents. However, the oversight agent can, through repeated play and observation, learn the requesting agent's behavioral patterns using standard Bayesian updating techniques, eventually gaining the ability to account for the bias in the requests. The oversight agent will also have a strong interest in learning this biasing behaviour quickly, as underage and overage costs accumulate over time.

From the point of view of the oversight agent, the problem can be modeled using a Markov decision process (see Puterman 1994 for a definitive overview of classical MDPs) with the added dimension of unknown transition probabilities. The state of the MDP is an aggregate summary of the past history of the game (for example, the last three requests submitted). The set of actions is the set of possible allocation quantities, and the reward process is given by the newsvendor payoff of the oversight agent. However, the probability of moving from one state to another

depends on the strategy of the requesting agent, which is unknown to the oversight agent. These transition probabilities must be estimated and improved over time as the game progresses. The oversight agent is thus faced with a classic exploration/exploitation tradeoff: it may be necessary to choose an action that appears to be sub-optimal, in the hope of collecting new information about the bias of the requesting agent that could help lower costs later on. The problem of making sequential decisions under an evolving belief structure is known as "optimal learning."

Despite a rich literature on single-agent newsvendor problems (see Petruzzi and Dada 1999 and Khouja 1999 for further references), few studies have considered coordination between two agents with newsvendor payoffs. Numerous studies on supply chain coordination analyze production decisions between retailers and manufacturers, and design contractual incentives to achieve optimal production levels (see e.g. Cachon 2003 for an overview). An alternate approach considers the effects of exchanging information on demand forecasts and production capabilities (see e.g. Chen 2003). However, these studies consider the perspective of a central planner optimizing costs over the entire supply chain, whereas the two-agent newsvendor problem deals with information acquisition and adaptive strategies on each player's aggregate costs. Information sharing does not consider the dimension of optimal learning.

Bayesian optimal learning has been widely studied in the context of simple problems such as ranking and selection (see Bechhofer, Santner, and Goldsman 1995 and Kim and Nelson 2006) and multi-armed bandits (see e.g. Gittins 1989 or Berry and Fristedt 1985). Optimal learning has also been studied in the context of the single-agent newsvendor problem, where it is necessary to learn an uncertain demand. Nahmias and Smith (1994) and Agrawal and Smith (1998) represent frequentist approaches. Bayesian methods for cases of fully observable demand can be found in Scarf (1959), Clark and Scarf (1960), and Azoury (1985). The case of censored demands is covered in Lariviere and Porteus (1999), Ding, Puterman, and Bisi (2002) and Bensoussan, Cakanyildirim, and Sethi (2007), and reviewed by Berk, Gürler, and Levine (2007). All of these studies focus on the issue of information collection, and do not involve a physical state variable.

An early approach to learning with a physical state (Bellman and Kalaba 1959) applied classical dynamic programming techniques after expanding the state variable to include the information about the transition probabilities (the "hyperstate" or "knowledge state"). However, the size of the state variable quickly becomes intractably large under this approach. The work by Cozzolino, Gonzalez-Zubieta, and Miller (1965) derives an optimal solution for a simple example with two physical states, but is also unable to handle larger problems. Similar approaches can be found in Martin (1967) and Satia and Lave (1973).

Duff and Barto (1996) applied concepts from the Bayesian optimal learning literature by placing Dirichlet priors on the transition probabilities, and viewing the problem of choosing an action as an instance of the multi-armed bandit problem. The resulting algorithm is only outlined, without much discussion of implementation. Later studies by Dearden, Friedman, and Russell (1998) and Mannor et al. (2007) place a Bayesian prior on the value function, whereas Dearden, Friedman, and Andre (1999) and Strens (2000) place the priors on the transition probabilities. Other Bayesian approaches have been considered in the literature on partially observable MDPs, with a detailed survey available in Ross et al. (2008).

We tackle the problem using the concept of knowledge gradients, which originally appeared in Gupta and Miescke (1996) as an approach to the ranking and selection problem. The knowledge gradient (KG) method chooses the action that maximizes the expected single-period improvement in the estimate of the optimal objective value. This approach was studied in greater detail by Frazier, Powell, and Dayanik (2008), and then extended to other classes of optimal learning problems, such as ranking and selection with correlated rewards (Frazier, Powell, and Dayanik 2009), ranking and selection with unknown measurement noise (Chick, Branke, and Schmidt 2010) and multi-armed bandits (Ryzhov, Powell, and Frazier 2009, Ryzhov and Powell 2009). The work by Ryzhov and Powell (2010) derives a KG policy for an offline learning problem where the objective is to solve a path-finding problem on a graph with unknown arc lengths. However, the physical structure of the graph only comes into play when solving the shortest-path problem, not when deciding what to measure.

In this paper, we propose a KG policy for the problem of learning on an MDP with unknown transition probabilities. We show how the expected single-period improvement can be computed exactly, and we also suggest how the computational cost of the policy can be reduced. We also give the implementation of the local bandit approximation method of Duff and Barto (1996), which was not done in the original paper. Finally, we present experimental results comparing KG to local bandit approximation and other methods.

2 MATHEMATICAL MODEL

Consider a demand process D that is unknown to both players. Further, suppose that the requesting agent can make a noisy observation of the demand at time n given by $\hat{D}^n \sim \mathcal{N}\left(D^n, \sigma^2\right)$. The variance σ^2 can be interpreted as the accuracy of the demand forecast made by the requesting agent.

