Operations Research Letters 41 (2013) 477–481

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

State observation accuracy and finite-memory policy performance

Olga L. Ortiz [a,*], Alan L. Erera [b], Chelsea C. White III [b]

[a] Frito-Lay North America, Plano, TX, United States

[b] School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, United States

ARTICLE INFO

Article history:
Received 10 November 2011
Accepted 24 May 2013
Available online 7 June 2013

Keywords:
State observation accuracy
Finite-memory policy
Partially observed Markov decision process

ABSTRACT

Assuming the probability of an inaccurate state observation is sufficiently small, we present a simple procedure for determining whether a stationary finite-memory controller will improve or degrade system performance, given improved state observation quality. The intent of this result is to lend insight into whether or not to invest in improved state observation quality.

1. Introduction

In this paper, we address the question: when state observations are reasonably accurate, under what conditions will a finite-memory controller provide better system performance if given more accurate state observations? It is a common assumption that more accurate state observations will result in improved system performance. For example, it is typically assumed that more accurate inventory counts will reduce inventory holding costs and/or the profit loss due to stock outs. We show that this common assumption does not hold in general for finite-memory controllers. For example, there are situations when higher quality inventory counts will degrade expected system performance under a reasonable finite-memory controller, such as an order-up-to policy. Thus, the inventory manager, who may want to initially ask "will better performance due to a more accurate inventory count justify the investment needed to improve the quality of the count?", should first ask "are we sure that inventory system performance will improve if the inventory is more accurately counted?".

This latter question is the focus of this paper. Clearly, if improved state observation quality degrades system performance, it would not make sense to invest in improved state observation quality. However, if improved state observation quality improves system performance, then whether or not to invest in improved state observation quality may warrant further investigation and analysis, which would presumably depend on the value and cost of improving observation accuracy and on what other opportunities for capital investment might exist.

2. Outline and summary of results

We now present an outline of the paper, and summarize its main results, which together identify a simple procedure for determining whether improved state observation quality will improve or degrade system performance.

In Sections 4 and 5 we define the partially observed Markov decision process (POMDP) and present preliminary results. We then present a large class of policies for the POMDP, denoted the finite-memory policies, in Section 6. A finite-memory policy selects actions on the basis of the most recent history of state observations and actions. For a finite-memory controller, the expected total discounted reward to be accrued over the infinite horizon, v, is shown to be the inner product of the current state probability mass vector x and a vector γ, i.e., $v = x\gamma$, where γ only depends on the most recent history of state observations and actions.

We then show in Section 7 that γ can be represented as a power series in ϵ, where ϵ is the probability of a state observation error. All of the coefficient vectors α in the power series are easy to compute, relative to determining the expected total discounted reward-to-go for a given finite-memory policy for the case where state observation is perfect. Thus, for small ϵ, v can be approximated by $x(\alpha^0 + \epsilon\alpha^1)$, which implies that whether or not improved state observation quality improves system performance is dependent on the signs of the scalar elements of the vector α^1. We remark that for inventory systems, observation quality is typically described as a percent likelihood that an inventory count equals the actual inventory level. For example, if inventory count accuracy is 97.5%, then $\epsilon = 1.000 - 0.975 = 0.025$.

We also show that if γ is generated by an optimal (necessarily, zero-memory) policy for the completely observed counterpart of the POMDP, then all scalar elements of α^1 are non-positive; therefore, decreasing ϵ, i.e., improving state observation quality,

* Corresponding author.
E-mail addresses: olga.ortiz@pepsico.com (O.L. Ortiz), alerera@isye.gatech.edu (A.L. Erera), cwhite@isye.gatech.edu (C.C. White III).

0167-6377/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.orl.2013.06.001

does not decrease v, i.e., does not degrade system performance. This is essentially equivalent to a well-known result described in [17] and elsewhere. Further, we show that if the finite-memory controller does not depend on the most recent history of state observations and actions, then $\alpha^1 = 0$; hence, system performance is independent of state observation quality for history invariant controllers.

