Optimal Policies for a Class of Restless Multiarmed Bandit Scheduling Problems with Applications to Sensor Management

R. B. WASHBURN
M. K. SCHNEIDER

We present verifiable sufficient conditions for determining optimal policies for finite horizon, discrete time Markov decision problems (MDPs) with terminal reward. In particular, a control policy is optimal for the MDP if (i) it is optimal at the terminal time, (ii) immediate decisions can be deferred to future times, and (iii) the probability transition functions are commutative with respect to different decisions. The result applies to a class of finite horizon restless multiarmed bandit problems that arise in sensor management applications, which we illustrate with a pair of examples.

Manuscript received January 26, 2006; revised February 25, 2008; released for publication May 2, 2008.

Refereeing of this contribution was handled by Chee Chong.

This material is based upon work supported by the United States Air Force under Contract No. F33615-02-C-1197.

Authors' addresses: R. B. Washburn, Parietal Systems Inc., 510 Turnpike Street, Suite 201, North Andover, MA 01845, robert.washburn@parietal-systems.com; M. K. Schneider, BAE Systems, 6 New England Executive Park, Burlington, MA 01803, michael.k.schneider@baesystems.com.

1557-6418/08/$17.00 © 2008 JAIF

1. INTRODUCTION

Consider the Markov decision problems (MDPs) arising in the areas of intelligence, surveillance, and reconnaissance in which one selects among different targets for observation so as to track their position and classify them from noisy data [9], [10]; medicine in which one selects among different regimens to treat a patient [1]; and computer network security in which one selects different computer processes for observation so as to find ones exhibiting malicious behavior [6]. These MDPs all have a special structure. Specifically, they are discrete-time MDPs in which one controls the evolution of a set of Markov processes. There are two possible transition probability functions for the processes. The control at a given time selects a subset of processes, which then transition independently according to the controlled transition probability; the remaining processes transition independently according to the uncontrolled transition probability. Rewards are additive across processes and accumulated over time. The control problem is one of determining a policy to select controls so as to maximize expected rewards. MDPs with this structure have been termed restless bandit problems [15]. Our particular interest in such problems is in developing methods for deriving optimal solutions to them. Such solutions may be important of themselves as a control solution or may be useful for analyzing a problem in the process of developing a good suboptimal controller.

Restless bandits problems are a variation of a classical stochastic scheduling problem called a multiarmed bandit problem. It differs from the restless bandits problems considered here in two key respects. The first is that the states of the unselected process in the multiarmed bandit problem do not change. Second, the rewards in a multiarmed bandit problem are accumulated over an infinite horizon, discounting future rewards. Note that this is a significant difference because the time remaining in the horizon is essentially a component of the state and does not change for a multiarmed bandit problem but does change for the finite horizon restless bandit problems considered here.

A number of techniques have been previously developed for computing solutions to restless bandits problems. For example, index rules have been shown to optimally solve classical multiarmed bandit scheduling problems [2], [4]. Generalizations of this result have been conjectured, and some of them have been proven to apply to other classes of restless bandit problems [14], [15]. Proofs establishing the optimality of controls for finite-horizon restless bandit problems with particular reward structures have also been presented [1], [3], [5]. Each of these results describes a set of conditions for a control to be optimal for a restless bandit problem.

This paper introduces a set of novel conditions that are sufficient for a control policy to be optimal for a finite-horizon MDP. The conditions are readily verified

for a specified control policy and are convenient for verifying the optimality of controls such as priority index rules [3], [4], [15]. We have been able to apply the conditions to verify the optimality of controls for a number of different restless bandit problems [12]. In particular, our conditions can be used to verify special cases of previous results on the optimality of controls for MDPs in [1] and [3]. We have also been able to apply our results to verify the optimality of controls for MDPs arising in sensor management [9], [10] applications for which no existing proofs of optimality existed. General conditions have not been previously developed that can verify optimality of strategies for such a range of examples. These sufficient conditions may prove useful in helping to identify and verify optimal policies for similar multiarmed bandit problems and for developing good suboptimal solutions for more complex problems. The latter is illustrated by the work in [3] and [13]. In both cases, a priority index rule is proven to be optimal for special cases of a more general restless bandit problem. Although the policies are not optimal for the general problems, empirical results reported in the papers demonstrate that the policies perform well even for the more general case.

An example of the type of sensor management problem we are interested in is that of managing an airborne sensor to collect data on ground targets. The goals could be either to collect kinematic data so as to track the kinematic state of the targets or to collect discriminative data so as to classify them. The control problem is one of selecting a subset of the targets for observation, subject to sensor field of view constraints, given current estimates of target and class. The objective is to optimize the quality of the data collected within a finite time horizon. Such sensor management problems are naturally modeled as restless bandit problems [8], [7], [11]. The quality of the data for each target can be modeled as a Markov process, which transitions differently depending on whether the target is selected for observation or not.

The details of our results and applications of them to sensor management problems are provided in the rest of the paper. Section 2 presents our results on sufficient conditions for a control to be optimal for a Markov decision problem. Section 3 applies the result to a general sensor management problem. Finally, examples of managing a sensor to perform binary classification and target tracking are presented in Section 4-A. For the reader's convenience, we have relegated the proofs for the main theorem and all propositions to the Appendix.

2. SUFFICIENT CONDITION

We will denote a MDP with terminal reward by the tuple $(\mathbb{X},\mathbb{U},p_u,R,T)$ where $\mathbb{X}$ denotes the discrete (finite or countable) state space of the Markov chain, $\mathbb{U}$ denotes the finite set of possible decisions, $\{p_u : u \in \mathbb{U}\}$ is the collection of transition probabilities parameterized by the decision u, R is the terminal reward function $R : \mathbb{X} \to \mathbb{R}$, and the integer T is the terminal time.

If $X(t)$, $0 \le t \le T$ is the Markov process with decisions $U(t)$, $0 \le t \le T-1$, and terminal reward $R(X(T))$, the MDP problem is to select U to maximize the expected value $E\{R(X(T))\}$ of the terminal reward. We assume that the decision $U(t)$ depends only on $X(0),\ldots,X(t)$ and that

$$\Pr\{X(t+1) = \xi \mid X(t) = x, U(t) = u\} = p_u(\xi \mid x). \tag{1}$$

The dynamic programming equations for the optimal reward-to-go function $V(x,t)$ for the MDP $(\mathbb{X},\mathbb{U},p_u,R,T)$ are given as follows. The terminal condition is

$$V(x,T) = R(x). \tag{2}$$

The recursion is

$$V(x,t) = \max_u \{V_u(x,t)\} \tag{3}$$

for times $0 \le t \le T-1$, where we define

$$V_u(x,t) := \sum_\xi V(\xi,t+1)p_u(\xi \mid x). \tag{4}$$

Also, any u that achieves the maximum in (3) is defined to be an optimal decision at time t when in state x.