2.1 Generation of the request

Let c_r^u and c_r^o be the underage and overage costs of the requesting agent. In the nth play of the game, the requesting agent's cost function for an allocation quantity $\tilde{x}^n$ is the standard newsvendor payoff:

$$C^R(D^n, \tilde{x}^n) = c_r^u (D^n - \tilde{x}^n)^+ + c_r^o (\tilde{x}^n - D^n)^+.$$

The requesting agent minimizes costs by ordering at the critical quartile (Arrow, Harris, and Marschak 1951), or

$$q^{n,*} = \hat{D}^n + \sigma\Phi^{-1}\left(\frac{c_r^u}{c_r^u + c_r^o}\right).$$

However, in most cases, the oversight agent is more conservative than the requesting agent, with a lower underage cost relative to overage cost. Letting c_a^u and c_a^o be the underage and overage costs of the oversight agent, we have $\frac{c_a^u}{c_a^o} < \frac{c_r^u}{c_r^o}$. From the point of view of the oversight agent, the optimal allocation quantity given $\hat{D}^n$ is

$$\tilde{x}^{n,*} = \hat{D}^n + \sigma\Phi^{-1}\left(\frac{c_a^u}{c_a^u + c_a^o}\right) < q^n. \tag{1}$$

Because the oversight agent prefers to allocate less than the requesting agent asks for, the latter may adopt a strategy of employing a bias or padding term β when making requests. Thus, the actual order quantity submitted by the requesting agent at time n is

$$Q^n = q^{n,*} + \beta^n. \tag{2}$$

In practice, the oversight agent does not see the observation $\hat{D}^n$ when making a decision, and only has access to the request Q^n. Then, (1) can be rewritten in terms of β^n as

$$\tilde{x}^{n,*} = Q^n + \sigma\left(\Phi^{-1}\left(\frac{c_a^u}{c_a^u + c_a^o}\right) - \Phi^{-1}\left(\frac{c_r^u}{c_r^u + c_r^o}\right)\right) - \beta^n.$$

We can express the cost incurred by the oversight agent in the nth play, given the decision $x^n \in \mathscr{X}$, is expressed in terms of the differential

$$x^n = \tilde{x}^n - \left(Q^n - \sigma\Phi^{-1}\left(\frac{c_r^u}{c_r^u + c_r^o}\right)\right) \tag{3}$$

that was allocated in relation to the request Q^n. The newsvendor cost for the oversight agent now depends on the bias, and is given by

$$C^O(S^n, x^n, \beta^n) = c_a^u[-(x^n + \beta^n)]^+ + c_a^o[x^n + \beta^n]^+. \tag{4}$$

With the representation in (4), we can restate the oversight agent's decision problem in terms of choosing the differential x^n rather than the allocation quantity $\tilde{x}^n$ itself. The request Q^n is now contained in the decision variable. By considering the differential (how much to over- or underfund the request) rather than the request itself, the bias becomes the only uncertain quantity in the problem.

We assume that the bias β^n is drawn from a finite set $\{b_1, ..., b_K\}$ where $b_K > ... > b_1$. The distribution of β^n depends on the degree of cooperation between agents. If the requesting agent expects full cooperation from the oversight agent, β^n will tend to be small; if the requesting agent believes that the oversight agent is likely to underfund the request, β^n will be more likely to be large. The requesting agent can infer the oversight agent's responsiveness from an aggregate expression of the past history of the moves made in the game up to time n, represented by a scalar

$$s^n = h^n(\beta^0, x^0, ..., \beta^{n-1}, x^{n-1}), \tag{5}$$

where h^n is a discretizing function mapping the previously observed costs into some finite set. For a given cost history s at time n, the bias β^n is a discrete random variable with probability mass function given by

$$P(\beta^n = b_k \,|\, s) = \rho_{s,k}.$$

If S is the number of possible cost histories, the biases can be completely characterized by an $S \times K$ matrix called ρ, where element $\rho_{s,k}$ is as given above. The bias process $(\beta^n)_{n=0}^{\infty}$ is thus Markovian with respect to the sigma-algebra $\mathscr{F}^n$ generated by the first n moves made by both agents.

2.2 Markov decision process model

Suppose that the matrix ρ is known to the oversight agent. This assumption is relaxed in Section 3, where we assume that the oversight agent holds imperfect beliefs about the biasing behaviour that can be improved through learning. We model the decision problem of the oversight agent using a Markov decision process.

The decision made by the oversight agent at time n depends on the history s^n of the game as given in (5). We thus refer to s^n as the *physical state* of the MDP. The decision itself is a quantity x^n, as defined in (3). We denote by $\mathscr{X}$ the set of all possible differentials available to the oversight agent, and assume that $\mathscr{X}$ is finite (e.g. a discretization of an interval). Once a decision is made, the next state is determined by a function $s^{n+1} = S^M(s^n, x^n, \beta^n)$ where S^M is determined by (5). The quantity β^n will not become known to the oversight agent until time $n+1$. However, we index it by n to reflect the fact that it is set by the requesting agent at time n.

The oversight agent must choose an allocation policy π to minimize the total discounted cost,

$$\inf_{\pi} \mathbb{E} \sum_{n=0}^{\infty} \gamma^n C^O\left(s^n, X^{\pi,n}(s^n), \beta^n\right), \tag{6}$$

where $X^{\pi,n}$ is a decision rule associated with the policy π that maps the state s^n to an action $X^{\pi,n}(s^n) \in \mathscr{X}$. The allocation policy that solves (6) can be found by solving Bellman's equation (see Puterman 1994) using value iteration, policy iteration, or linear programming methods.