We then present two examples that illustrate the variety of possible behaviors a finite-memory controller can exhibit. In Section 8, we demonstrate that a near-optimal zero-memory controller can produce a vector α^1 with all positive scalar elements. Hence, as $\epsilon > 0$ gets smaller (i.e., the state observation accuracy improves) the expected total discounted reward over the infinite horizon generated by the controller also gets smaller. Thus, it is possible that what appears to be a good sub-optimal design can degrade system performance, given improved state observation quality. We also show that zero-memory policies exist where the vector α^1 includes both positive and negative scalar elements, indicating that the impact of improved state observation quality on system performance may depend on x.

3. Literature review

Our initial motivation for addressing the topic of this paper was inventory control. Inventory levels are usually observed accurately, but not perfectly. Our interest in finite-memory controllers is due to the fact that such controllers can serve as good, easily computed sub-optimal designs for the class of models of sequential decision making under uncertainty that serves as the basis for our analysis; optimal controllers are often impossible or impractical to deploy. We use the infinite horizon, total discounted reward POMDP as the basis for analysis since the POMDP, an extension of the (standard) MDP, can model noise corrupted, incomplete, or costly observations of the state process. The MDP assumes the system state is perfectly observed without cost and hence is an inadequate model for our investigation.

The superior modeling validity of the POMDP, relative to the MDP, is in contrast to the superior tractability of the MDP, relative to the POMDP. Although not the focus of the research reported in this paper, the determination of optimal or good sub-optimal policies for the POMDP has been a source of considerable interest. Computational procedures for determining an optimal policy for the POMDP can be found in [16,18,5,6,10,12]. Finite-memory approaches for the POMDP can be found in [19,11,13,1]. The finite-memory approach presented in [19] is the approach taken in this paper. More general results and discussion concerned with whether or not improved state observation will improve the performance of a given controller can be found in [3] on p. 192, and in [17].

In spite of the fact that inventory counts are rarely perfect, the MDP, rather than the POMDP, has served until recently as the basis for analysis for inventory control problems in part due to the aforementioned tractability of the MDP, relative to the POMDP. This tractability advantage is further amplified for inventory problems by the optimality of the computationally useful (s, S), continuous review (Q, R), and order-up-to policy structures for large classes of inventory control problems, policy structures that appear to have no counterparts for the POMDP.

There has been, however, growing recent interest in inventory control with inaccurate records. An excellent introduction to this area is provided in [9], from the viewpoint of attempting to understand the benefits that radio-frequency identification (RFID) technology may hold for supply chain inventory management. Please see [4,7,15,8], for related discussion. Numerical results presented in this literature indicate that effective policies can substantially reduce the costs of inventory count inaccuracies. Implicit in this literature is the intuitive (albeit false) assumption that improved inventory count invariably improves the performance of the inventory replenishment strategy.

4. Problem definition

Let $\{s(t), t = 0, 1, \ldots\}$, $\{z(t), t = 1, 2, \ldots\}$, and $\{a(t), t = 0, 1, \ldots\}$ be the *state*, *observation*, and *action* processes respectively. Assume that the state space S, the observation space Z, and the action space A are each finite and that these three processes are related by the given probabilities

$$p_{ij}(z, a) = P[z(t+1) = z, s(t+1) = j | s(t) = i, a(t) = a],$$

where $P(z, a) = \{p_{ij}(z, a)\}$ is a sub-stochastic matrix which is such that $\sum_z P(z, a)$ is stochastic. We assume the problem horizon is countably infinite and that for decision epoch $t \in \{0, 1, \ldots\}$, action $a(t)$ can be selected based on $d(t)$, where $d(0) = \{x(0)\}$ and for $t > 1$, $d(t) = \{z(t), \ldots, z(1), a(t-1), \ldots, a(0), x(0)\}$. Vector $x(t) = \{x_i(t)\}$ is a probability mass vector (pmv) over the state space S, where $x_i(t) = P[s(t) = i | d(t)]$; note that $x(0)$ is an *a priori* pmv. A *policy* at decision epoch t is a function $\delta_t : \{d(t)\} \to A$; a *strategy* is an ordered sequence of policies $\pi = \{\delta_t, t = 0, 1, \ldots\}$.