DEFINITION 1 Suppose that the MDP $(\mathbb{X},\mathbb{U},p_u,R,T)$ has the probability transition functions $p_u(\xi \mid x)$ for $x,\xi \in \mathbb{X}$, $u \in \mathbb{U}$, and terminal reward $R(x)$ for $x \in \mathbb{X}$. If $\Phi(x,t) \subset \mathbb{U}$ for each $x \in \mathbb{X}$ and $0 \le t \le T-1$, we say that Φ is a **strategy set** for the MDP.

DEFINITION 2 If Φ is a strategy set for the MDP $(\mathbb{X},\mathbb{U},p_u,R,T)$ and if for each $x \in \mathbb{X}$, the expected value for selecting each $u \in \Phi(x,T-1)$ achieves the maximum value, *i.e.*,

$$\sum_y R(y)p_u(y \mid x) = \max_v \sum_y R(y)p_v(y \mid x), \tag{5}$$

we say that the strategy set Φ is **terminally optimal** for the MDP.

DEFINITION 3 More generally, for, $0 \le t \le T-1$ and $x \in \mathbb{X}$, define $\Phi^*(x,t)$ to be the **set of optimal strategies**

$$\Phi^*(x,t) = \left\{u : V_u(x,t) = \max_w V_w(x,t)\right\}. \tag{6}$$

What follows is a pair of definitions for properties of the strategy set and MDP as well as a theorem concerning the optimality of the strategy set when these conditions hold. Note that the properties in the definitions are abstract at this point and are illustrated later in this section with an example.

DEFINITION 4 If Φ is a strategy set for the MDP $(\mathbb{X},\mathbb{U},p_u,R,T)$, and if for each t such that $0 \le t \le T-2$, each $x \in \mathbb{X}$,

$$\begin{aligned} &u \in \Phi(x,t), \qquad V_v(x,t) > V_u(x,t), \qquad \text{and} \\ &p_v(y \mid x) > 0 \quad \text{imply} \quad u \in \Phi(y,t+1), \end{aligned} \tag{7}$$

then we say that **decisions are deferrable** in the strategy set Φ.

REMARK 1 Definition 4 gives conditions under which if u is in the decision set at the current time but a different decision v is made, then u is still in the decision set at the next time. This condition allows using an interchange argument to prove the optimality of the decision set (Theorem 1). Unfortunately, Definition 4 is too hard to check in practice. However, it is implied by various stronger conditions that are easier to check. For example, if for each t such that $0 \le t \le T-2$, each $x \in \mathbb{X}$, and for all u, v, y,

$$\begin{aligned} &u \in \Phi(x,t),\ v \neq u, \qquad \text{and} \\ &p_v(y \mid x) > 0 \quad \text{imply} \quad u \in \Phi(y, t+1), \end{aligned} \tag{8}$$

then decisions are deferrable in the strategy set Φ. This condition is stronger than the definition, since $V_v(x,t) > V_u(x,t)$ obviously implies that $v \neq u$. At the end of this section we prove another stronger condition for problems with symmetry.

DEFINITION 5 We say that the probability transition functions $p_u(\xi \mid x)$ are **commutative** if for all $u, v \in \mathbb{U}$,

$$\sum_{\eta} p_u(\xi \mid \eta) p_v(\eta \mid x) = \sum_{\eta} p_v(\xi \mid \eta) p_u(\eta \mid x) \tag{9}$$

for all $x, \xi \in \mathbb{X}$.

THEOREM 1 *Suppose that Φ is a strategy set for an MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ with commutative transition probability functions p_u, such that Φ is terminally optimal and decisions in Φ are deferrable. Then the strategy set Φ is optimal in the sense that any decision $U(t) \in \Phi(X(t), t)$ for $0 \le t \le T-1$, is an optimal decision for $(\mathbb{X}, \mathbb{U}, p_u, R, T)$.*

REMARK 2 If $\Phi^*(x,t)$ is the optimal strategy set for $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ as defined in (6), then Φ^* is necessarily terminally optimal. It also necessarily satisfies the condition for deferrable decisions, simply because the hypothesis of the condition,

$$u \in \Phi^*(x,t),\ V_v(x,t) > V_u(x,t), \tag{10}$$

is always false. As we indicated in Remark 1, this condition is difficult to check in practice, but we can replace it with stronger conditions which do not refer to the optimal reward function. With these stronger conditions, it is important to have the third condition, commutativity of the transition probabilities, to prove the optimality of a proposed strategy set.

To conclude this section we will prove another stronger condition for deferrable decisions in Φ based on symmetric MDP problems. Note that for these problems, the state space of the Markov chain $\mathbb{X}$ is a product space $\mathbf{X}^n$, and the ith component of an element $x \in \mathbb{X}$ is denoted by x_i.

DEFINITION 6 The MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ is symmetric if for some $n > 1$

$$\mathbb{X} = \mathbf{X}^n, \tag{11}$$

$$\mathbb{U} = \{1, \ldots, n\}, \tag{12}$$

$$p_{\pi(i)}(\pi y \mid \pi x) = p_i(y \mid x) \tag{13}$$

and

$$R(x) = R(\pi x) \tag{14}$$

where π permutes the components of x, y, namely

$$\pi x = (x_{\pi(1)}, \ldots, x_{\pi(n)}), \tag{15}$$

for any permutation π of $\{1, \ldots, n\}$ and all $x \in \mathbf{X}^n$.

REMARK 3 Note that the symmetry conditions in Definition 6 all hold for multiarmed bandit scheduling problems. However, the symmetric scheduling problem considered here still differs from the multiarmed bandit problem in two key respects. First, the states for unobserved processes may change, whereas, for multiarmed bandit problems, the states of unobserved processes remain the same. Second, the horizon here is finite whereas the horizon for multiarmed bandit problems is infinite.

PROPOSITION 1 *Suppose that the MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ is symmetric. Then if for $0 \le t \le T-2$ and all $x \in \mathbb{X}$*

$$\begin{aligned} &u \in \Phi(x,t),\ x_v \neq x_u \qquad \textit{and} \\ &p_v(y \mid x) > 0 \quad \textit{imply} \quad u \in \Phi(y, t+1), \end{aligned} \tag{16}$$

decisions are deferrable in Φ.