2.3 A learning model for transition probabilities

Suppose now that the oversight agent does not know the biasing beliefs ρ exactly. Following the precedent of Silver (1963) and Martin (1967), we use Dirichlet priors to capture the oversight agent's uncertainty about ρ. Given a set of unknown transition probabilities $p_1, \ldots, p_K$ for K outcomes and a vector of parameters $\alpha = (\alpha_1, \ldots, \alpha_K)$ with all $\alpha_k \geq 0$, the Dirichlet density is given by

$$f(p_1, \ldots, p_K; \alpha_1, \ldots, \alpha_K) = \frac{\Gamma\left(\prod_{k=1}^K \alpha_k\right)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K p_k^{\alpha_k - 1}$$

for all $p_1, \ldots, p_K$ satisfying $p_1 + \ldots + p_K = 1$. The notation Γ refers to the gamma function. The resulting marginal estimate of each individual probability p_k is given by (DeGroot 1970)

$$\mathbb{E}(p_k) = \frac{\alpha_k}{\sum_{k'=1}^K \alpha_{k'}}.$$

A sample $\hat{p} \sim Dir(\alpha)$ can be generated (Gelman et al. 2004) by first drawing $A_k \sim Gamma(\alpha_k, 1)$ for $k = 1, \ldots, K$ and letting $\hat{p}_k = A_k / \sum_{k'=1}^K A_{k'}$.

Let $\rho_s = (\rho_{s,1}, \ldots, \rho_{s,K})$ denote the biasing behaviour for a given cost history. This vector is unknown to the oversight agent, but we assume that our beliefs about ρ_s follow a Dirichlet distribution with parameter vector $\alpha_s^0 = \left(\alpha_{s,1}^0, \ldots, \alpha_{s,K}^0\right)$, denoted by $\rho_s \sim Dir\left(\alpha_s^0\right)$. We assume that ρ_s and $\rho_{s'}$ are independent for $s \neq s'$. The collection $\alpha^0 = \left(\alpha_s^0 \mid s\right)$ is referred to as the *knowledge state*, to distinguish it from the physical state of the MDP.

Every time we choose an action x under a cost history s, we observe a random transition to a new state. The observation W_s^{n+1} is determined using the true, unknown bias probabilities ρ_s, and thus provides information that can be used to update our beliefs. The random transition can be viewed as a multinomial random variable with 1 trial (the single random transition) and K different categories (the possible values of the bias). The pmf of the random observation is given by $P\left(W_s^{n+1} = e_k\right) = \rho_{s,k}$, where e_k is a vector of zeros with 1 at position k. Note that the random transition depends on the bias β^n that was already chosen by the requesting agent before the action x^n is selected. Thus, W_s^{n+1} depends on s, but not on the choice of action out of s.

Each random transition changes our beliefs about the bias probabilities. Let α_s^n denote our beliefs about ρ_s at time n. If, at time n, the physical state is s^n, the updating equations for our beliefs are given by

$$\alpha_s^{n+1} = \begin{cases} \alpha_s^n + \hat{W}_s^{n+1} & \text{if } s = s^n \\ \alpha_s^n & \text{otherwise.} \end{cases} \tag{7}$$

Because the vectors ρ_s are independent, observing a random transition out of a state with history s only changes our beliefs about ρ_s. This change is made by incrementing the corresponding component of α_s^n by 1. A derivation of (7) can be found in DeGroot (1970).

In fact, given α^n and the game history s, we can compute the conditional distribution of α_s^{n+1}. Despite the complexity of the Dirichlet density, this conditional distribution has a simple form that will allow us to present a computable learning policy in Section 3. Proposition 1 gives this result.

Proposition 1. *Suppose that, at time n, the physical state of the MDP is s. The conditional distribution of α_s^{n+1} given α^n is discrete, with pmf*

$$\rho_{s,k}^n = P^n\left(\alpha_s^{n+1} = \alpha_s^n + e_k\right) = \frac{\alpha_{s,k}^n}{\sum_{k'=1}^K \alpha_{s,k'}^n}$$

where P^n denotes a conditional probability given α^n.

Proof: The value of α_s^{n+1} depends on the outcome of the random transition W_s^{n+1}. We know that W_s^{n+1} takes values e_k for $k = 1,\ldots,K$ and follows a multinomial distribution. Thus, we can apply the tower property of conditional expectation to write

$$\rho_{s,k}^n = \mathbb{E}^n\left(P^n\left(\alpha_s^{n+1} = \alpha_s^n + e_k \,|\, \rho\right)\right) = \mathbb{E}\left(\rho_{s,k} \,|\, \alpha^n\right) = \frac{\alpha_{s,k}^n}{\sum_{k'=1}^K \alpha_{s,k'}^n}.$$

The notation $\mathbb{E}^n$ denotes a conditional expectation given α^n.

3 LEARNING POLICIES

We describe three heuristics for optimal learning of transition probabilities in an MDP. The first is the pure exploitation policy, which follows a greedy strategy and does not consider the learning component of the problem when making decisions. The second is the local bandit approximation policy of Duff and Barto (1996), which attempts to reduce the problem to a multi-armed bandit problem. Finally, we apply the knowledge gradient concept of Gupta and Miescke (1996) and Frazier, Powell, and Dayanik (2008) to this setting.