Let $r(i, a)$ be the reward accrued at decision epoch t, given $s(t) = i$ and $a(t) = a$. The criterion is

$$E_{x(0)} \left\{ \sum_{t=0}^{\infty} \beta^t r[s(t), a(t)] \right\},$$

where E_x is the expectation operator, conditioned on pmv x. In order to insure that the criterion is well defined, assume throughout that $\beta < 1$ (see [14]). The problem objective is to determine a strategy that maximizes the criterion.

5. Preliminary results

It is well known that $\{x(t), t = 0, 1, \ldots\}$ is a sufficient statistic for the POMDP [2]. Let $X = \{x \geq 0 : \sum_{i \in S} x_i = 1\}$, $\| \cdot \|$ be the supremum norm on X, and V be the set of all bounded, real-valued functions on X. Define $H_\delta : V \to V$ for a given policy δ and $H : V \to V$ as follows:

$$[H_\delta v](x) = xr[\delta(x)] + \beta \sum_z \sigma(z, x, \delta(x)) v[\lambda(z, x, \delta(x))],$$

$$[Hv](x) = \max_{a \in A} \left\{ xr(a) + \beta \sum_z \sigma(z, x, a) v[\lambda(z, x, a)] \right\},$$

where $y\mathbf{1} = \sum_i y_i$, $\sigma(z, x, a) = xP(z, a)\mathbf{1}$, and $\lambda(z, x, a) = xP(z, a)/\sigma(z, x, a)$ when $\sigma(z, x, a) \neq 0$. We remark that $x(t+1) = \lambda[z(t+1), x(t), a(t)]$ and $\sigma(z, x, a) = P(z(t+1) = z | x(t) = x, a(t) = a)$. We remark that $Hv = \sup_\delta H_\delta v$.

It is shown in [14] and elsewhere that H and H_δ are contraction operators on the Banach space $(V, \| \cdot \|)$. Thus, there exists a unique $v^* \in V$ such that $v^* = Hv^*$ and $\lim_{n \to \infty} \| v^n - v^* \| = 0$ for $v^{n+1} = Hv^n$, for $v^0 \in V$. (An analogous statement can be made about v^*_δ for each δ.) Further, if $H_\delta v^* = Hv^*$, then the (stationary) strategy $\pi = \{\delta, \delta, \ldots\}$ is an optimal strategy.

6. Truncated data sequences

We now define a finite-memory policy and present a key structural result for such a policy. For a given finite integer $m > 0$, let $\zeta_t^m = (z_t^m, a_{t-1}^m)$, where $z_t^m = \{z(t), \ldots, z(t-m+1)\}$ and $a_{t-1}^m = \{a(t-1), \ldots, a(t-m)\}$. Thus, ζ_t^m represents the present and most recent past observation and action selection data over m decision epochs, where m can be thought of as a design parameter. Let $\bar{\zeta}$ be such that $\zeta_{t+1}^m = \bar{\zeta}(z(t+1), \zeta_t^m, a(t))$. Hence, $\bar{\zeta}$ transforms ζ_t^m and $\{z(t+1), a(t)\}$ into ζ_{t+1}^m and thus is independent

of $\{z(t-m+1), a(t-m)\}$. For notational simplicity assume $\zeta = \zeta_t^m$ and define

$$P(\zeta) = P[z(t-m+1), a(t-m)] \times \cdots \times P[z(t), a(t-1)],$$
$$\sigma^m(\zeta, x) = xP(\zeta)\mathbf{1},$$
$$\lambda^m(\zeta, x) = \frac{xP(\zeta)}{\sigma^m(\zeta, x)}$$

if $\sigma^m(\zeta, x) \neq 0$. We remark that $(\zeta_t^m, x(t-m))$ is a sufficient statistic since $x(t) = \lambda^m[\zeta_t^m, x(t-m)]$. Any policy $\delta : \{\zeta_t^m\} \to A$ is said to be *finite-memory* policy. For stationary finite-memory strategy $\pi = \{\delta, \delta \cdots\}$, we note (with a slight abuse of notation) that

$$[H_\delta v](\zeta, x') = xr(\delta(\zeta)) + \beta \sum_z \sigma(z, x, \delta(\zeta))v \times [\overline{\zeta}(z, \zeta, \delta(\zeta)), \lambda(z', x', a')],$$

where $x' = x(t-m)$, $x = x(t) = \lambda(\zeta, x')$, $z' = z(t-m+1)$, $a' = a(t-m)$, and $\zeta = \zeta_t^m$. A simple induction argument, based on the fact that $\lim_{n\to\infty} \|v_\delta^n - v_\delta^*\| = 0$, implies the following preliminary result.

