Non-Myopic Multi-Aspect Sensing with Partially Observable Markov Decision Processes

Shihao Ji, Ronald Parr and Lawrence Carin, *Fellow, IEEE*

***Abstract*— We consider the problem of sensing a concealed or distant target by interrogation from multiple sensors situated on a single platform. The available actions that may be taken are selection of the next relative target-platform orientation and the next sensor to be deployed. The target is modeled in terms of a set of states, each state representing a contiguous set of target-sensor orientations over which the scattering physics is relatively stationary. The sequence of states sampled at multiple target-sensor orientations may be modeled as a Markov process. The sensor only has access to the scattered fields, without knowledge of the particular state being sampled, and therefore the problem is modeled as a *partially observable* Markov decision process (POMDP). The POMDP yields a policy, in which the belief state at any point is mapped to a corresponding action. The non-myopic policy is compared to an approximate myopic approach, with example results presented for measured underwater acoustic scattering data.**

***Index Terms*— Multi-aspect sensing, partially observable Markov decision processes (POMDPs), hidden Markov models (HMMs), non-myopic algorithms.**

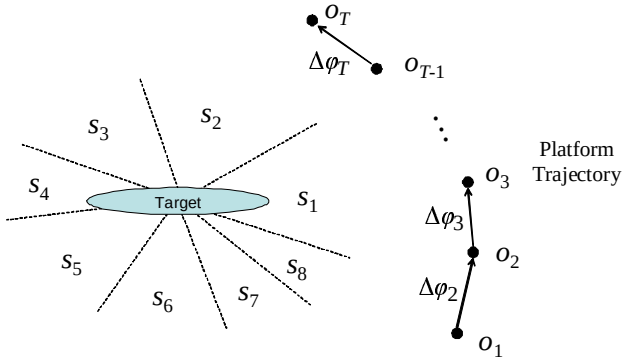

Fig. 1. Multi-aspect sensing of a concealed target. The kth state s_k is a contiguous set of target-sensor orientations over which the scattered fields are approximately stationary ($K=8$ states are shown). Here T observations are performed, $\{o_1, o_2, \cdots, o_T\}$, as performed at a sequence of *relative* sensor angular positions, where $\Delta\varphi_{t+1} = \varphi_{t+1} - \varphi_t$ are orientations.

I. Introduction

The integration of sensing and processing is of increasing importance for many applications, including new unmanned sensing platforms that have the capacity to adapt to their environment [1]. The problem may be posed as one of sensor management. Specifically, given particular sensor assets and previously collected data, one may ask which data should be collected next to best advance a sensing mission. Techniques that have been applied to this problem include ideas from the theory of optimal experiments [2], wherein one may be interested in minimizing the uncertainty (entropy) of parameters estimated within an inversion [3]–[5]. Most of these previous studies have been myopic [2]–[5], in that they seek to perform the next measurement that is most informative, for example in terms of a measure of entropy [2], without considering how the measurement may affect those that come subsequently.

Partially observable Markov decision processes (POMDPs) [6]–[9] are well suited to non-myopic sensing problems, when the underlying physics supports a Markov representation. It has been demonstrated previously, with fixed sensor actions, that sensing a target from multiple target-sensor orientations may be modeled via a hidden Markov model (HMM) [10]. Each state of the HMM corresponds to a contiguous set of target-sensor orientations for which the scattering physics is relatively stationary (see Fig. 1). When the sensor interrogates a given target from a sequence of target-sensor orientations, it inherently samples different target states. For most problems of interest the target is either distant or concealed, and therefore the underlying states are "hidden". The sensor does have access to the associated scattered fields, thereby motivating an HMM representation. In the HMM formulation the sensing actions are assumed fixed (e.g., a constant relative change in target-sensor orientation), and therefore there is limited opportunity for adaptive sensing. In this paper we extend the HMM formalism to a POMDP, yielding a natural and flexible adaptive-sensing framework.

We assume access to a single platform that may carry multiple sensors. The objective is to classify the target under interrogation based upon multi-aspect sensing data. There are three questions that may be asked, based on previous observations and on underlying models of the targets of interest: (i) which relative target-platform orientation should be considered next, (ii) which sensor should be deployed next, and (iii) when should the sensing be terminated and a classification decision made?

It is assumed that a model is available for all targets that may be interrogated, and the objective is to perform classification (the model is designed assuming access to training data, and HMM learning algorithms are employed [10]). For this classification task the POMDP is here formulated in terms of Bayes risk [11], with C_{uv} representing the cost of declaring target u when actually the target under interrogation is target v. Using the same units as associated with C_{uv}, we also define a cost for each class of sensing action. After a set of sensing actions and observations the sensor may utilize the belief state [7] to quantify the probability that the target under interrogation corresponds to target u. The POMDP

S. Ji, L. Carin are with the Department of Electrical and Computer Engineering, and R. Parr is with the Department of Computer Science, Duke University, Box 90291, Durham, NC 27708-0291.

yields a non-myopic policy for the optimal sensor action given the belief state, where here the sensor actions correspond to defining the next relative target-sensor orientation and the next sensor to deploy. In addition, the POMDP gives a policy for when the belief state indicates sensing should be terminated and a classification decision made. The latter is implicitly manifested when the expected future reduction in Bayes risk is not justified by the expected future cost in sensing actions. A summary of actions, states and costs is provided in Table I.

Learning a non-myopic policy may be computationally expensive [6], [7], [12]; however, this is an offline calculation based on the underlying target models. Once the policy has been so learned, the actual sensing may be performed in near real time, limited only by the time required to compute the belief state [7] and then map this into the corresponding (pre-optimized) sensor action.