An example of a strategy for a symmetric MDP that is terminally optimal, deferrable, and commutative is as follows. Suppose $\mathbf{X} = \mathbb{N}$ and

$$R(x) = \sum_{i=1}^{n} \delta(x_i) \tag{17}$$

where

$$\delta(x_i) = \begin{cases} 1 & \text{if} \quad x_i = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{18}$$

Moreover, define the transition probabilities as follows. If $x_i > 0$

$$p_i(y \mid x) = \begin{cases} 1/2 & \text{if} \quad y_i = x_i \pm 1 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

and if $x_i = 0$,

$$p_i(y \mid x) = \begin{cases} 1 & \text{if} \quad y_i = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This is a MDP in which there are n independent Markov processes x_i evolving on the non-negative integers. A process x_i transitions only if it is selected by the control and is equally likely to increase or decrease. The value 0 is a trapping state. The objective is to drive as many

processes as possible to the trapping state. Define the strategy set to be the non-zero processes of minimal value

$$\Phi(x,t) = \arg\min_i \{x_i > 0\}. \tag{21}$$

This strategy set is terminally optimal because selecting a process with value 1 is optimal at the last stage. The strategy set is deferrable because the condition of Proposition 1 holds. Specifically, if $u \in \Phi(x,t)$ and $v \in \mathbb{U}$ is such that $x_v \neq x_u$, then $x_v > x_u$ and $p_v(y \mid x) > 0$ implies $y_v \geq x_u$ so that

$$u \in \arg\min_i \{y_i > 0\} = \Phi(y, t+1). \tag{22}$$

Finally, the probability transitions are commutative because

$$\sum_\eta p_u(\xi \mid \eta) p_v(\eta \mid x)$$

$$= \begin{cases} 1/4 & \text{if} \quad \xi_u = x_u \pm 1, \xi_v = x_v \pm 1 \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

$$= \sum_\eta p_v(\xi \mid \eta) p_u(\eta \mid x). \tag{24}$$

Thus, the strategy is optimal for this problem by Proposition 1.

Applications of the results in this section to sensor management problems follow.

3. APPLICATIONS TO SENSOR MANAGEMENT PROBLEMS

The results are stated for a very general situation in Section 2, where few assumptions are made concerning the statistics of the Markov chain. However, the optimality conditions are expected to be useful for analyzing special cases of more general problems, in part to develop good heuristics for the general case. The purpose of this section is to specialize the optimality conditions to problems where the Markov chain is a product of independent, identically distributed chains, which is a common situation arising in some important special cases of sensor management problems.

Specifically, consider the sensor management problem where there are n targets and we can only observe one target at a time. In the simplest case, the decision $U(t)$ to make at each time t is only which target $i = 1, \ldots, n$ to observe. There is a Markov chain $X_i(t)$ corresponding to each target i, where $X_i(t)$ represents the information state of target i at time t. Typically, we assume that the chains $X_i(t)$ are independent and identically distributed, and that the selected (i.e., observed) chain transitions according to $p(\xi \mid x)$ and the $n-1$ unobserved chains transition according to $q(\xi \mid x)$. Moreover, the reward is typically additive over the n targets, namely

$$R(X_1(T), \ldots, X_n(T)) = \sum_{i=1}^n r(X_i(T)). \tag{25}$$

The resulting MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ has special structure where

$$\mathbb{X} = \mathbf{X}^n \text{ and } \mathbf{X} \text{ is the state space of one Markov chain } X_i \tag{26}$$

$$\mathbb{U} = \{1, \ldots, n\} \tag{27}$$

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} q(\xi_j \mid x_j) \qquad \text{for} \quad i \in \mathbb{U},\ x, \xi \in \mathbf{X}^n \tag{28}$$

$$R(x) = \sum_{i=1}^n r(x_i) \qquad \text{for} \quad x \in \mathbf{X}^n. \tag{29}$$

REMARK 4 If $s = |\mathbf{X}|$ is the number of states for each single Markov chain, then the computational complexity of the dynamic programming solution is $O(ns^{2n}T)$. Thus, for fixed s and T, the complexity is exponential in n. Furthermore, the memory requirements are exponential, namely $O(s^nT)$. In some cases we can find an optimal strategy of the form $U(t) \in \Phi((X_1(t), \ldots, X_n(t)), t)$ where

$$\Phi(x,t) = \{i : M_i(x_i, t) = \max_j M_j(x_j, t)\}. \tag{30}$$

This is what we call a priority index rule strategy. The $M_i(x_i, t)$ are indices that can be computed for each target with complexity $O(s^2T)$ (i.e., equivalent to solving the dynamic program for one target). Thus, the complexity of the n target strategy is $O(ns^2T)$ rather than $O(ns^{2n}T)$, linear in n rather than exponential in n.

For the class of transition probabilities $p_i(\xi \mid x)$ with structure (28), commutativity is equivalent to the commutativity of the transition functions p and q, as the following simple result shows.

PROPOSITION 2 *If the transition probability functions $p_i(\xi \mid x)$ defined for $\xi, x \in \mathbb{X}^n$ and $i \in \{1, \ldots, n\}$ satisfy*

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} q(\xi_j \mid x_j), \tag{31}$$

and if for all $\xi_1, x_1 \in X$,

$$\sum_{\eta_1} q(\xi_1 \mid \eta_1) p(\eta_1 \mid x_1) = \sum_{\eta_1} p(\xi_1 \mid \eta_1) q(\eta_1 \mid x_1), \tag{32}$$

then $p_i(\xi \mid x)$ are commutative transition probability functions for $\xi, x \in \mathbb{X}^n$.

REMARK 5 Note that commutativity always holds if p or q is the identity transition $\delta(\xi_i \mid x_i) = 1$ for $\xi_i = x_i$ and 0 otherwise. Note that $q = \delta$ is assumed true in (nonrestless) multiarmed bandit problems. Also, classification sensor management problems often satisfy $q = \delta$ (i.e., the classification information state remains unchanged while the target is unobserved).

REMARK 6 Transition probabilities of the form

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} \delta(\xi_j \mid x_j) \tag{33}$$

and reward functions

$$R(x) = \sum_{i=1}^{n} r(x_i) \tag{34}$$

are obviously symmetric.

For this class of MDPs corresponding to sensor management problems, the general result (Theorem 1) becomes the following.