Lemma 1. *For each ζ, there exists a vector $\gamma(\zeta)$ such that $v_\delta^*(\zeta, x') = \lambda^m(\zeta, x')\gamma(\zeta)$, where γ is the (unique) solution of*

$$\gamma(\zeta) = r(\delta(\zeta)) + \beta \sum_z P(z, \delta(\zeta))\gamma(\overline{\zeta}(z, \zeta, \delta(\zeta)))$$

for stationary, finite-memory policy δ.

This result, and a concomitant policy improvement algorithm, can also be found in [5]. We remark that it seems intuitive that as m gets larger, the supremum of the expected total discounted reward accrued by finite-memory controllers would increase. The dependence on m for a finite-state controller is an interesting topic for future research.

7. Policy performance and observation quality

We now investigate the impact of state observation quality on finite-memory policy performance. Henceforth, assume $S = Z$, the observation probability $q(z|j, i, a) = P[z(t+1) = z | s(t+1) = j, s(t) = i, a(t) = a]$ is independent of i and a and

$$q(z|j) = \begin{cases} 1-\epsilon & \text{if } z = j \\ \sigma_{jz}\epsilon & \text{if } z \neq j, \end{cases}$$

where $\sigma_{jz} \geq 0$, $\sigma_{jj} = 0$, and $\sum_z \sigma_{jz} = 1$ for all j, and $\epsilon > 0$ represents the probability of an inaccurate state observation.

We remark that the stochastic matrix $\{q(z|j)\}$ represents better state observation quality if it depends on ϵ rather than ϵ' and $\epsilon < \epsilon'$. See [17] for another closely related description of state observation quality. The description and analysis of other definitions of state observation quality is a topic for future consideration.

Note that $p_{ij}(z, a) = q(z|j)p_{ij}(a)$, where $p_{ij}(a) = \sum_z p_{ij}(z, a) = P[s(t+1) = j | s(t) = i, a(t) = a]$ which is often referred to as the *transition* probability.

Intuitively, we would expect $v_\delta^*(\zeta, x')$ to increase (or at least not decrease) as ϵ gets smaller. We show below that this characteristic is not in general true but present conditions that guarantee it holds. We now show that for small ϵ, the expected total discounted reward for a finite-memory stationary strategy can be represented by a power series in ϵ.

Proposition 1. *Assume $\epsilon < \frac{(1-\beta)}{2\beta}$. Then, $\gamma(\zeta) = \sum_{l=0}^\infty \epsilon^l \alpha^l(\zeta)$, where:*

$$\alpha^0(i, \zeta) = r(i, \delta(\zeta)) + \beta \sum_j p_{ij}(\delta(\zeta))\alpha^0(j, \overline{\zeta}(j, \zeta, \delta(\zeta))),$$

$$\Delta^l(i, \zeta) = \beta \sum_j p_{ij}(\delta(\zeta)) \left[\sum_{z \neq j} \sigma_{jz} \alpha^l(j, \overline{\zeta} \times (z, \zeta, \delta(\zeta))) - \alpha^l(j, \overline{\zeta}(j, \zeta, \delta(\zeta))) \right], \quad l \geq 0$$

$$\alpha^l(i, \zeta) = \Delta^{l-1}(i, \zeta) + \beta \sum_j p_{ij}(\delta(\zeta))\alpha^l \times (j, \overline{\zeta}(j, \zeta, \delta(\zeta))), \quad l \geq 1.$$