The related technique of multi-armed bandits [13], [14] has been considered for similar sensing problems. POMDPs have also been applied to the problem of multi-sensor target tracking [8]. The general problem of employing POMDPs for multi-sensor classification of multiple targets is discussed in [15], [16], applied to large problems but with simplified target models. Krishnamurthy and colleagues [17], [18] have also recently employed POMDPs for several other sensing problems. Our application of the POMDP framework to the problem of underwater acoustic scatter from five elastic targets is, to our knowledge, the first successful application of POMDPs to a problem of this scale using actual, measured scattering data. This work has significant utility in the context of sensing underwater targets via sensors deployed on unmanned underwater vehicles (UUVs).

The remainder of the paper is organized as follows. In Sec. II we introduce the POMDP formulation for multi-aspect, multi-sensor interrogation of a concealed or distant target. The non-myopic formulation is compared to an approximate myopic framework. In Sec. III we present example results for acoustic sensing of underwater elastic targets, based on measured scattering data, with comparisons between myopic and non-myopic strategies. Conclusions are provided in Sec. IV.

II. POMDP Formulation

A. Markov model of target sensing

The scattering physics from a complex target is typically a strong function of the target-sensor orientation [10]. However, there are generally contiguous ranges of target-sensor orientations for which the scattering physics is relatively stationary. Each such set of angles is termed a state [10]. To simplify notation we assume that the targets of interest are rotationally symmetric, and the scattered fields are observed in a plane bisecting the axis of symmetry (see Fig. 1). Consequently, at a fixed radial distance r from the target center the scattered fields are characterized by a single angle φ. For a target with K states, the states are defined by the set of angles $\{\varphi_0, \varphi_1, \cdots, \varphi_K\}$, with the kth state corresponding to the contiguous range of angles $\varphi \in [\varphi_{k-1}, \varphi_k]$.

Assume a sequence of measurements is performed, at a sequence of target-sensor orientations. The particular state being interrogated at a given time is "hidden", because the target is distant or concealed. It is assumed that, on consecutive measurements, the probability of transitioning from any given state to another state may be modeled as a Markov process. The corresponding state-transition probabilities are modeled as follows. Let $d_{i,j}$ represent the shortest angular distance to travel in a prescribed direction, i.e. clockwise or counter-clockwise, from the center of state s_i to the center of state s_j. Further, assume that $\Delta\varphi \geq 0$ represents the change in the relative angular position on consecutive measurements, performed in the same angular direction as used to define $d_{i,j}$. The probability of transitioning from state s_i to state s_j on consecutive measurements separated by $\Delta\varphi$ is defined as

$$p(s_j|s_i, \Delta\varphi) \equiv \frac{w_j(d_{i,j} - \Delta\varphi)}{\sum_{j=1}^{K} w_j(d_{i,j} - \Delta\varphi)} \tag{1}$$

where

$$w_j(\varphi) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{\varphi^2}{2\sigma_j^2}\right) \tag{2}$$

with $\sigma_j = (\varphi_j - \varphi_{j-1})/2$. To simplify the above analysis we have assumed that the sensor always moves in a fixed direction (clockwise or counter-clockwise). However, in practice the actual direction of sensor motion may be dictated by which path is shortest, e.g. it is easier to move 5° counter clockwise than clockwise 355°.

Considering (1) and (2) in greater detail, we note that the likelihood of transitioning from state i to state j, is maximized when $\Delta\varphi = d_{i,j}$, corresponding to transitioning an angular distance commensurate with the distance between the centers of these two states. Assume now that measurement t is performed in state i, and that the next measurement at time $t+1$ is performed at an angular displacement $\Delta\varphi \neq d_{i,j}$. As $|\Delta\varphi - d_{i,j}|$ increases, the likelihood of transitioning from state i to state j diminishes, as defined in (1) and (2). The rate of which this likelihood diminishes is dictated by the angular extent of state j relative to $|\Delta\varphi - d_{i,j}|$, since for simplicity it is assumed that the measurement at time t was performed in the center of state i.

By construction the target states constitute a range of target-sensor orientations for which the scattering physics is stationary. This is represented by defining $p(o|s_k, m)$, quantifying the probability of observing o in state s_k, given that sensor m was deployed, where $m \in \{1, 2, \cdots, M\}$ for M sensors. We have assumed that the state decomposition is independent of the sensor deployed. For the problem considered in Sec. III this is a valid assumption, but it may require generalization for disparate sensor types.

When performing sensing in the above setting the sensor has the opportunity to choose among different actions a, where here an action corresponds to selecting a relative angle $\Delta\varphi \geq 0$ for movement of the platform, and deployment of one of the M sensors for collection of scattering data (see Table I). We may therefore define the probabilities $p(s_j|s_i, a)$ and $p(o|s_k, a)$, which generalize the expressions introduced above. In addition, we may introduce the probability π_k to represent the probability of being in state s_k on the first observation. If

TABLE I

SUMMARY OF THE POMDP ACTIONS, STATES AND COSTS. THE SENSING COSTS $c(m)$ AND CLASSIFICATION COSTS C_{uv} MUST BE IN THE SAME UNITS. WHEN A CLASSIFICATION DECISION IS MADE THE MODEL RANDOMLY TRANSITIONS TO A NEW TARGET AND ORIENTATION (RESET FORMULATION) OR IT TRANSITIONS TO AN ABSORBING STATE (SEE FIG. 2), AND THEREFORE THE TERMINAL CLASSIFICATION STATES s_{uv} ARE NOTIONAL.

Actions	States	Cost
Sensing Action: • Move platform angle $\Delta\varphi$; • Perform measurement with one of M sensors.	$\mathcal{S} = \{s_k^{(n)}, \forall\, k, n\}$ Target states k across all targets $n = \{1, 2, \cdots, N\}$.	$c(m)$, m representing one of the M possible sensors (independent of target state visited).
Classification Action: • Stop sensing, declare object under test to be one member from set $\{1, 2, \cdots, N\}$.	s_{uv}, corresponding to declaring target u when in reality target v is being sensed; both u and v are members of the set $\{1, 2, \cdots, N\}$.	C_{uv}, for classification state s_{uv}. In terms of target states s in $\mathcal{S}$, $c(s, a = u) = C_{uv}$ for all s associated with target v.

each target orientation is uniformly likely, we may define

$$\pi_k = \frac{\varphi_k - \varphi_{k-1}}{2\pi} \tag{3}$$

B. Multi-target belief state

In the above discussion we have introduced the statistical parameters needed to characterize a single target. In general each target will have a distinct number of states and a distinct state decomposition. We therefore employ the notation $p(s_j^{(n)}|s_i^{(n)}, a)$, $p(o|s_k^{(n)}, a)$ and $\pi_k^{(n)}$ to represent the parameters for target n; for example, $s_k^{(n)}$ represents the kth state of target n.