COROLLARY 1 *Suppose that the MDP* $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ *has special symmetric structure where*

$$\mathbb{X} = \mathbf{X}^n \text{ and } \mathbf{X} \text{ is the state space of one one Markov chain } X_i \tag{35}$$

$$\mathbb{U} = \{1, \ldots, n\} \tag{36}$$

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} q(\xi_j \mid x_j) \qquad \text{for} \quad i \in \mathbb{U},\ x, \xi \in \mathbf{X}^n \tag{37}$$

$$R(x) = \sum_{i=1}^{n} r(x_i) \text{ for } x \in \mathbf{X}^n. \tag{38}$$

Then the strategy set Φ *is optimal if the following three conditions are met. The first condition is that p and q are commutative so that*

$$\sum_{\eta} p(\xi \mid \eta) q(\eta \mid x) = \sum_{\eta} q(\xi \mid \eta) p(\eta \mid x). \tag{39}$$

Suppose that $\Phi(x,t)$ *is a strategy set for* $x = (x_1, \ldots, x_n)$. *Then, the second condition is that* $i \in \Phi(x, T-1)$ *implies*

$$\sum_{y_i} r(y_i) p(y_i \mid x_i) - r(x_i) \geq \sum_{y_j} r(y_j) p(y_j \mid x_j) - r(x_j) \tag{40}$$

for all $j \neq i$, *and the third condition is that*

$$i \in \Phi(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n, t),\ x_i \neq x_j, \qquad \text{and} \qquad p(y_j \mid x_j) > 0 \tag{41}$$

implies that

$$i \in \Phi(x_1, \ldots, x_i, \ldots, y_j, \ldots, x_n, t+1). \tag{42}$$

PROOF The condition on $p_i(\xi \mid x)$ implies that it is commutative. The second condition implies that Φ is terminally optimal for the terminal reward $R(x)$, and the third condition implies that decisions in Φ are deferrable (Proposition 1). The result follows from Theorem 1.

4. SENSOR MANAGEMENT EXAMPLES

What follows are two examples that illustrate the application of the conditions presented in the paper. The first is a binary classification example, which is a type of finite horizon sensor management problem for which the states of unobserved processes (other than the time remaining until the end of the horizon) do not change. The second is a tracking problem, for which the states of unobserved processes do change. The examples illustrate how the novel conditions presented in this paper apply to a large class of problems that includes these two. The utility of such an analysis is that it sheds insight into sensor management problems and suggests heuristics that could be used for more general sensor management problems.

A. Binary Classification Problem

This problem is to classify as many of n objects as possible over a finite time horizon T given binary measurements of the objects. This problem is similar to the classical treasure hunting problem [2]. In that problem, one selects among a finite number of areas to search for treasure, but the treasure may be missed with a fixed probability. This is a special type of a multiarmed bandit problem in which a so-called deteriorating condition holds so that the optimal policy is a greedy policy. The binary classification problem considered here differs from the treasure hunting problem in two respects. The first is that the horizon here is finite whereas it is infinite for the treasure hunting problem. The second is that processes in the two problems represent different quantities and so have different transition probabilities. In the problem here, each process represents the probability that a target is of a particular type. In the treasure hunting problem, each process represents the probability that the treasure is present at a particular location. The details of the binary classification problem follow.

First, note that this problem is a partially observed Markov decision process (POMDP) that can be interpreted as an MDP with a countable state space. Suppose there are n random variables Z_i with values 0, 1 and that $\Pr\{Z_i = 1\} = p$ for all $i = 1, \ldots, n$. Suppose that the $Y_i(t)$ are $0, 1$ observations of Z_i, and $Y_i(t)$ are independent and identically distributed conditioned on Z_i with

$$\Pr\{Y_i(t) = y \mid Z_i = z\} = (1-\varepsilon) \cdot \delta_{y,z} + \varepsilon \cdot (1 - \delta_{y,z}), \tag{43}$$

where we use the notation $\delta_{y,z} = 1$ if $y = z$ and 0 otherwise. We assume that $\varepsilon < \frac{1}{2}$. Note that ε is the probability of classification error for one measurement.

Define the information state $X_i(t)$ as the conditional probability

$$X_i(t) = \Pr\{Z_i = 1 \mid Y_i(1), \ldots, Y_i(t)\}. \tag{44}$$

The objective of the problem is to maximize the expected reward

$$E\left\{\sum_{i=1}^{n} r(X_i(T))\right\} \tag{45}$$

at the terminal time T, where $r(x_i)$ is the individual reward

$$r(x_i) = \max_{d_i = 0,1} \{r(d_i, 1) x_i + r(d_i, 0)(1 - x_i)\} \tag{46}$$

and $r(d,z)$ are the rewards for the different types of outcomes (i.e., deciding d_i when the true state of i is z_i).

The processes $X_i(t)$ satisfy

$$X_i(0) = p \tag{47}$$

and for $t \geq 0$,

$$X_i(t+1) = \begin{cases} \dfrac{(1-\varepsilon)X_i(t)}{(1-2\varepsilon)X_i(t)+\varepsilon} \\ \qquad \text{with probability } (1-2\varepsilon)X_i(t)+\varepsilon \\ \dfrac{\varepsilon X_i(t)}{(2\varepsilon-1)X_i(t)+1-\varepsilon} \\ \qquad \text{with probability } (2\varepsilon-1)X_i(t)+1-\varepsilon. \end{cases} \tag{48}$$

Note that although $X_i(t)$ take values in $\mathbb{R}$, there are only a countable number of possible values they can take. Thus, $X_i(t) \in \mathbf{X} \subset \mathbb{R}$ where $\mathbf{X}$ is a countable set. Thus, we have an MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ where

$$\mathbb{X} = \mathbf{X}^n \tag{49}$$

$$\mathbb{U} = \{1,2,\ldots,n\} \tag{50}$$

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} \delta(\xi_j \mid x_j) \qquad \text{for} \quad i \in \mathbb{U},\ x, \xi \in \mathbf{X}^n \tag{51}$$

$$R(x) = \sum_{i=1}^{n} r(x_i) \qquad \text{for} \quad x \in \mathbf{X}^n \tag{52}$$

where $p(\xi_i \mid x_i)$ is defined by

$$p\left(\frac{(1-\varepsilon)x_i}{(1-2\varepsilon)x_i+\varepsilon} \mid x_i\right) = (1-2\varepsilon)x_i + \varepsilon \tag{53}$$

$$p\left(\frac{\varepsilon x_i}{(2\varepsilon-1)x_i+1-\varepsilon} \mid x_i\right) = (2\varepsilon-1)x_i + 1 - \varepsilon \tag{54}$$

and $r(x_i)$ is defined by (46). We will consider the special case for which $r(1,1) = r(0,0) = 1$ and $r(0,1) = r(1,0) = 0$ so that

$$r(x_i) = \tfrac{1}{2} + |x_i - \tfrac{1}{2}|, \tag{55}$$

and we will assume that the prior probability $p = \frac{1}{2}$. Note that if $p = \frac{1}{2}$, then

$$\mathbf{X} = \left\{ \frac{1}{1+\left(\frac{\varepsilon}{1-\varepsilon}\right)^m} : \quad m = 0, \pm 1, \pm 2, \ldots \right\}. \tag{56}$$

PROPOSITION 3 *The strategy set Φ defined by*

$$\Phi(x) = \{i : |x_i - \tfrac{1}{2}| = \min_j |x_j - \tfrac{1}{2}|\} \tag{57}$$

is optimal for the binary classification problem with $r(1,1) = r(0,0) = 1$, $r(0,1) = r(1,0) = 0$, and prior probability $p = \frac{1}{2}$ for each object i.