Proof. By successive approximations. It follows from Lemma 1 that $\lim_{n\to\infty} \|\gamma_n(\zeta) - \gamma(\zeta)\| = 0$, where

$$\gamma_{n+1}(\zeta) = r(\delta(\zeta)) + \beta \sum_z P(z, \delta(\zeta))\gamma_n(\overline{\zeta}(z, \zeta, \delta(\zeta))),$$

and where $\gamma_0 = 0$. It follows from the definition of $q(z|j)$ that

$$\gamma_{n+1}(i, \zeta) = r(i, \delta(\zeta)) + \beta \sum_j p_{ij}(\delta(\zeta))\gamma_n \times (j, \overline{\zeta}(j, \zeta, \delta(\zeta))) + \epsilon \Delta_n(i, \zeta),$$

where

$$\Delta_n(i, \zeta) = \beta \sum_j p_{ij}(\delta(\zeta)) \left[\sum_{z \neq j} \sigma_{jz} \gamma_n(j, \overline{\zeta}(z, \zeta, \delta(\zeta))) - \gamma_n(j, \overline{\zeta}(j, \zeta, \delta(z))) \right].$$

It is then straightforward to show that $\gamma_n(i, \zeta) = \sum_{l=0}^n \epsilon^l \alpha_n^l(i, \zeta)$, where:

$$\alpha_{n+1}^0(i, \zeta) = r(i, \delta(\zeta)) + \beta \sum_j p_{ij}(\delta(\zeta))\alpha_n^0(j, \overline{\zeta}(j, \zeta, \delta(\zeta)))$$

$$\alpha_{n+1}^l(i, \zeta) = \Delta_n^{l-1}(i, \zeta) + \beta \sum_j p_{ij}(\delta(\zeta))\alpha_n^l(j, \overline{\zeta}(j, \zeta, \delta(\zeta)))$$

for $l = 1, \ldots, n$, where $\alpha_{n+1}^{n+1}(i, \zeta) = \Delta_n^n(i, \zeta)$, and for $l = 0, \ldots, n$,

$$\Delta_n^l(i, \zeta) = \beta \sum_j p_{ij}(\delta(\zeta)) \left[\sum_{z \neq j} \sigma_{jz} \alpha_n^l \times (j, \overline{\zeta}(z, \zeta, \delta(\zeta))) - \alpha_n^l(j, \overline{\zeta}(j, \zeta, \delta(\zeta))) \right].$$

Letting $n \to \infty$ gives the result, assuming $\lim_{n\to\infty} \sum_{l=0}^n \epsilon^l \alpha_n^l(i, \zeta)$ exists.

We now show that the infinite sum is well-defined if $\epsilon < \frac{(1-\beta)}{2\beta}$. Since $|r(i, a)| \leq M$ for all i and a, it follows that $\|\alpha^0\| \leq \frac{M}{1-\beta}$, where $\|\cdot\|$ is the supremum norm. A similar argument implies that for $l \geq 1$, $\|\alpha^l\| \leq \frac{\|\Delta^{l-1}\|}{1-\beta}$, where we note that $\|\Delta^{l-1}\| \leq 2\beta\|\alpha^{l-1}\|$. Thus, $\|\alpha^l\| \leq [\frac{2\beta}{1-\beta}]^l(\frac{M}{1-\beta})$. Convergence of the infinite series is then guaranteed if $\frac{2\beta\epsilon}{1-\beta} < 1$. □

We note that determining α^0 is computationally identical to computing the expected total discounted reward to be accrued over the infinite horizon generated by δ for the completely observed (i.e., MDP) case. Computing Δ^{l-1} and α^l, for each $l \geq 1$, has substantially more modest computational requirements.

Our numerical results thus far indicate that $\|\alpha^l\|$ approaches zero quickly as l grows large. Thus, $\frac{1-\beta}{2\beta}$ appears to be a very conservative upper bound on ϵ in order to guarantee that the limit $\lim_{n\to\infty} \sum_{l=0}^n \epsilon^l \alpha^l$ exists. Further, for small ϵ (i.e., reasonably

accurate observation quality), $\alpha^0 + \epsilon\alpha^1$ appears to be a good approximation of γ. Recall by Lemma 1 that the expected total discounted reward over the infinite horizon, v, is such that $v = \lambda\gamma$, where λ is a pmv. Hence, if $\alpha^1 \leq 0$ ($\alpha^1 \geq 0$), then v is non-decreasing (non-increasing) as ϵ gets smaller. Thus, if the finite-memory policy is such that $\alpha^1 \leq 0$, then there may be value in improving state observation quality in order to improve expected system performance. However, if $\alpha^1 \geq 0$, then improving state observation quality will never improve, and may degrade expected system performance. If α^1 is neither non-positive nor non-negative, then whether there is value in improving state observation quality depends on the sign of $\lambda\alpha^1$ and hence the value of λ.