3.1 Value iteration and pure exploitation

With the addition of learning into the problem, the policy π in (6) is a set of decision rules $X^{\pi,n}$ mapping a physical state s^n and a knowledge state α^n to a point in the action space $\mathscr{X}$. We will now give a simple example of such a learning policy.

Suppose that we have made n measurements and stopped learning entirely. We will continue to make transitions and collect rewards after time n, but we will not be allowed to use these transitions to update our beliefs, so $\alpha^{n'} = \alpha^n$ for all $n' > n$. Then, our best guess of the optimal policy is the policy produced by the classic value iteration algorithm (an overview of this and other classic MDP algorithms is available in Puterman 1994), assuming that the true transition probabilities are given by the time-n beliefs α^n. This procedure, which we refer to as α^n-value iteration, initializes $v^0(s) = 0$ for all possible physical states s, and iterates

$$v^m(s) = \min_{x\in\mathscr{X}} \sum_{k=1}^K \rho_{s,k}^n \left(C^O(s,x,b_k) + \gamma v^{m-1}\left(s^M(s,x,b_k)\right)\right) \tag{8}$$

for all s until $\max_s \left|v^m(s) - v^{m-1}(s)\right|$ is within some specified tolerance level, then returns the policy π_n that chooses actions by solving (8) using the final approximation v^{π_n} obtained from the procedure. If desired, value iteration can be replaced by any classic MDP algorithm.

The *pure exploitation* policy is defined to be the policy that makes a decision according to the policy π_n obtained from α^n-value iteration, but then proceeds to update the beliefs using (7), and uses α^{n+1}-value iteration to recompute the policy that seems to be the best under the new beliefs. The pure exploitation decision rule is

$$X^{Exp,n}(s^n,\alpha^n) = \arg\min_{x\in\mathscr{X}} \sum_{k=1}^K \rho_{s^n,k}^n \left[C^O(s^n,x,b_k) + \gamma v^{\pi_n}\left(S^M(s^n,x,b_k)\right)\right]. \tag{9}$$

3.2 Local bandit approximation

The local bandit approximation (LBA) policy by Duff and Barto (1996) removes the physical state from the problem, and uses techniques from the multi-armed bandit literature to make decisions. Suppose that we are at time n. We can use α^n-value iteration to obtain the policy π_n. If we fix this policy, the MDP reduces to a Markov chain Y_n whose

transition probabilities are given by

$$P\left(Y_{n+1} = S^M\left(s, X^{\pi_n}(s), b_k\right) | Y_n = s\right) = \rho^n_{s^n,k}.$$

When we make a random transition out of state s, we incur a cost of the form

$$f\left(s, S^M\left(s, X^{\pi_n}(s), b_k\right)\right) = C^O\left(s, X^{\pi_n}(s), b_k\right).$$

We can let

$$\tau_{s,s'} = \min\left\{n \geq 0 | Y_n = s', \quad Y_0 = s\right\}$$

be the number of transitions required for this Markov chain to reach state s' for the first time, given that it starts in state s. The quantities

$$\begin{aligned} R\left(s,s'\right) &= \mathbb{E}\left(\sum_{n=0}^{\tau_{s,s'}-1} \gamma^n f\left(Y_n, Y_{n+1}\right) | Y_0 = s\right) \\ T\left(s,s'\right) &= \mathbb{E}\left(\sum_{n=0}^{\tau_{s,s'}-1} \gamma^n | Y_0 = s\right) \end{aligned}$$

represent the expected discounted cost incurred and time elapsed up to time $\tau_{s,s'}$, and can be computed using first-transition analysis on Y_n.

Then, given that the MDP is in state s^n, define

$$g_x\left(s^n, \alpha^n\right) = \frac{\sum_{k=1}^K \rho^n_{s^n,k}\left[C^O\left(s^n, x, b_k\right) + \gamma R\left(S^M\left(s^n, x, b_k\right), s^n\right)\right]}{1 + \gamma \sum_{k=1}^K \rho^n_{s^n,k} T\left(S^M\left(s^n, x, b_k\right), s^n\right)}. \tag{10}$$

If we choose action x out of state s^n, we will transition to state $S^M\left(s^n, x, b_k\right)$ with a certain probability, after which we will follow the policy π_n. When we do so, we will make $\tau_{S^M(s^n,x,b_k),s^n}$ transitions before we return to the state s^n. Thus, (10) represents the expected total discounted cost per unit of expected discounted time that we receive during the sojourn between visits to s^n.

Equation (10) is analogous to the reward-per-unit-time representation of Gittins indices for multi-armed bandits. Gittins indices were first developed by Gittins and Jones (1974), and this particular representation is discussed e.g. by Duff (1995). Thus, $g_x\left(s^n, \alpha^n\right)$ is like a Gittins index for the action x, with the sojourn costs and times determined by policy π_n after the first transition. The LBA policy then makes the decision with the smallest cost-per-unit-time,

$$X^{LBA,n}\left(s^n, \alpha^n\right) = \arg\min_{x \in \mathscr{X}} g_x\left(s^n, \alpha^n\right). \tag{11}$$

If we were maximizing a reward instead of minimizing a cost, we would take the argmax in (11) rather than the argmin.