B. Tracking Problem

The following is an example in which one is managing a sensor to track targets. Specifically, one is tracking targets over a finite horizon with a noisy sensor. At the end of the time horizon, the tracks are to be handed over to another sensor. The handover is successful if the track mean square error is smaller than the required level. The objective is to maximize the number of tracks that are successfully handed over.

Note that this example differs from the binary classification one in that unobserved chains have nontrivial dynamics. Specifically, the dynamics are those of the track error covariances. The conditions are used to verify the optimality of a strategy for an approximate model of the track error covariance where the increase in error when a track is unobserved is given by that of a Kalman filter, but the error reduction is approximated as being constant, independent of the initial error. The details of this example are as follows.

Consider the one-dimensional tracking problem in which there are n targets each of which is moving as a one-dimensional Brownian motion with process noise variance Λ_p. Location measurements have additive noise with variance Λ_m. The state of each track i at time t, for the purposes of sensor management, is the error variance $X_i(t)$. All tracks are initialized with the same error variance Λ_0, and all have the same desired error variance Λ_h at the end of the horizon T. The objective of the problem is to maximize the expected reward

$$E\left\{\sum_{i=1}^{n} r(X_i(T))\right\} \tag{58}$$

at the terminal time T, where $r(x_i)$ is the individual reward

$$r(x_i) = \begin{cases} 1 & \text{if} \quad x_i \leq \Lambda_h \\ 0 & \text{otherwise.} \end{cases} \tag{59}$$

Now suppose that the track error is approximated so that the track error reduction for observed tracks is constant, given by the error reduction from the desired value Λ_h. That is, if the error variance is initially Λ_h, then after one measurement update and one prediction, it is reduced by the amount $\Lambda_p - \Lambda_h^2/(\Lambda_h + \Lambda_m)$. Then, the dynamics of the processes $X_i(t)$ satisfy

$$X_i(0) = \Lambda_0 \tag{60}$$

and for unobserved processes for $t \geq 0$,

$$X_i(t+1) = X_i(t) + \Lambda_p. \tag{61}$$

For the process observed at $t \geq 0$,

$$X_i(t+1) = X_i(t) + \Lambda_p - \frac{\Lambda_h^2}{\Lambda_h + \Lambda_m}. \tag{62}$$

We will assume that $\Lambda_p < \Lambda_h^2/(\Lambda_h + \Lambda_m)$ so that the error for observed processes is always decreasing. This is an MDP $(\mathbb{X}, \mathbb{U}, p_u, R, T)$ with

$$\mathbb{X} = \mathbb{R}^n \tag{63}$$

$$\mathbb{U} = \{1, 2, \ldots, n\} \tag{64}$$

$$p_i(\xi \mid x) = p(\xi_i \mid x_i) \prod_{j \neq i} q(\xi_j \mid x_j) \qquad \text{for} \quad i \in \mathbb{U},\ x, \xi \in \mathbb{X} \tag{65}$$

$$R(x) = \sum_{i=1}^{n} r(x_i) \qquad \text{for} \quad x \in \mathbb{X} \tag{66}$$

where $p(\xi_i \mid x_i)$ is defined by

$$p\left(x_i + \Lambda_p - \frac{\Lambda_h^2}{\Lambda_h + \Lambda_m} \mid x_i\right) = 1 \tag{67}$$

$q(\xi_i \mid x_i)$ is defined by

$$q(x_i + \Lambda_p \mid x_i) = 1, \tag{68}$$

and $r(x_i)$ is defined by (46).

PROPOSITION 4 *The strategy set* Φ *defined by*

$$\Phi(x) = \left\{ i : x_i = \min_j \{x_j : x_j > \Lambda_h - \Lambda_p\} \right\} \tag{69}$$

is optimal for this tracking problem.

REMARK 7 Note that under this strategy, the approximate error variance $X_i(t) \geq 0$ because $X_i(t)$ will decrease only if the process is chosen for observation, which will only occur if $X_i(t) > \Lambda_h - \Lambda_p$ and

$$X_i(t) + \Lambda_p - \frac{\Lambda_h^2}{\Lambda_h + \Lambda_m} \geq \Lambda_h - \frac{\Lambda_h^2}{\Lambda_h + \Lambda_m} > 0. \tag{70}$$

5. CONCLUSION

Thus, the sufficient conditions stated in Section 2 are useful for establishing optimality of sensor management strategies. Note that the optimal strategy for the binary classification and tracking examples presented in Section 4-A are priority index rule strategies, as defined in Section 3. Priority index rules are optimal strategies for other sensor management problems including those in [1], [5], [3]. However, the conditions in this paper do not imply optimality of these strategies except for some special cases of the sensor management problem being solved. Whether there exists a generalization of the results in this paper that implies optimality of priority index rules for general sensor management problems and other restless bandit problems is an open question.

APPENDIX. PROOFS OF RESULTS

A. Proof of Theorem 1

For, $0 \leq t \leq T-1$ and $x \in \mathbb{X}$, let $\Phi^*(x,t)$ be the set of optimal strategies, as defined in Definition 3. We want to prove that

$$\Phi(x,t) \subset \Phi^*(x,t). \tag{71}$$

The terminal optimality condition is equivalent to

$$\Phi(x, T-1) \subset \Phi^*(x, T-1). \tag{72}$$

Thus, assume that $\Phi(x,t+1) \subset \Phi^*(x,t+1)$ is true for $t < T-1$ and prove (71) from it. Suppose that $u \in \Phi(x,t)$ and $u \notin \Phi^*(x,t)$. Clearly $\Phi^*(x,t) \neq \emptyset$ and there is $v \in \Phi^*(x,t)$ such that $V_v(x,t) > V_u(x,t)$. The condition that decisions in Φ are deferrable implies that $u \in \Phi(X(t+1), t+1)$ where $X(t+1)$ results from using $U(t) = v$. The induction hypothesis implies that

$$\Phi(X(t+1), t+1) \subset \Phi^*(X(t+1), t+1), \tag{73}$$

so that $U(t+1) = u$ is an optimal decision.