We now present a finite-memory policy that guarantees $\alpha^1 \leq 0$. This policy is a zero-memory policy (i.e., $a(t)$ depends only on $z(t)$), is identical to the optimal policy for the case where $\epsilon = 0$ (i.e., the perfect state observation case and hence a MDP), and is thus relatively easy to calculate and implement. The following result is essentially equivalent to a well-known result described in [17] and elsewhere.

Corollary 1. *Let $\delta^* : Z \to A$ be a policy that achieves the maximum in*

$$\alpha^*(i) = \max_{a\in A}\left\{r(i,a) + \beta\sum_j p_{ij}(a)\alpha^*(j)\right\}.$$

Then, $\alpha^*(i) = \alpha^0(i, \overline{\zeta}(i, \zeta, \delta^*(\zeta))) \geq \alpha^0(i, \overline{\zeta}(z, \zeta, \delta^*(\zeta))) = \alpha^0(i, z)$, for all i and z, and hence $\alpha^1(i, \zeta) \leq 0$ for all i and ζ.

Proof. Let $\{\alpha_n^*\}$ and $\{\delta_n^*\}$ be such that

$$\begin{aligned}\alpha_{n+1}^*(i) &= \max_{a\in A}\left\{r(i,a) + \beta\sum_j p_{ij}(a)\alpha_n^*(j)\right\}\\ &= r(i, \delta_n^*(i)) + \beta\sum_j p_{ij}(\delta_n^*(i))\alpha_n^*(j),\end{aligned}$$

where $\alpha_0^*(i) = 0$. Then, $\lim_{n\to\infty}\|\alpha_n^* - \alpha^*\| = 0$ and hence $\alpha^*(i) = \alpha^0(i, \overline{\zeta}(i, \zeta, \delta^*(\zeta)))$. It then follows from Proposition 1 that for the case where $\zeta = \zeta_t^n$ and $z(t) = z \neq i$, that

$$\begin{aligned}\alpha^0(i,\zeta) &= r(i, \delta^*(z)) + \beta\sum_j p_{ij}(\delta^*(z))\alpha^*(j)\\ &\leq r(i, \delta^*(i)) + \beta\sum_j p_{ij}(\delta^*(i))\alpha^*(j) = \alpha^*(i).\end{aligned}$$

Clearly, $\alpha^0(i, \zeta) = \alpha^*(i, z)$. The fact that $\alpha^1(i, \zeta) \leq 0$ follows from the definition of α^1 in terms of Δ^0 and hence in terms of α^0. □

We remark that we showed in the proof of Corollary 1 that $\alpha^*(i)$ is an upper bound on $\alpha^0(i, \zeta)$ for any ζ, where α^0 is generated using δ^*. More generally, α^* is an upper bound on α^0 for any ζ and any finite-memory policy, which is true due to the fact that for the completely observed Markov decision process (where $z(t) = s(t)$), $s(t)$ is a sufficient statistic for ζ_t^n. We also remark that if we select δ^* to achieve the minimum, rather than the maximum, in the optimality equation in Corollary 1, then $\alpha^1(i, \zeta) \geq 0$ for all i and ζ.

As we are assured by Corollary 1 that there exists a finite-memory (specifically zero-memory) policy whose performance will not degrade as state observation quality improves, we are also assured that there exists a zero-memory policy whose performance will not improve as state observation quality improves. More generally, if a policy tends to accrue greater benefit from accurate observations than from inaccurate observations (i.e., if $\alpha^0(j, \overline{\zeta}(j, \zeta, \delta(\zeta))) > \alpha^0(j, \overline{\zeta}(z, \zeta, \delta(\zeta)))$, where z does not equal j; reference the definitions of $\Delta^0(i, \zeta)$ and $\alpha^1(i, \zeta)$ in Proposition 1), then the policy is likely to be such that improved state observation quality will increase system performance. Indeed, if a policy does not possess this characteristic, it may be decidedly sub-optimal, and hence it may be worth switching to a more productive controller.