At each time step, LBA views the process as stateless; the downstream rewards are incorporated into the calculation of (10), and the problem reduces to a choice between $|\mathscr{X}|$ different reward processes, each of which leads back to the present location. To compute Gittins indices for all processes, we must first compute π_n, then solve $2 \cdot |\mathscr{X}|$ systems of $S \times S$ linear equations.

3.3 The knowledge gradient policy

We apply the knowledge gradient concept, studied by Gupta and Miescke (1996) and Frazier, Powell, and Dayanik (2008) originally for the ranking and selection problem, to the setting of MDPs with unknown transition probabilities. In Ryzhov and Powell (2009), this concept is stated as "choosing the measurement that would be optimal if it were the last chance to learn." We assume that we are in state s^n at time n, and the $(n+1)$st transition will be the last one to impact our beliefs. That is, $\alpha^{n'} = \alpha^{n+1}$ for $n' > n+1$. Then, we need to choose one action at time n, and we will switch to the policy π_{n+1} starting at time $n+1$.

Suppose that we make a decision x at time n. By Proposition 1, we know that $\alpha^{n+1}_{s^n}$ is discrete with K possible values, and $\alpha^{n+1}_{s'}$ is known for all $s' \neq s$. Let $\alpha^{n+1,k}_{s^n} = \alpha^n_{s^n} + e_k$ denote the kth possible value of $\alpha^{n+1}_{s^n}$. By Proposition 1, the conditional probability of this outcome given α^n is $\rho^n_{s^n,k}$.

For each outcome $k = 1, \ldots, K$, we can run $\alpha^{n+1,k}$-value iteration to obtain a vector $v^{\pi^k_{n+1}}$ representing the infinite-horizon value of being in each state, given that we see outcome k in the random transition out of s^n at time n. We can

take an expectation over all possible outcomes of this transition to obtain

$$\mathbb{E}^n_{s^n,x} v^{\pi_{n+1}} = \sum_{k=1}^{K} \rho^n_{s^n,k} v^{\pi^k_{n+1}} \left(S^M\left(s^n,x,b_k\right)\right), \tag{12}$$

where the expectation $\mathbb{E}^n_{s^n,x}$ is given α^n and the state-action pair (s^n,x). If we observe the kth outcome on the random transition out of s^n after making the decision x, it follows that $s^{n+1} = S^M(s^n,x,b_k)$. The time-$n+1$ infinite-horizon value of being in this state is then $v^{\pi^k_{n+1}}\left(S^M\left(s^n,x,b_k\right)\right)$.

The knowledge gradient policy chooses the optimal action to take at time n, under the assumption that we will stop learning at time $n+1$. We can do this by solving a modified version of Bellman's equation given by

$$X^{KG,n}(s^n,\alpha^n) = \arg\min_{x\in\mathscr{X}} \sum_{k=1}^{K} \rho^n_{s^n,k} \left[C^O\left(s^n,x,b_k\right) + \gamma v^{\pi^k_{n+1}}\left(S^M\left(s^n,x,b_k\right)\right)\right].$$

The KG decision rule can be rewritten as

$$\begin{aligned} X^{KG,n}(s^n,\alpha^n) = \arg\min_{x\in\mathscr{X}} \sum_{k=1}^{K} \rho^n_{s^n,k} C^O\left(s^n,x,b_k\right) &+ \gamma \sum_{k=1}^{K} \rho^n_{s^n,k} \left[v^{\pi^k_{n+1}}\left(S^M\left(s^n,x,b_k\right)\right) - v^{\pi_n}\left(S^M\left(s^n,x,b_k\right)\right)\right] \\ &+ \gamma \sum_{k=1}^{K} \rho^n_{s^n,k} v^{\pi_n}\left(S^M\left(s^n,x,b_k\right)\right). \end{aligned}$$

Define the *knowledge gradient* (KG) value of action x in state s^n to be

$$\begin{aligned} \nu^{KG,n}_{s^n,x} &= \mathbb{E}^n_{s^n,x} v^{\pi_n}\left(s^{n+1}\right) - v^{\pi_{n+1}}\left(s^{n+1}\right) \\ &= \sum_{k=1}^{K} \rho^n_{s^n,k} \left[v^{\pi_n}\left(S^M\left(s^n,x,b_k\right)\right) - v^{\pi^k_{n+1}}\left(S^M\left(s^n,x,b_k\right)\right)\right]. \end{aligned} \tag{13}$$

Then, the KG decision rule becomes

$$X^{KG,n}(s^n,\alpha^n) = \arg\min_{x\in\mathscr{X}} \left\{ \sum_{k=1}^{K} \rho^n_{s^n,k} \left[C^O\left(s^n,x,b_k\right) + \gamma v^{\pi_n}\left(S^M\left(s^n,x,b_k\right)\right)\right] \right\} - \gamma \nu^{KG,n}. \tag{14}$$