After performing a sequence of T actions and making T observations (the first action only involves selecting a sensor and making a measurement, with the original target-sensor orientation uniformly distributed as above), we may compute the belief state for any state $s \in \mathcal{S} = \{s_k^{(n)}, \forall\, k, n\}$ as

$$b_T(s|o_1, \cdots, o_T, a_1, \cdots, a_T) = p(s|o_T, a_T, b_{T-1}) \tag{4}$$

where (4) reflects that the belief state b_{T-1} is a sufficient statistic for $\{a_1, \cdots, a_{T-1}, o_1, \cdots, o_{T-1}\}$ [6], [7]. Note that the belief state is defined across the states from all targets, and it may be computed via [7]

$$\begin{aligned} b_T(s') &= \frac{p(o_T|s', a_T, b_{T-1})p(s'|a_T, b_{T-1})}{p(o_T|a_T, b_{T-1})} \\ &= \frac{p(o_T|s', a_T, b_{T-1})\sum_s p(s'|a_T, b_{T-1}, s)p(s|a_T, b_{T-1})}{p(o_T|a_T, b_{T-1})} \\ &= \frac{p(o_T|s', a_T)\sum_s p(s'|a_T, s)b_{T-1}(s)}{p(o_T|a_T, b_{T-1})} \end{aligned} \tag{5}$$

where the denominator $p(o_T|a, b_{T-1})$ may be viewed as a normalization constant, independent of s', allowing $b_T(s')$ to sum to one.

After T actions and observations we may use (5) to compute the probability that a given state, across all N targets, is being observed. The belief state in (5) may also be used to compute the probability that target n is being interrogated, with the result

$$p(n|o_1, \cdots, o_T, a_1, \cdots, a_T) = p(n|b_T) = \sum_{s \in S_n} b_T(s) \tag{6}$$

where S_n denotes the set of states associated with target n.

Let C_{uv} denote the cost of declaring the object under interrogation to be target u, when in reality it is target v, where u and v are members of the set $\{1, 2, \cdots, N\}$, defining the N targets of interest. After T actions and observations, target classification may be effected by minimizing the Bayes risk, i.e., we declare the target

$$\begin{aligned} \text{Target} &= \arg\min_u \textstyle\sum_{v=1}^N C_{uv} p(v|b_T) \\ &= \arg\min_u \textstyle\sum_{v=1}^N C_{uv} \sum_{s \in S_v} b_T(s) \end{aligned} \tag{7}$$

Therefore, a classification may be performed at any point in the sensing process using the belief state $b_T(s)$.

As discussed in detail below, the POMDP sensing construction is designed to weigh the expected cost of performing future sensing actions with the expected future reduction in the Bayes risk. When the cost of sensing is not justified by the expected reduction in risk, a classification decision is made, using the belief state.

C. Bayes-risk POMDP formulation

In addition to the aforementioned sensing actions (selection of the relative platform displacement $\Delta\varphi$ and selection of the sensor), we introduce a distinct set of terminal actions, where here the action is defined by terminating sensing and declaring that the target under interrogation comes from one of the targets $n \in \{1, 2, \cdots, N\}$.

Costs are now defined for the sensing and classification actions. The sensing actions are defined by the cost of deploying the associated sensor and the cost of moving a relative angle displacement $\Delta\varphi$. There are many ways this cost may be defined, and further details are provided in Sec. III when presenting example results. Note, however, that the sensing cost is independent of which particular target state is being interrogated, since our ultimate objective is target classification; we do not have a goal of visiting particular target states (which, in other settings, is a common objective of robot navigation by using POMDPs [7], [9], [19]). With regard to the terminal classification action, there are N^2 terminal states that may be visited. Terminal state s_{uv} is defined by taking the action of declaring that the object under interrogation is target u when in reality it is target v; the cost of state s_{uv} is

C_{uv}, as defined in the context of the Bayes risk in (7). The sensing costs and Bayes-risk costs must be in the same units.

Making the above discussion quantitative, let $c(s,a)$ represent the immediate cost of performing action a when in state s. For the sensing actions indicated above $c(s,a)$ is independent of the target state being interrogated (independent of s) and is only dependent on the type of sensing action taken. For the terminal classification action, defined by taking the action of declaring target u, we have

$$c(s, a = u) = C_{uv}, \forall\ s \in S_v \tag{8}$$

The expected immediate cost of taking action a in belief state $b(s)$ is

$$C(b,a) = \sum_s b(s)c(s,a) \tag{9}$$

For sensing actions, that have a cost independent to s, the expected cost is simply the known cost of performing the measurement. For the terminal classification action the expected cost is

$$C(b, a = u) = \sum_{v=1}^{N} \sum_{s \in S_v} b(s) C_{uv} = \sum_{v=1}^{N} C_{uv} p(v|b) \tag{10}$$

and therefore the optimal terminal action for a given belief state b is to choose that target u that minimizes the Bayes risk. Of interest, therefore, is to learn a policy that defines when a belief state b warrants taking such a terminal classification action; when classification is not warranted, the desired policy defines what sensing actions should be executed for the associated belief state b. The POMDP parameters are summarized in Table I.