Finally, we examine a class of history invariant policies, and show that such policies are independent of state observation quality. Thus, improving or degrading state observation quality will have no effect on the performance of a stationary history invariant policy.

Corollary 2. *Assume δ is a such that $\delta(\zeta) = a$ for all ζ. Then $\alpha^l = 0$ for all $l \geq 1$ and hence the expected total discounted reward is independent of ϵ.*

Proof. If $\delta(\zeta) = a$ for all ζ, then $\alpha^0 = (I - \beta P(a))^{-1} r(a)$, where $P(a) = \{p_{ij}(a)\}$ and $r(a) = \{r(i, a)\}$. Thus, α^0 is independent of ζ, which implies $\Delta^0 = 0$ and hence $\alpha^1 = 0$. An induction argument then implies that $\Delta^{l-1} = 0$ and hence $\alpha^l = 0$ for all $l \geq 1$. □

8. Examples

We now present two inventory control examples that only differ in the observation probability distribution. We restrict attention to order-up-to policies and calculate the signs of all $\alpha^1(i, j)$ for all i and j. The first example demonstrates that for sufficiently small ϵ, some order-up-to policies will improve system performance as state observation quality improves, and that some order-up-to policies will degrade system performance as state observation quality improves. The policies that degrade performance in the first example might be considered unreasonable, since they specify order-up-to quantities much larger than the optimal policy for the completely observed counterpart (when $\epsilon = 0$). The second example shows that a more reasonable sub-optimal order-up-to policy is guaranteed not to improve system performance if given improved state observation quality; in this example, this policy only increases the order-up-to quantity from the completely-observed optimum by one.

Let the per unit per period holding cost $h = 1999$, the per unit ordering cost $c = 1000$, the per unit selling price $p = 3000$ and the discount factor $\beta = 0.9$. Thus $r(i, a) = -hi - ca + p[\sum_{j\leq i+a} jP(j) + \sum_{j>i+a}(i + a)P(j)]$, where $P(j)$ is the demand probability distribution and is given by:

$$P(j) = \begin{cases} 0.5 & \text{if } j = 1\\ 0.45 & \text{if } j = 2\\ 0.05 & \text{if } j = 10\\ 0 & \text{otherwise.}\end{cases}$$

For the first example, the observation probability matrix is

$$q(z|j) = \begin{cases} 1-\epsilon & \text{if } z = j\\ \dfrac{\epsilon}{3|j-z|} & \text{if } 2 \leq j \leq 8, \quad z \neq j,\ j-2 \leq z \leq j+2\\ \dfrac{\epsilon}{3|j-z|} & \text{if } j = 1,\ 2 \leq z \leq 3\\ \dfrac{1}{2}\epsilon & \text{if } j = 1,\ z = 0\\ \dfrac{2\epsilon}{3|j-z|} & \text{if } j = 0,\ 1 \leq z \leq 2\\ \dfrac{\epsilon}{3(j-z)} & \text{if } j = 9,\ 7 \leq z \leq 8\\ \dfrac{1}{2}\epsilon & \text{if } j = 9,\ z = 10\\ \dfrac{2\epsilon}{3(j-z)} & \text{if } j = 10,\ 8 \leq z \leq 9\\ 0 & \text{otherwise.}\end{cases}$$

We now examine the behavior of zero-memory policies of the following form:

$$\delta(z) = \begin{cases} k - z & \text{if } z \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Numerical calculations indicate that $\alpha^1(i, j) \leq 0$ for all i, j for $k \in \{0, 1, 2\}$, implying that these policies improve system performance as state observation quality increases for sufficiently small ϵ. We also observe that $\alpha^1(i, j) \geq 0$ for all i, j for $k \in \{9, 10\}$, implying that these policies decrease system performance as state observation accuracy increases. For $k \in \{3, 4, 5, 6, 7, 8\}$ we observe $\alpha^1(i, j) \leq 0$ for some i, j and we also observe $\alpha^1(i, j) \geq 0$ for other i and j implying that the impact of these policies, given improved state observation accuracy, is inconclusive. We remark that the optimal completely observed MDP order up to level is 2, which is consistent with Corollary 1. We also note that $\alpha^1(i, j) = 0$ for all i, j for $k = 0$ which is consistent with Corollary 2.