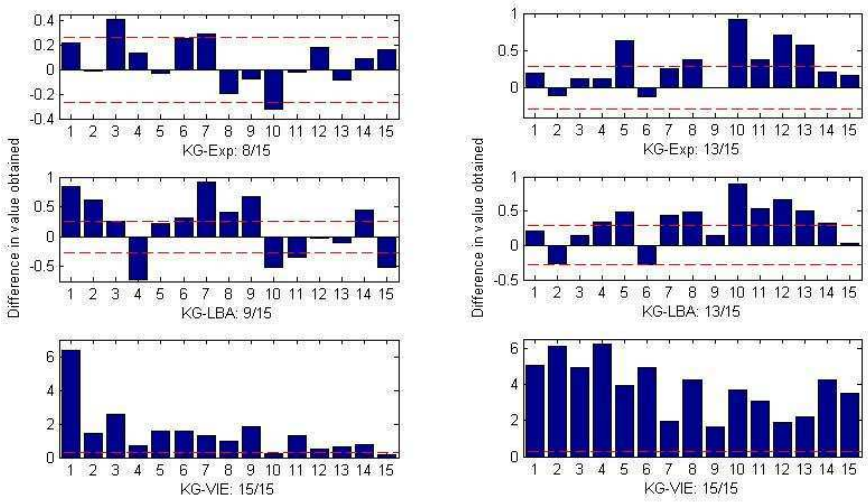
This expression is very similar to Bellman's equation for infinite-horizon MDPs, with one crucial difference. In addition to the one-period contribution function and the downstream reward, we also consider the expected improvement $\nu^{KG,n}_{s^n,x}$ in our estimate of the downstream cost that we obtain as a result of choosing action x. To put it another way, the KG factor $\nu^{KG,n}_{s^n,x}$ is a bonus for action x representing the value of the information that we can obtain by choosing this action out of state s^n at time n. Note that our algorithm computes this quantity exactly (unlike e.g. the value of perfect information quantity of Dearden, Friedman, and Andre 1999, which must be estimated using Monte Carlo sampling).

For each action available to us in state s^n, we must compute $v^{\pi^k_{n+1}}$ for all outcomes k in order to compute the KG factor using (13). Thus, it is necessary to solve $K\cdot|\mathscr{X}|$ value iteration problems in each time step. At the same time, the computational cost depends only on the size of the action space and the set of possible biases, not on the size of the space of all possible knowledge states. Furthermore, it is possible to substantially reduce the computational cost of the KG policy by "fast-starting" the value iteration algorithm in each time step. Instead of initializing $v^0 = 0$ when running (8), we can initialize value iteration at time $n+1$ with the output of value iteration at time n.

Table 1 shows the running times of the first three iterations of a MATLAB implementation of the KG policy for problems of various sizes. Because KG requires us to solve $K\cdot|\mathscr{X}|$ value iteration problems, we measure problem size in terms of this quantity. The first iteration (time $n=0$) is the most computationally expensive. Starting at $n=1$, however, we are able to fast-start value iteration with the results from the previous time step. We can observe a speedup by a factor of five or more between $n=0$ and $n=1$. This allows KG to handle problems whose size ranges in the hundreds of thousands, after the hurdle of the first time step.

4 SIMULATION STUDY

Two learning policies can be compared by running them on randomly generated MDPs, in which the true transition probabilities ρ are generated along with the other parameters. The policies are not allowed to see the true values ρ when making decisions. However, the random transition made by an MDP after an action has been chosen can be determined using the true probabilities. Thus, the discounted long-run reward collected by a policy is based on the



(a) (b)

Figure 1: Results for (a) 15 truth-from-prior problems, and (b) 15 equal-prior problems.

true probabilities, though the decisions themselves are based on the beliefs α^n. Our performance measure for a given problem is the difference in discounted long-run reward collected by two policies on that problem, averaged over 10^4 sample paths. These sample paths are divided into groups of 500 to obtain approximately normal estimates of the performance measure, which can then be used to obtain standard errors. It is necessary to generate many sample paths in order to make a decisive comparison of two policies. However, this substantially increases the computation time, and requires us to use small problems for our numerical experiments. If one were to use a policy in practice, one would only run it once, and the computation time would be closer to the numbers in Table 1.

We considered four policies: the KG policy from (14), pure exploitation policy from (9), the LBA policy from (11), and the Value of Information Exploration (VIE) policy of Dearden, Friedman, and Andre (1999). The VIE policy approximates the expectation of the true value function by taking Monte Carlo samples from the prior distribution. For every Monte Carlo sample, we then solve a value iteration problem. To ensure a fair comparison, we set the number of Monte Carlo samples in VIE equal to the number of value iteration problems solved by KG. Thus, both VIE and KG required approximately the same computational effort. We ran every policy for 30 iterations.

In our experiments, the history function h^n from (5) was defined in such a way as to discretize the cost incurred by the requesting agent in the most recent game,

$$\tilde{C}^R\left(x^{n-1},\beta^{n-1}\right)=c_r^u\left[-\left(x^{n-1}+\beta^{n-1}\right)\right]^+ + c_r^o\left[x^{n-1}+\beta^{n-1}\right]^+,$$

Table 1: Running time of KG policy with fast-starting.

Size ($K\cdot\lvert\mathcal{X}\rvert$)	$n=0$	$n=1$	$n=2$
20	0.068s	0.010s	0.019s
72	0.164s	0.057s	0.054s
272	0.613s	0.185s	0.170s
1056	2.688s	0.755s	0.610s
4160	15.432s	3.784s	1.977s
16512	2m 2s	26.131s	8.041s
65792	19m 31s	3m 49s	3m 42s
262656	4h 30m	46m 48s	43m 45s

into five levels. From the point of view of the requesting agent, it is natural to adopt less of a bias when this historical value is small, and more of a bias when it is large. Every problem we considered had randomly generated cost structures, all of which obeyed the property $\frac{c_a^u}{c_a^o} < \frac{c_r^u}{c_r^o}$.