D. Non-myopic policy estimation

The goal of a policy is to minimize the expected discounted infinite-horizon cost $\mathbb{E}[\sum_{k=0}^{\infty} \gamma^k C(b_k, a_k)]$, which yields Bellman's dynamic programming recursion

$$\chi(b) = \min_a \left[C(b,a) + \gamma \sum_{b' \in \mathcal{B}} p(b'|b,a) \chi(b') \right] \tag{11}$$

where $\gamma \in [0,1)$ is a discount factor that quantifies the degree to which future costs are discounted with respect to immediate costs, and $\mathcal{B}$ defines the set of all possible belief states. In (11) the action a that minimizes $\chi(b)$ defines the optimal policy (i.e., the mapping from belief states to actions, $b \rightarrow a$). When optimized exactly for a finite number of iterations, the cost function is piecewise linear and concave in the belief space [6], [7].

After t consecutive iterations of (11) we have

$$\chi_t(b) = \min_a \left[C(b,a) + \gamma \sum_{b' \in \mathcal{B}} p(b'|b,a) \chi_{t-1}(b') \right] \tag{12}$$

where $\chi_t(b)$ represents the cost of taking the optimal action for belief state b at t steps from the horizon. One may show that $\chi_t(b) = \min_{\alpha \in \Gamma_t} \sum_{s \in \mathcal{S}} \alpha(s) b(s)$, where the α-vectors come from a set $\Gamma_t = \{\alpha_1, \alpha_2, \cdots, \alpha_r\}$, where in general r is not known *a priori* and is a function of t. Each α-vector defines an $|S|$-dimensional hyperplane, and each is associated with an action, defining the best immediate policy assuming optimal behavior for the following $t-1$ steps. The cost at iteration t may be computed by "backing up" one step from the solution $t-1$ steps from the horizon [6], [7], [9]. Recalling that $\chi_{t-1}(b) = \min_{\alpha \in \Gamma_{t-1}} \sum_{s \in \mathcal{S}} \alpha(s) b(s)$, we have

$$\chi_t(b) = \min_{a \in \mathcal{A}} \Bigg[C(b,a) + \gamma \sum_{o \in \mathcal{O}} \min_{\alpha \in \Gamma_{t-1}} \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} p(s'|s,a) \\ p(o|s',a) \alpha(s') b(s) \Bigg] \tag{13}$$

where $\mathcal{A}$ represents the set of possible actions (both for sensing and making classifications), and $\mathcal{O}$ represents the set of possible observations. As discussed in Sec. III when presenting results, the set of actions is discretized, as are the observations, such that both constitute a finite set.

Iterative solution of (13) corresponds to sequential updating of the set of α-vectors, via a sequence of backup steps away from the horizon. There has been much research directed toward development of approximate techniques (e.g. [9], [19], [20]) for solving (13), since exact solutions (e.g. [6], [7], [12]) are only possible for problems composed of a small number of actions and states. The focus of this paper is on developing a POMDP formulation for multi-aspect, multi-sensor target classification, not on the details of approximately solving (13). We note, however, that in the results to follow we have utilized the point-based value iteration (PBVI) algorithm [9], which has demonstrated excellent policy design on complex benchmark problems. As discussed in Sec. III, PBVI allows development of good policies for problems of interest here (the examples considered involve a relatively large number of actions and observations, and a modest number of states). It is known that PBVI has a time complexity of $O(|B||\mathcal{S}||\mathcal{A}||\mathcal{O}||\Gamma_{t-1}|$ [9] for the t-th iteration, where $|V|$ denotes the size of a set V, and $|B|$ is the size of belief-point set that is used for approximated planning.

E. Random reset vs. absorbing state

There are many ways in which one may formulate the POMDP, two of which are indicated in Fig. 2. In one formulation, after making a classification decision the underlying model randomly selects another target (at a random target-sensor orientation) and the sensing process proceeds. In the other formulation after performing a classification the model transitions into an absorbing state, and no further sensing is performed. It is interesting to examine how these two formulations differ.

One may view the sensing process as a sequence of questions asked of the unknown target by the sensor, with the scattering physics providing the question answers. Specifically, the sensor asks: "For this unknown target, what would the scattered fields look like if I moved over there to perform a measurement?". To obtain the answer to this question the sensor physically moves and performs the associated measurement. The sensor recognizes that the ultimate objective is to perform classification, and that a cost is assigned to each question. The objective is therefore to ask the fewest number of sensing questions, with the goal of minimizing the ultimate

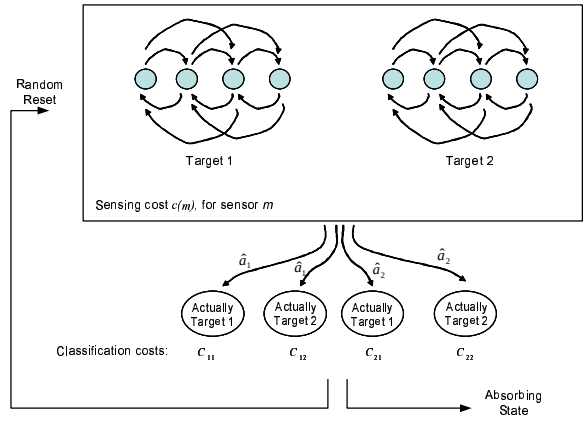

Fig. 2. Schematic of the POMDP formulation used in policy design, for simple case of two targets. The sensing actions (top box) correspond to sampling data from a sequence of target states. Actions $\hat{a}_n$ correspond to stopping sensing and declaring target n. Depending on the formulation, the classification action is followed by an absorbing state, or the algorithm randomly resets on another target and target-sensor orientation (see Sec. II-E).

cost of the classification decision (accounting for the costs of inaccurate classifications).

The reset formulation in Fig. 2 gives the sensor more flexibility in optimally asking questions and performing classifications, within a cost budget. Specifically, the sensor may discern that at give classification problem is very "hard" (i.e., prior to sensing it may be known that the object under test is one of N targets, and after a sequence of measurements the sensor may have winnowed this down to two possible targets; however discerning between these final two targets may be a significant challenge, requiring many sensing actions). Once the complexity of the "problem" is understood, the optimal thing to do within this formulation may be to stop asking questions and give the best classification answer possible, moving on to the next (randomly selected) classification problem, with the hope that it is "easier". While the sensor may not do as well in classifying the "hard" classification problems, overall it may reduce costs.