We now consider the same parameter values except that the observation probability matrix is given by:

$$q(z|j) = \begin{cases} 1 - \epsilon & \text{if } z = j \\ \dfrac{\epsilon}{10} & \text{otherwise.} \end{cases}$$

For this example we examine policies of the same form described above. Numerical calculations imply that $\alpha^1(i, j) \leq 0$ for all i, j for all $k \in \{0, 1, 2\}$, implying that the system performance of these policies improves with improved observation quality for sufficiently small ϵ. We also observe that $\alpha^1(i, j) \geq 0$ for all i, j for all $k \geq 3$, implying that these policies will degrade system performance if given improved state observation quality. We remark that the optimal completely observed order up to level is again 2. We also note that $\alpha^1(i, j) = 0$ for all i, j for $k = 0$.

We recall from Corollary 1, the "order up to 2" policy is guaranteed to improve system performance if given improved state observation quality. Since state observations may be inaccurate, it would seem reasonable to use an "order up to Y" policy for some $Y > 2$. Interestingly, our numerical results indicate that if we did so, then the resulting policy would be guaranteed not to improve system performance if given improved state observation quality.

References

[1] D.A. Aberdeen, Policy-gradient algorithms for partially observable Markov decision processes, Ph.D. Thesis, The Australian National University, 2003.
[2] K.J. Åström, Optimal control of Markov processes with incomplete state information, Journal of Mathematical Analysis and Applications 10 (1965) 174–205.
[3] D.P. Bertsekas, Dynamic Programming and Optimal Control, Academic Press, New York, NY, 1976.
[4] M. Fisher, A. Raman, A. McClelland, Rocket science retailing is almost here—are you ready? Harvard Business Review 74 (2000) 115–124.
[5] E.A. Hansen, An improved policy iteration algorithm for POMDPs, in: Proceedings of the Tenth Neural Information Processing Systems Conference, Denver, CO, 1997, pp. 1015–1021.
[6] L.P. Kaelbling, M.L. Littman, A.R. Cassandra, Planning and acting in partially observable stochastic domains, Artificial Intelligence 101 (1998) 99.
[7] Y. Kang, S.B. Gershwin, Information inaccuracy in inventory systems: stock loss and stockout, IIE Transactions 37 (2005) 843–859.
[8] A.G. Kok, K. Shang, Inspection and replenishment policies for systems with record inaccuracy, Fuqua School of Business. Duke University Working Paper.
[9] H.L. Lee, O. Ozer, Unlocking the value of RFID, Production and Operations Management 16 (2007) 40–64.
[10] W.S. Lovejoy, A survey of algorithmic methods for partially observed Markov decision processes, Annals of Operations Research 28 (1991) 47–66.
[11] N. Meuleau, L. Peshkin, K. Kim, L.P. Kaelbling, Learning finite-state controllers for partially observable environments, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers, 1999, pp. 427–436.
[12] G.E. Monahan, A survey of partially observable Markov decision processes: theory, models, and algorithms, Management Science 28 (1982) 1–16.
[13] P. Poupart, C. Boutilier, Bounded finite state controllers, in: S. Thrun, L. Saul, B. Schölkopf (Eds.), Advances in Neural Information Processing Systems 16, MIT Press, Cambridge, MA, 2004.
[14] M.L. Puterman, Markov Decision Processes, John Wiley and Sons Inc., New York, 1994.
[15] C. Uckun, F. Karaesmen, S. Savas, Investment in improved inventory accuracy in a decentralized supply chain, Graduate Schools of Sciences and Engineering. Koc University Working Paper.
[16] C.C. White, A survey of solution techniques for the partially observed Markov decision process, Annals of Operations Research 32 (1991) 215–230.
[17] C.C. White, D. Harrington, Application of Jensen's inequality for adaptive suboptimal design, Journal of Optimization Theory and Applications 32 (1980) 89–99.
[18] C.C. White, W. Scherer, Solution procedures for partially observed Markov decision processes, Operations Research 37 (1989) 791–797.
[19] C.C. White, W. Scherer, Finite-memory suboptimal design partially observed Markov decision processes, Operations Research 42 (1994) 439–455.