Two types of problems were generated. In the *truth-from-prior* problems, the true transition probabilities ρ are generated from the prior α^0, as assumed by the model. This represents a setting where the prior beliefs are reasonably accurate, and provide useful starting information. The initial priors themselves were chosen in such a way as to reflect the intuitive biasing behaviour explained above. In the *equal-prior* problems, all elements of α^0 are set to 1, and the true probabilities ρ are chosen to reflect the intuitive behaviour. This represents a setting where we have no prior belief about which transitions are more likely. We generated 15 MDPs of each type. Each MDP had five states and four actions. The small problem size is dictated by the necessity of running each policy many times to obtain an accurate comparison.

Figures 1(a) and 1(b) show the simulation results for both problem types. The bars represent the difference in discounted long-run reward collected by the two policies in each comparison. Bars above zero represent problems where KG outperformed a competing policy. For example, "KG-Exp: 8/15" means that the KG policy outperformed pure exploitation on 8 out of 15 problems. The dotted lines represent ± 2 times the average standard error across all comparisons.

We see that KG is generally competitive against pure exploitation. KG loses on seven truth-from-prior problems, but only one of these losses is statistically significant. Similarly, of the eight problems where KG wins, two are statistically significant. In the equal-prior case, KG has more of an edge, achieving a statistically significant margin of victory on six out of fifteen problems, while never losing by a statistically significant amount. Pure exploitation yields good performance on truth-from-prior problems because they consider a setting in which the prior beliefs are accurate on average. Thus, the action that seems to be the best based on the current beliefs usually is the best. However, in the equal-prior case, the performance of pure exploitation suffers noticeably.

KG also performs competitively against LBA. While KG suffers four statistically significant losses in the truth-from-prior case, it achieves a statistically significant margin of victory on seven problems. On the equal-prior problems, KG outperforms LBA by a significant margin nine times, while being (barely) significantly outperformed only once.

The VIE policy is outperformed by KG in all of the experiments, almost always by a significant margin. This is especially noteworthy since both KG and VIE solve the same number of value iteration problems in every time step. However, KG consistently yields significantly better performance with the same computational effort.

We conclude that KG is competitive against a variety of learning policies under both types of priors. If the prior reveals enough information, a simple policy like pure exploitation tends to achieve good results. However, KG seems to have an especial advantage in a setting where the prior reveals little information about the problem. It is important to note that this is precisely the type of problem where it is important to learn well and quickly.

5 CONCLUSION

We have presented a stochastic model for the two-agent newsvendor problem from the point of view of the oversight agent. The problem is modeled using a Markov decision process in which the transition probabilities represent the biasing behaviour of the requesting agent. The oversight agent has the ability to adaptively learn this behaviour using a Bayesian model with Dirichlet priors and multinomial observations. We have proposed a policy for optimal learning in this setting, based on the knowledge gradient concept from the ranking and selection literature. This work is the first to apply this concept to a problem with a physical state. The KG policy is intuitive and easy to implement.

The computational complexity of the policy depends on the size of the physical state space, but does not grow over time. Like the Value of Information Exploration policy of Dearden et al. (1999), the KG policy solves a number of value iteration problems in order to choose an action. However, experimental results suggest that, when VIE and KG solve the same number of value iteration problems, KG consistently yields significantly better performance with the same computational effort. Our experiments also show that KG is competitive against other policies such as the local bandit approximation policy of Duff and Barto (1996).

While computational cost remains an issue in very large problems, one can use fast-starting for a large speed boost in all but the first iteration. With this modification, the algorithm can potentially handle problems with large state spaces, as long as the action space is not too large. Additional methods for decreasing running time are an interesting subject for future work. However, we believe that the KG approach as we have presented it offers an interesting new perspective on the problem of learning on an MDP with unknown transition probabilities.

ACKNOWLEDGMENTS

This research was supported in part by AFOSR contract FA9550-08-1-0195 through the Center for Dynamic Data Analysis.

REFERENCES

Agrawal, N., and S. Smith. 1998. Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics* 43 (6): 839–861.

Arrow, K., T. Harris, and T. Marschak. 1951. Optimal inventory policy. *Econometrica* 19 (3): 250–272.

Azoury, K. 1985. Bayes solution to dynamic inventory models under unknown demand distribution. *Management Science* 31 (9): 1150–1160.

Bechhofer, R., T. Santner, and D. Goldsman. 1995. *Design and analysis of experiments for statistical selection, screening and multiple comparisons*. New York: John Wiley and Sons.

Bellman, R., and R. Kalaba. 1959. On adaptive control processes. *IRE Transactions on Automatic Control* 4 (2): 1–9.

Bensoussan, A., M. Cakanyildirim, and S. Sethi. 2007. A multiperiod newsvendor problem with partially observed demand. *Mathematics of Operations Research* 32 (2): 322–344.

Berk, E., Ü. Gürler, and R. Levine. 2007. Bayesian demand updating in the lost sales newsvendor problem: A two-moment approximation. *European Journal of Operational Research* 182 (1): 256–281.

Berry, D. A., and B. Fristedt. 1985. *Bandit problems*. London: Chapman and Hall.

Cachon, G. 2003. Supply chain coordination with contracts. In *Handbooks of Operations Research and Management Science, vol. 11: Supply Chain Management*, ed. S. Graves and A. de Kok, 229–340. North-Holland Publishing, Amsterdam.