By contrast, if the sensor transitions into an absorbing state after performing classification (see Fig. 2), it cannot "opt out" of a "hard" sensing problem, with the hope of being given an "easier" problem subsequently. Therefore, with the absorbing-state formulation we might expect that the sensor will on average ask more questions (perform more sensing actions), with the goal of reducing costs on the ultimate classification task (which it cannot opt out of if it is "hard").

The above characteristics of the two non-myopic POMDP formulations are observed in Sec. III, when presenting results on measured acoustic-scattering data. The appropriateness of these formulations depends on the problem. The reset formulation may be appropriate for scenarios in which there are many targets to be classified and for which there is a finite sensing budget. This formulation may lead to more errors on the "hard" classification examples, but this may be counterbalanced by the benefit of visiting and correctly classifying more targets overall in a given amount of time.

F. Myopic alterative

The POMDP formulation discussed above yields a non-myopic policy, mapping belief states into actions. It is of interest to consider an alternative myopic approach, for comparison.

After T sensing actions and observations there is a belief state $b_T(s)$, with which one may compute the Bayes risk associated with making a classification decision, as defined in (7). The expected Bayes risk after new *sensing* action a_{T+1} may be computed as

$$R_E(b_T, a_{T+1}) = \sum_{o\in\mathcal{O}} \min_u \Bigg[\sum_{v=1}^{N} C_{uv} \sum_{s'\in S_v} \sum_{s\in\mathcal{S}} p(o|s', a_{T+1}) \; p(s'|s, a_{T+1}) b_T(s)\Bigg] \tag{14}$$

One may compute the myopic cost associated with a sensing action a_{T+1} as $\hat{C}(b_T, a_{T+1}) = c(a_{T+1}) - [R(b_T) - R_E(b_T, a_{T+1})]$, and action a_{T+1} is selected as to minimize $\hat{C}$, with $R(b_T) = \min_u \left[\sum_{v=1}^{N} C_{uv} \sum_{s\in S_v} b_T(s)\right]$. When $\hat{C}$ is positive the costs of sensing exceed the expected reduction in risk; the sensing is then terminated and a classification made. For the results in Sec. III this myopic strategy is performed exactly: for each b_T and a_{T+1} under consideration, (14) is computed directly before selecting the next action (there is not a policy learned "offline", as in the non-myopic case).

III. Example Results

A. Targets and sensor considered

We consider the problem of classifying between five elastic targets, based on underwater acoustic scattering data. This problem is of importance for multi-aspect sensing and classification of underwater targets via an unmanned underwater vehicle (UUV). Details on the targets, scattering data and features are provided in [10], [21]. We here provide a brief summary. The physical descriptions of the five targets are provided in Fig. 3. The scattered fields are observed as a function of angle, with data sampled in 1° increments. The time-domain scattered fields from each target are processed using matching pursuits [10], [21], [22], from which a set of features are extracted. The feature vectors are aggregated across all target-sensor orientations and target types, and vector quantization (VQ) is performed [23]. When performing POMDP design and policy implementation, a feature vector under test is mapped to one of the codes. Therefore, the observations o are discrete elements from the associated VQ codebook, and $p(o|s_k^{(n)}, a)$ is represented via a probability mass function (pmf). The discount factor is set at $\gamma = 0.95$ for all POMDP results presented below.

Due to the target symmetry reflected in Fig. 3, the target states are uniquely defined over a 90° segment (see Fig. 4), with the remaining states at other angles manifested by employing symmetry. For the acoustic frequencies considered, we have found that each of the five targets is represented well by five distinct states, as indicated schematically in Fig. 5; the proper number of states may be computed using model-selection techniques [24]. The target-dependent state-transition

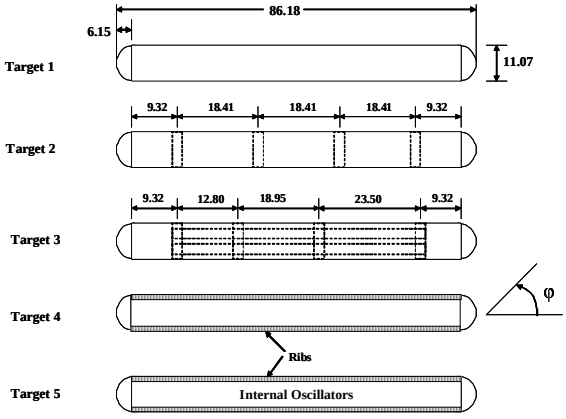

Fig. 3. Five elastic shell targets, with all units in meters. The scattered fields are observed as a function of angle at a fixed distance from the target center, in a plane bisecting the target axis. The nominal external sizes and shapes of the targets are the same, with the acoustic scattering distinguished by the internal structure.

probabilities $p(s_j^{(n)}|s_i^{(n)}, \Delta\varphi)$ and $\pi_k^{(n)}$ initial-state probabilities and may be computed using HMM training procedures, such as the Baum-Welch and Viterbi algorithms [25], [26]; however, for the data considered here we have found that the representations in (1) and (3) are effective (the change in these parameters after training is minimal). We therefore assume that training data are available *a priori* for design of the model parameters needed for the POMDP.

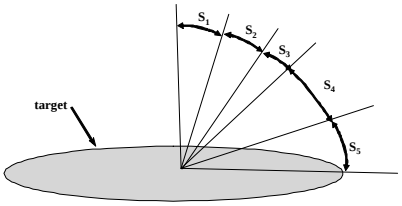

Fig. 4. Example state decomposition for targets.