We now can use the commutativity of the transitions p_w to show that the sequence of decisions $U(t) = u$, $U(t+1) = v$ has the same expected value-to-go as the sequence of decisions $U(t) = v$, $U(t+1) = u$ and must be optimal too. Specifically, note that starting from $X(t)$, if $X(t+2)$ is the state resulting from $U(t) = v$, $U(t+1) = u$ and $\tilde{X}(t+2)$ is the state resulting from $U(t) = u$, $U(t+1) = v$, then commutativity implies that $X(t+2)$ and $\tilde{X}(t+2)$ have the same distribution. By assumption (induction) the decisions $U(t) = v$, $U(t+1) = u$ are optimal and have the value-to-go

$$V(X(t), t) = E\{V(X(t+2), t+2) \mid X(t)\}. \tag{74}$$

Commutativity implies that

$$E\{V(X(t+2), t+2) \mid X(t)\} = E\{V(\tilde{X}(t+2), t+2) \mid X(t)\}, \tag{75}$$

which implies that $U(t) = u$, $U(t+1) = v$ must also be optimal decisions. Thus, u is optimal, contrary to assumption and we must have $u \in \Phi^*(x,t)$.

B. Proof of Proposition 1

First, we show by induction that the symmetry assumption implies that the optimal reward-to-go satisfies

$$V_v(x,t) = V_{\pi(v)}(\pi x, t) \tag{76}$$

for all permutations π. Let x denote a vector in $\mathbf{X}^n = \mathbb{X}$. By definition of symmetry

$$V_{\pi(u)}(\pi x, T-1) = \sum_y R(\pi y) p_{\pi(u)}(\pi y \mid \pi x) \tag{77}$$

$$= \sum_y R(y) p_u(y \mid x) \tag{78}$$

$$= V_u(x, T-1). \tag{79}$$

Now, assume that

$$V_v(x,t+1) = V_{\pi(v)}(\pi x, t+1) \tag{80}$$

for all x, π and prove it for t. Note that the induction hypothesis implies that

$$V(x,t+1) = \max_v V_v(x,t+1) \tag{81}$$

$$= \max_v V_{\pi(v)}(\pi x, t+1) \tag{82}$$

$$= V(\pi x, t+1). \tag{83}$$

By symmetry assumptions,

$$V_u(x,t) = \sum_y V(y,t+1) p_u(y \mid x) \tag{84}$$

$$= \sum_y V(\pi y, t+1) p_{\pi(u)}(\pi y \mid \pi x) \tag{85}$$

$$= \sum_y V(y,t+1) p_{\pi(u)}(y \mid \pi x) \tag{86}$$

$$= V_{\pi(u)}(\pi x, t) \tag{87}$$

which completes the induction.

Now, to prove the statement of Proposition 1, suppose that $u \in \Phi(x,t)$, $V_v(x,t) > V_u(x,t)$ and $p_v(y \mid x) > 0$. We have just shown that

$$V_v(x,t) = V_{\pi(v)}(\pi x, t) \tag{88}$$

for all permutations π. Let π be the permutation that interchanges v and u. Then if $x_v = x_u$, $V_v(x,t) = V_u(x,t)$. Thus, $V_v(x,t) > V_u(x,t)$ implies that $x_v \neq x_u$. By the proposition's assumption, it follows that $u \in \Phi(y, t+1)$, which proves the result.

C. Proof of Proposition 2

Note that for $i \neq j$,

$$\sum_\eta p_i(\xi \mid \eta) p_j(\eta \mid x) \tag{89}$$

$$= \sum_\eta p(\xi_i \mid \eta_i) \prod_{k \neq i} q(\xi_k \mid \eta_k) p(\eta_j \mid x_j) \prod_{k \neq j} q(\eta_k \mid x_k) \tag{90}$$

$$= \sum_{\eta_j} q(\xi_j \mid \eta_j) p(\eta_j \mid x_j) \sum_{\eta_i} p(\xi_i \mid \eta_i) q(\eta_i \mid x_i) \times \prod_{k \neq i,j} \sum_{\eta_k} q(\xi_k \mid \eta_k) q(\eta_k \mid x_k). \tag{91}$$

By assumption

$$\sum_{\eta_j} q(\xi_j \mid \eta_j) p(\eta_j \mid x_j) = \sum_{\eta_j} p(\xi_j \mid \eta_j) q(\eta_j \mid x_j) \tag{92}$$

and

$$\sum_{\eta_i} q(\xi_i \mid \eta_i) p(\eta_i \mid x_i) = \sum_{\eta_i} p(\xi_i \mid \eta_i) q(\eta_i \mid x_i). \tag{93}$$

Thus,

$$\sum_{\eta_j} q(\xi_j \mid \eta_j) p(\eta_j \mid x_j) \sum_{\eta_i} p(\xi_i \mid \eta_i) q(\eta_i \mid x_i) \times \prod_{k \neq i,j} \sum_{\eta_k} q(\xi_k \mid \eta_k) q(\eta_k \mid x_k) \tag{94}$$

$$= \sum_{\eta_j} p(\xi_j \mid \eta_j) q(\eta_j \mid x_j) \sum_{\eta_i} q(\xi_i \mid \eta_i) p(\eta_i \mid x_i) \times \prod_{k \neq i,j} \sum_{\eta_k} q(\xi_k \mid \eta_k) q(\eta_k \mid x_k) \tag{95}$$

$$= \sum_\eta p_j(\xi \mid \eta) p_i(\eta \mid x), \tag{96}$$

proving that

$$\sum_\eta p_i(\xi \mid \eta) p_j(\eta \mid x) = \sum_\eta p_j(\xi \mid \eta) p_i(\eta \mid x). \tag{97}$$

D. Proof of Proposition 3

The proof verifies that the three conditions of Corollary 1 hold.

First, the transition probabilities $p_i(\xi \mid x)$ are obviously commutable and symmetric, and the reward function $R(x)$ is obviously symmetric.