Chen, F. 2003. Information sharing and supply chain coordination. In *Handbooks of Operations Research and Management Science, vol. 11: Supply Chain Management*, ed. S. Graves and A. de Kok, 341–421. North-Holland Publishing, Amsterdam.

Chick, S., J. Branke, and C. Schmidt. 2010. Sequential Sampling to Myopically Maximize the Expected Value of Information. *INFORMS J. on Computing* 22 (1): 71–80.

Clark, A., and H. Scarf. 1960. Optimal policies for a multi-echelon inventory problem. *Management Science* 6 (4): 475–490.

Cozzolino, J., R. Gonzalez-Zubieta, and R. Miller. 1965. Markov decision processes with uncertain transition probabilities. Technical Report 11, Operations Research Center, MIT.

Dearden, R., N. Friedman, and D. Andre. 1999. Model-based Bayesian Exploration. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 150–159.

Dearden, R., N. Friedman, and S. Russell. 1998. Bayesian Q-learning. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, 761–768.

DeGroot, M. H. 1970. *Optimal Statistical Decisions*. Hoboken, NJ: John Wiley and Sons.

Ding, X., M. Puterman, and A. Bisi. 2002. The censored newsvendor and the optimal acquisition of information. *Operations Research* 50 (3): 517–527.

Duff, M. 1995. Q-learning for bandit problems. Technical Report 95-26, Department of Computer Science, U. Mass Amherst.

Duff, M., and A. Barto. 1996. Local bandit approximation for optimal learning problems. *Advances in Neural Information Processing Systems* 9:1019–1025.

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

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

Gelman, A., J. Carlin, H. Stern, and D. Rubin. 2004. *Bayesian data analysis*. CRC Press.

Gittins, J. 1989. *Multi-armed bandit allocation indices*. New York: John Wiley and Sons.

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

Gupta, S., and K. Miescke. 1996. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of statistical planning and inference* 54 (2): 229–244.

Khouja, M. 1999. The single-period (newsvendor) problem: literature review and suggestions for future research. *OMEGA-International Journal of Management Science* 27 (5): 537–553.

Kim, S., and B. Nelson. 2006. Selecting the best system. In *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, ed. S. Henderson and B. Nelson, 501–534. North-Holland Publishing, Amsterdam.

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

Mannor, S., D. Simester, P. Sun, and J. Tsitsiklis. 2007. Bias and variance approximation in value function estimates. *Management Science* 53 (2): 308–322.

Martin, J. 1967. *Bayesian Decision Problems and Markov Chains*. John Wiley and Sons.

Nahmias, S., and S. Smith. 1994. Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science* 40 (5): 582–596.

Petruzzi, N., and M. Dada. 1999. Pricing and the newsvendor problem: A review with extensions. *Operations Research* 47 (2): 183–194.

Puterman, M. L. 1994. *Markov decision processes*. New York: John Wiley & Sons.

Ross, S., J. Pineau, S. Paquet, and S. Brahim. 2008. Online planning algorithms for POMDPs. *Journal of Artificial Intelligence* 32:663–704.

Ryzhov, I. O., and W. B. Powell. 2009. The knowledge gradient algorithm for online subset selection. In *Proceedings of the 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Nashville, TN*, 137–144.

Ryzhov, I. O., and W. B. Powell. 2010. Information collection on a graph. *Operations Research (to appear)*.

Ryzhov, I. O., W. B. Powell, and P. I. Frazier. 2009. The knowledge gradient algorithm for a general class of online learning problems. *Submitted for publication*.

Satia, J., and R. Lave. 1973. Markov decision processes with imprecise transition probabilities. *Operations Research* 21 (3): 755–763.

Scarf, H. 1959. Bayes solutions of the statistical inventory problem. *The Annals of Mathematical Statistics* 30 (2): 490–508.

Silver, E. 1963. Markovian decision processes with uncertain transition probabilities or rewards. Technical Report 1, Operations Research Center, MIT.

Strens, M. 2000. A Bayesian framework for reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, 943–950.

AUTHOR BIOGRAPHIES

ILYA O. RYZHOV is a Ph.D. candidate in the Department of Operations Research and Financial Engineering at Princeton University. He received a B.S. degree in Computer Science and an M. Eng. degree in Operations Research and Industrial Engineering from Cornell University, as well as an M.Sc. degree in Management Science from Stanford University. His research focuses on developing algorithms for solving the exploration/exploitation dilemma in stochastic optimization problems, e.g. multi-stage resource allocation problems, using optimal learning techniques. His email address for these proceedings is <iryzhov@princeton.edu>.

MARTIN R. VALDEZ-VIVAS is a Ph.D. student in the Department of Management Science and Engineering at Stanford University. He received a B.S.E. in Operations Research and Financial Engineering from Princeton University, where he became interested in optimal learning over the course of his senior thesis research. Currently, he is interested in applications of optimal learning in economics and game theory. His email address for these proceedings is <mvv@stanford.edu>.

WARREN B. POWELL is a Professor in the Department of Operations Research and Financial Engineering at Princeton University, and director of CASTLE Laboratory (<www.castlelab.princeton.edu>). He has coauthored over 150 refereed publications in stochastic optimization, stochastic resource allocation and related applications. He is the author of the book *Approximate Dynamic Programming: Solving the curses of dimensionality*, published by John Wiley & Sons. Currently, he is involved in applications in energy, transportation, finance and homeland security. His email address for these proceedings is <powell@princeton.edu>.