Concerning the scattering data, in Fig. 5 we plot the magnitude of the measured scattered fields, for each of the five targets, as a function of sensing angle φ (over 360° of target-sensor orientations). Note that the scattering data are relatively similar across targets. We underscore that the classification is not performed using all of the data in Fig. 5, but rather a small number of angular samples from these data. The original scattering data were measured over the frequency band 7.5–45 KHz. To simulate the use of multiple sensors, in the results to follow four distinct "sensors" are manifested by filtering the full frequency spectrum into four subbands using wavelets (details below). While in principle all data may be collected, stored and then processed subsequently, by employing a policy to select from among the subbands there is savings in sensor cost (time and energy for the data collection) and in storage (computer memory). The basic idea also extends naturally to other classes of sensors.

In the results to follow we keep the sensing cost $c(a)$ to a constant, $c(a) = C_s$, for all sensing actions (subband selection and choice of $\Delta\varphi$). The algorithm readily generalizes

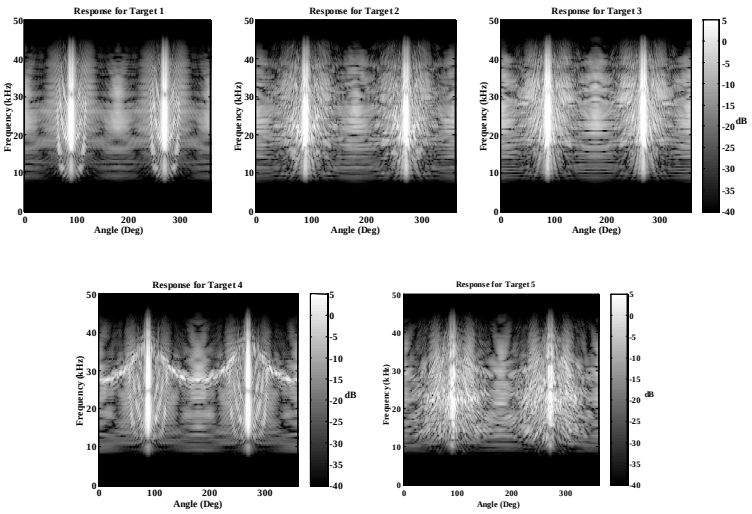

Fig. 5. Scattered fields (magnitude) as a function of sensing angle, for the targets in Fig. 3. The vertical axis represents the frequency dependence of the scattered fields, and the horizontal axis represents the sensor position relative to the target (see Fig. 1)

if one wishes to make $c(a)$ a function of $\Delta\varphi$, for example to favor less extensive angular displacements. One may also make $c(a)$ a function of the particular sensor considered (some sensors may be more expensive to implement than others). For example, in some applications passive sensors are preferred to active sensors, since the latter reveal the sensing asset to the target. The classification costs C_{uv} are represented as C_{uv}=C_c for all $u \neq v$, and $C_{uu} = -10$ (a reward of 10 is obtained upon correct classification). We consider compromises in the classification performance vs. the number of sensor actions by considering different choices of C_s and C_c (implicitly changing the underlying policies).

B. Myopic and non-myopic sensing results, fullband data

We first present a comparison of three algorithms, using the original full-band data: (i) a POMDP policy as computed via PBVI [9], (ii) the greedy (myopic) algorithm in Sec. II-F with a stopping criterion as defined in that section, and (iii) the greedy algorithm with a fixed number of sensing actions T prior to making a classification. With regard to (i), two formulations are considered, one based upon the reset after classification, and the other employing an absorbing state after classification (see Fig. 2 and Sec. II-E). Since the full spectrum of data are used in these results, the only sensing action is selection of the angular displacement $\Delta\varphi$. For these results the number of VQ codes (possible observations) is 25, and the action space is discretized in angular displacements of 5°, with a maximum displacement of 50°.

Classification results are shown in Fig. 6(a) as a function of the average number of actions employed before making a classification decision. Each point for methods (i) and (ii) corresponds to selecting costs C_c. These costs range from $C_c = 15$ for the smallest number of actions, and $C_c = 150$ for the largest number of actions, with $C_s = 1$. The results for method (iii) are for integer number of actions, since the

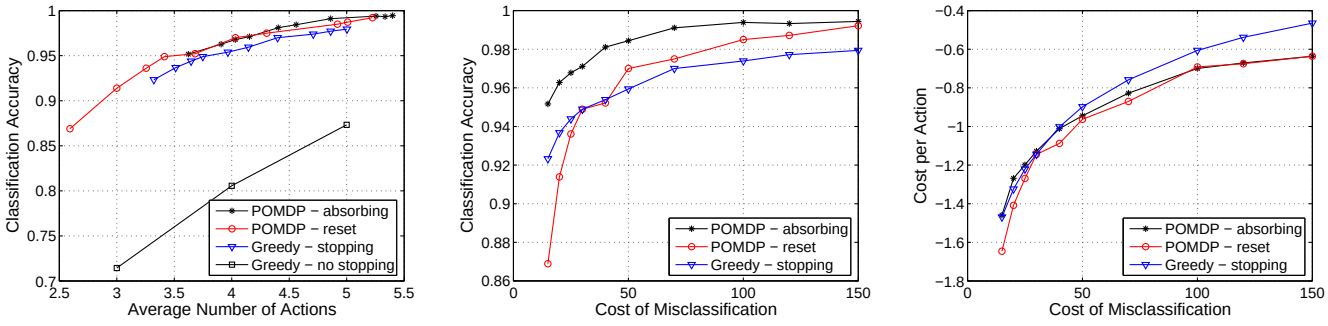

Fig. 6. (a) Classification performance of the four different formulations, as a function of the average number of actions; (b) Classification accuracy of the two POMDP formulations and of the myopic (greedy) formulation with a stopping criterion, as a function of the cost of misclassification; (c) Average cost per action, as a function of the misclassification cost C_c, for the example in Fig. 6(b).

number of actions in this case is fixed (not determined adaptively). The results in Fig. 6(a) are averaged across all possible initializations of the target and target-sensor orientation under consideration.