Now, note that

$$\sum_{y_i} [R(y_i) - R(x_i)] p(y_i \mid x_i) \tag{98}$$

$$= -\tfrac{1}{2} - |x_i - \tfrac{1}{2}| \tag{99}$$

$$+ \left(\frac{1}{2} + \left| \frac{(1-\varepsilon)x_i}{(1-2\varepsilon)x_i + \varepsilon} - \frac{1}{2} \right| \right) ((1-2\varepsilon)x_i + \varepsilon) \tag{100}$$

$$+ \left(\frac{1}{2} + \left| \frac{\varepsilon x_i}{(2\varepsilon - 1)x_i + 1 - \varepsilon} - \frac{1}{2} \right| \right) \times ((2\varepsilon - 1)x_i + 1 - \varepsilon). \tag{101}$$

This simplifies to

$$\sum_{y_i} [R(y_i) - R(x_i)] p(y_i \mid x_i) = -|x_i - \tfrac{1}{2}| + \tfrac{1}{2}|x_i - \varepsilon| + \tfrac{1}{2}|(1 - x_i) - \varepsilon|, \tag{102}$$

which is equivalent to

$$\sum_y [R(y) - R(x_i)] p(y \mid x_i) = \begin{cases} 0 & \text{for} \quad 0 \le x_i \le \varepsilon \\ x_i - \varepsilon & \text{for} \quad \varepsilon \le x_i \le \frac{1}{2} \\ 1 - x_i - \varepsilon & \text{for} \quad \frac{1}{2} \le x_i \le 1 - \varepsilon \\ 0 & \text{for} \quad 1 - \varepsilon \le x_i \le 1 \end{cases}. \tag{103}$$

Note that because

$$x_i = \frac{1}{1 + \left(\frac{\varepsilon}{1-\varepsilon}\right)^m}, \tag{104}$$

if $m < 0$, then

$$x_i \le \frac{1}{1 + \left(\frac{\varepsilon}{1-\varepsilon}\right)^{-1}} = \varepsilon \tag{105}$$

and if $m > 0$, then

$$x_i \ge \frac{1}{1 + \left(\frac{\varepsilon}{1-\varepsilon}\right)} = 1 - \varepsilon. \tag{106}$$

Thus,

$$\sum_{y_i} [R(y_i) - R(x_i)] p(y_i \mid x_i) = \begin{cases} 0 & \text{for} \quad x_i \neq \frac{1}{2} \\ \frac{1}{2} - \varepsilon & \text{for} \quad x_i = \frac{1}{2} \end{cases}. \tag{107}$$

In particular,

$$\max_j \left\{ \sum_{y_j} [R(y_j) - R(x_j)] p(y_j \mid x_j) \right\} = \begin{cases} 0 & \text{if all} \quad x_i \neq \frac{1}{2} \\ \frac{1}{2} - \varepsilon & \text{if some} \quad x_i = \frac{1}{2} \end{cases} \tag{108}$$

and if

$$|x_i - \tfrac{1}{2}| = \min_j |x_j - \tfrac{1}{2}|, \tag{109}$$

then

$$\sum_{y_i} [R(y_i) - R(x_i)] p(y_i \mid x_i) = \max_j \left\{ \sum_{y_j} [R(y_j) - R(x_j)] p(y_j \mid x_j) \right\}. \tag{110}$$

This shows that the second condition of Corollary 1 holds.

To show that the third condition of Corollary 1 holds, suppose that $i \in \Phi(x)$ so that

$$|x_i - \tfrac{1}{2}| = \min_k |x_k - \tfrac{1}{2}| \tag{111}$$

and suppose $x_j \neq x_i$, $p(y_j \mid x_j) > 0$. Thus,

$$y_j = \frac{(1-\varepsilon)x_j}{(1-2\varepsilon)x_j + \varepsilon} \tag{112}$$

or

$$y_j = \frac{\varepsilon x_j}{(2\varepsilon - 1)x_j + 1 - \varepsilon}. \tag{113}$$

If $|x_j - \frac{1}{2}| > |x_i - \frac{1}{2}|$, then it is easy to see that $|y_j - \frac{1}{2}| \ge |x_i - \frac{1}{2}|$ and therefore $i \in \Phi(x_1, \ldots, y_j, \ldots, x_n)$. However, it is possible that $|x_j - \frac{1}{2}| = |x_i - \frac{1}{2}|$ and $x_i \neq x_j$. If $x_i \neq \frac{1}{2}$, then the conclusion is not true, because one of the two values of y_j is closer to $\frac{1}{2}$ than x_i.

However, we can easily extend the proposition to cover this case. Note that if $|x_j - \frac{1}{2}| = |x_i - \frac{1}{2}|$, then $x_j = 1 - x_i$. The classification problem is invariant under the transformation $x_i \to 1 - x_i$, and in particular,

$$V(\ldots, x_i, \ldots, \tau) = V(\ldots, 1 - x_i, \ldots, \tau). \tag{114}$$

Furthermore, if $p(y_i \mid x_i) > 0$, then $p(1 - y_i \mid 1 - x_i) > 0$. It follows that

$$V_i(\ldots, x_i, \ldots, \tau) = V_i(\ldots, 1 - x_i, \ldots, \tau). \tag{115}$$

As a consequence of this and the symmetry of V, we find that $|x_j - \frac{1}{2}| = |x_i - \frac{1}{2}|$ implies that

$$V_i(x, \tau) = V_j(x, \tau). \tag{116}$$

Thus, $i, j \in \Phi(x)$ implies that $V_i(x,\tau) = V_j(x,\tau)$. This is sufficient to extend the proposition because if $V_j(x,t) > V_i(x,t)$, then both $x_i \neq x_j$ and $|x_j - \frac{1}{2}| \neq |x_i - \frac{1}{2}|$. Thus, we can apply the earlier argument to show that $i \in \Phi(x_1, \ldots, y_j, \ldots, x_n)$. Consequently, the third and final condition of Corollary 1 holds so that Φ is optimal.

E. Proof of Proposition 4

The optimality of Proposition 4 will be established by verifying that the three conditions of Corollary 1 hold.