By comparing the two greedy (myopic) algorithms, one using a stopping criterion and the other with a fixed number of actions, we observe significantly improved performance of the former. If the number of actions is T, the algorithm with a fixed number of actions always performs T measurements (each selected optimally, in a myopic sense). By contrast, the algorithm with the stopping criterion takes an *average* number of actions T. For the "easy" classification decisions the algorithm will often terminate sensing after less than T actions, and for the "harder" cases greater than T actions may be taken (T actions are only taken in an average sense). Based upon the results in Fig. 6(a), this adaptive stopping is important.

From Fig. 6(a) we observe that, for the same number of average actions, the non-myopic POMDP with PBVI [9] yields slightly better performance than the greedy algorithm with an adaptive stopping criterion (using both non-myopic formulations). As the probability of detection increases (increasing C_c), there are an increasing number of average steps required of the myopic algorithm with stopping criterion to achieve the same classification performance as the non-myopic POMDP formulations.

It is also of interest to compare the performance of the two non-myopic POMDP formulations, based on a reset and absorbing terminal state. As suggested in Sec. II-E, it appears that for relatively low classification-error costs (small C_c), when presented with a "hard" sensing problem the reset POMDP formulation stops sensing early, makes the best classification decision it can, and moves to the next target, with the hope that its classification is "easier". Therefore, for small C_c, the reset formulation leads to fewer classification actions and less-accurate classification than the absorbing-state POMDP formulation. As the cost of classification error increases (increasing C_c), the two POMDP formulations become more comparable, since there is a considerable cost associated with opting out of "hard" classification problems and making a less-accurate classification.

Figure 6(b) considers the same problem as Fig. 6(a), but now the horizontal axis quantifies C_c rather than the number of average actions. From Fig. 6(b) we observe a substantial difference in the classification accuracy of the two POMDP formulations for small C_c; while this appears to undermine the utility of the reset formulation, note that Fig. 6(a) indicates that this reduced classification performance has the attendant property of fewer average actions (of importance for sensing many targets with a finite sensing budget).

While the motivation for employing POMDPs is accurate classification within a sensing budget, with this issue addressed in Figs. 6(a) and 6(b), the POMDP is actually formulated to reduce costs (with classification performance coming indirectly, parametrized by the cost of classification errors C_c and the cost of correct and incorrect classification, C_{uu} and C_{uv}, respectively). In Fig. 6(c) we present the average cost per action for the two POMDP formulations and for the myopic (greedy) algorithm with a stopping criterion. Recall that $C_{uu} = -10$, implying that negative average costs reflect that on average correct classifications are being made at a rate that justifies the sensing costs $C_s = 1$. As C_c increases the advantages of the non-myopic POMDP formulations become evident *vis-à-vis* the myopic approach with a stopping criterion. Also note that for small C_c the POMDP with reset provides minimal average cost per action (but higher classification errors), while for large C_c the two POMDPs yield similar average costs per action.

Another important distinction between the two POMDP formulations is that the algorithm with absorbing terminal state actually yields a finite-horizon algorithm, with horizon length dictated by the complexity of the classification task (defined by the target under interrogation and the initial target-sensor orientation). For the absorbing-terminal-state POMDP formulation one may *not* wish to use discounting. When the discount factor was set to $\gamma = 1$ (no discounting), we observed little difference in the absorbing-state formulation performance *vis-à-vis* the $\gamma = 0.95$ used in all other cases. This may be attributed to the relatively small number of sensing actions required to make a classification, for the problem considered.

All computer codes employed in this study were implemented in unoptimized Matlab. However, to give a sense of the

computational complexity, the offline POMDP policy design required 3 hours of CPU with the PBVI algorithm [9], using a Pentium IV with 2.8 GHz CPU. In these computations the PBVI dealt with a total of 25 target states (and an absorbing state, when that formulation was used), 15 possible actions, and 25 possible observations. Selection of the actions when sensing was essentially instantaneous, based on the policy. The myopic algorithm required 0.02 seconds of CPU per action.

C. HMMs, myopic, non-myopic and subband selection

We now consider a case in which four different classes of "sensors" are synthesized from the original fullband data in Fig. 5. The POMDP results are here computed using the reset formulation (see Sec. II-E and Fig. 2); the relative performance of the reset and absorbing-state POMDP formulations are as indicated in the previous subsection. The four subband data come from the low-low, low-high, high-low, and high-high (LL, LH, HL and HH, respectively) outputs of a wavelet transform based on the Daubechies-4 wavelets [27]. We present results for the original fullband data, for each of the subbands considered separately, and when the adaptive algorithm selects from among the four subbands. For cases in which the fullband data are considered, the number of possible observations (VQ codebook size) is 25, while when subband data are used a total of 100 observations (codes) are possible.

We now compare three different algorithms: (i) the POMDP of Sec. II with reset, (ii) a myopic adaptive strategy (Sec. II-F) with a fixed stop criterion of four actions, and (iii) an HMM in which the number of observations is fixed at five. Note that the first observation is performed randomly (by selecting from among the possible targets and the corresponding target-sensor orientations); for the two adaptive algorithms all subsequent measurements are performed adaptively. Therefore, for the adaptive algorithms, when $T-1$ actions are performed, there are a total of T observations.

For (i) and (ii) above we consider several examples of adaptivity: (a) the full or subband sensor spectrum is fixed, and adaptivity occurs in selection of the relative angle $\Delta\varphi$; (b) the relative displacement is fixed at $\Delta\varphi = 5^\circ$ and the adaptive algorithm can select from among the four subbands; and (c) both the frequency subband and angular displacement $\Delta\varphi$ can be selected adaptively. For the POMDP the cost of sensing and classification actions are fixed at $C_s = 1$ and $C_c = 40$.