First, note that the distributions p and q are commutative since

$$\sum_{\eta} p(\xi \mid \eta) q(\eta \mid x) = \sum_{\eta} q(\xi \mid \eta) p(\eta \mid x) = \begin{cases} 1 & \text{if} \quad \xi = x + 2\Lambda_p - \dfrac{\Lambda_h^2}{\Lambda_h + \Lambda_p} \\ 0 & \text{otherwise.} \end{cases} \tag{117}$$

Moreover,

$$\sum_{y_i} r(y_i) p(y_i \mid x_i) - r(x_i) = \begin{cases} 0 & \text{if} \quad x_i \le \Lambda_h - \Lambda_p \\ 1 & \text{if} \quad \Lambda_h - \Lambda_p < x_i \le \Lambda_h - \Lambda_p + \dfrac{\Lambda_h^2}{\Lambda_h + \Lambda_p} \\ 0 & \text{if} \quad \Lambda_h - \Lambda_p + \dfrac{\Lambda_h^2}{\Lambda_h + \Lambda_p} < x_i. \end{cases} \tag{118}$$

Thus, if $i \in \Phi$, then for all $j \neq i$, either $x_i \leq x_j$ or $x_j \leq \Lambda_h - \Lambda_p$ so that

$$\sum_{y_i} r(y_i)p(y_i \mid x_i) - r(x_i) \geq \sum_{y_j} r(y_j)p(y_j \mid x_i) - r(x_i), \tag{119}$$

and the second condition of Corollary 1 holds. Finally, if

$$i \in \Phi(x_1, \ldots, x_i, \ldots, x_j, \ldots, x_n, t) \tag{120}$$

$x_j \neq x_i$ then $p(y_j \mid x_j) > 0$ implies one of the following. Either $x_i < x_j$ or $x_j \leq \Lambda_h - \Lambda_p$. In the first case, there exists $m \in \{0, 1, 2, \ldots\}$ so that $y_j - x_i = m\Lambda_h^2/(\Lambda_h + \Lambda_p) \geq 0$. In the second case, $y_j \leq x_j \leq \Lambda_h - \Lambda_p$. In either case,

$$i \in \Phi(x_1, \ldots, x_i, \ldots, y_j, \ldots, x_n, t+1). \tag{121}$$

and the third and final condition of Corollary 1 holds, and the strategy set is optimal.

REFERENCES

[1] D. A. Berry and B. Fristedt
Bandit Problems: Sequential Allocation of Experiments.
Chapman and Hall, 1985.

[2] D. P. Bertsekas
Dynamic Programming and Optimal Control.
Athena Scientific, Belmont, MA, 2001.

[3] D. A. Castanon
Optimal search strategies in dynamic hypothesis testing.
IEEE Transactions on Systems, Man, and Cybernetics, **25**, 7 (1995), 1130–1138.

[4] J. C. Gittins
Bandit processes and dynamic allocation indices.
Journal of the Royal Statistical Society: Series B (Methodological), **41**, 2 (1979), 148–177.

[5] S. Howard, S. Suvorova and B. Moran
Optimal policy for scheduling of Gauss-Markov systems.
In *Proceedings of the International Conference on Information Fusion*, 2004.

[6] O. P. Kreidl and T. M. Frazier
Feedback control applied to survivability: A host-based autonomic defense system.
IEEE Transactions on Reliability, **53**, 1 (2004), 148–166.

[7] V. Krishnamurthy
Algorithms for optimal scheduling and management of hidden Markov model sensors.
IEEE Transactions on Signal Processing, **50**, 6 (2002), 1382–1397.

[8] V. Krishnamurthy and R. J. Evans
Hidden Markov model multiarm bandits: A methodology for beam scheduling in multitarget tracking.
IEEE Transactions on Signal Processing, **49**, 12 (2001), 2893–2908.

[9] S. Musick and R. Malhotra
Chasing the elusive sensor manager.
In *Proceedings of the IEEE National Aerospace and Electronics Conference*, vol. 1, 1994, 606–613.

[10] R. Popoli
The sensor management imperative.
In Yaakov Bar-Shalom (Ed.), *Multitarget-Multisensor Tracking: Applications and Advances*, vol. 2, Artech House, 1992, 325–392.

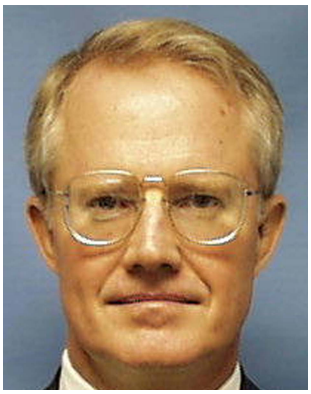
[11] M. K. Schneider, G. L. Mealy and F. M. Pait
Closing the loop in sensor fusion systems: Stochastic dynamic programming approaches.
In *Proceedings of the American Control Conference*, 2004.

[12] M. K. Schneider, A. Nedich, D. A. Castanon and R. B. Washburn
Approximation methods for Markov decision problems in sensor management.
BAE Systems, Burlington, MA, Technical Report TR-1620, 2005.

[13] R. B. Washburn, M. K. Schneider and J. J. Fox
Stochastic dynamic programming based approaches to sensor resource management.
In *Proceedings of 5th Iternational Conference on Information Fusion*, 2002, 608–615.

[14] R. R. Weber and G. Weiss
On an index policy for restless bandits.
Journal of Applied Probability, **27** (1990), 637–648.

[15] P. Whittle
Restless bandits: Activity allocation in a changing world.
Journal of Applied Probability, **25A** (1988), 287–298.



Robert B. Washburn received his B.S. in 1973 from Yale University in mathematics and his Ph.D. in 1979 from the Massachusetts Institute of Technology in applied mathematics.

He has been a principal scientist at Parietal-Systems, Inc. (PSI) since January 2006. At PSI he is developing algorithms for constraint-based data association, sensor resource management, and pattern recognition for information fusion applications. Before joining PSI, Dr. Washburn worked at ALPHATECH for 23 years, where he was a Chief Scientist and during which time he led research and development in several areas of information fusion and sensor management. His work included the development of multiple hypothesis tracking (MHT) and correlation algorithms for several applications, and development and application of multiresolution statistical image fusion algorithms and performance estimation theory for model-based target recognition algorithms. He led the efforts on sensor resource management, using approximate dynamic programming approaches to develop efficient control algorithms for pointing sensors and selecting sensor modes for sensing multiple targets. Dr. Washburn is a member of the IEEE, American Mathematical Society, and Society for Industrial and Applied Mathematics.

Michael K. Schneider received his B.S.E. degree in electrical engineering and Certificate in applied and computational mathematics (Summa Cum Laude) from Princeton University in 1994, and his M.S. and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology in 1996 and 2001, respectively.

He is a lead research engineer at BAE Systems Advanced Information Technologies (BAE AIT) where, since 2001, he has been working on problems in information fusion. At BAE AIT, he has been developing combinatorial optimization algorithms for sensor resource management as well as algorithms for signature-aided tracking and analysis of patterns in historical track data. He is a member of IEEE, SIAM, and INFORMS.