The results of this study are presented in Table II. Considering first the example for which the subband is fixed, we observe that best HMM results are manifested in the LL band; this same level of relative performance among the frequency subbands is also observed for the two classes of adaptive algorithms. It is interesting to observe that for fixed angular sampling $\Delta\varphi = 5^\circ$ there is substantial advantage found in choosing the sensor bandwidth adaptively. Specifically, the best HMM performance occurs with LL data (86.11% correct classification), while the myopic and POMDP algorithms yield respective correct classification of 90.72% and 94.72% by adaptively selecting from the four subbands. Table II also indicates that when the sensor bandwidth is fixed at LL, LH, HL, HH or fullband the adaptivity in angle provides substantial gains *vis-à-vis* the HMM with fixed angular sampling $\Delta\varphi = 5^\circ$.

TABLE II

PROBABILITY OF CORRECT CLASSIFICATION FOR AN HMM (FIVE OBSERVATIONS, FIXED ACTIONS AND NO ADAPTIVITY), THE POMDP, AND A GREEDY ALGORITHM (SEC. II-F) WITH A FIXED NUMBER OF FOUR ACTIONS (FIVE OBSERVATIONS). THE HMM RESULTS ARE SHOWN IN THE MIDDLE COLUMN (WITH ONLY A SINGLE RESULT IN EACH CASE), FOR FIXED 5° ANGULAR SAMPLING. WHERE TWO RESULTS ARE SHOWN ACTIVE SENSING IS EMPLOYED, WITH THE RESULTS AT LEFT CORRESPONDING TO MYOPIC SENSING WITH FOUR ACTIONS, AND THE RESULTS AT RIGHT CORRESPONDING TO THE POMDP (THE PARENTHESIS DENOTE THE AVERAGE NUMBER OF ACTIONS FOR THE LATTER). THE LL, HL, LH, AND HH DENOTE THE SUBBAND OUTPUTS FROM THE DAUBECHIES-4 [27] WAVELET FILTER APPLIED TO THE FULLBAND DATA (CORRESPONDING HERE TO FOUR DISTINCT "SENSORS"). FOR THE POMDP, THE COST OF PERFORMING A CLASSIFICATION IS $C_c = 40$, WHILE THE COST OF SENSING IS $C_s = 1$.

	Fixed Angular, 5°	Angle Selection
Fixed subband: LL	86.11%	91.94% – 91.33% (2.34)
Fixed subband: HL	72.67%	74.28% – 86.28% (5.08)
Fixed subband: LH	73.72%	80.72% – 91.22% (4.94)
Fixed subband: HH	77.72%	87.50% – 92.67% (3.89)
Fixed Fullband	76.50%	84.67% – 94.61% (4.08)
Subband Selection	90.72% – 94.72% (3.14)	93.50% – 97.17% (2.51)

The results in the bottom-right part of Table II reflect results when both the relative angle and the sensor subband are selected, either non-myopically via the POMDP or myopically as discussed in Sec. II-F. In almost half the number of actions on average, the POMDP manifests a relatively large improvement in classification performance *vis-à-vis* the myopic approach with four actions (five observations).

IV. CONCLUSIONS

A formulation has been presented for adaptive sensing and classification of multiple targets, based on viewing the distant or concealed object from a sequence of orientations. In addition to selecting the relative platform position, the algorithm allows adaptive selection from among a suite of sensors. The partially observable Markov decision process (POMDP) yields a policy, balancing the future costs of sensing with the future expected reduction in the Bayes risk of making a classification decision. In addition to determining a policy for the optimal sensing actions, the policy defines when to stop sensing and make a classification decision.

While the results reported here appear promising, there are several directions for further work. For example, the POMDP classification formulation assumes that models are available for all targets that may be observed when sensing. Underlying models were employed for each of the five targets considered in this study. In many practical applications one may come across targets that have not been seen previously. The questions that may be asked in this setting are: (i) is the target under test one seen previously (for which a model exists);

(ii) if so, which target is it; (iii) if not, which measurements should be performed to learn more about the new target (and possibly build a model for it)? This problem requires a balance of exploration and exploitation, the former interested in learning new information, the latter interested in utilizing the existing information toward a desired end. This balance has been investigated for related problems in reinforcement learning, associated with Markov decision processes (MDPs) [28]. In an MDP the underlying states are observable at all times, and it is of interest to extend these ideas to POMDPs, for which the underlying states are unobservable.

ACKNOWLEDGEMENT

The authors appreciate funding from NSF IIS award 0209088, SAIC, the Sloan Foundation, and a DARPA-ARO MURI on adaptive sensing.

REFERENCES

[1] L. J. Guibas, "Sensing, tracking and reasoning with relations," *IEEE Signal Processing Magazine*, vol. 19, pp. 73–85, Mar. 2002.

[2] V. V. Fedorov, *Theory of Optimal Experiments*. Academic Press, 1972.

[3] Y. Zhang, X. Liao, and L. Carin, "Detection of buried targets via active selection of labeled data: application to sensing subsurface UXO," *IEEE Trans. Geoscience Remote Sensing*, vol. 42, pp. 2535–2543, Nov. 2004.

[4] X. Liao and L. Carin, "Application of the theory of optimal experiments to adaptive electromagnetic-induction sensing of buried targets," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 26, pp. 961–972, Aug. 2004.

[5] P. Whaite and F. P. Ferrie, "Autonomous exploration: driven by uncertainty," *IEEE Trans. Pattern Analysis Mach. Intell.*, vol. 19, no. 3, pp. 193–205, Mar. 1997.

[6] E. J. Sondik, "The optimal control of partially observable Markov processes," Ph.D. dissertation, Standford University, 1971.

[7] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial Intelligence*, vol. 101, pp. 99–134, 1998.

[8] C. Kreucher, K. Kastella, and A. Hero, "Sensor management using an active sensing approach," *Signal Processing*, vol. 85, no. 3, pp. 607–624, Mar. 2005.

[9] J. Pineau, G. Gordon, and S. Thrun, "Point-based value iteration: An anytime algorithm for POMDPs," in *Proc. of Int. Joint Conf. on Artificial Intelligence (IJCAI)*, 2003, pp. 1025–1032.

[10] P. Runkle, P. Bharadwaj, and L. Carin, "Hidden Markov model multi-aspect target classification," *IEEE Trans. Signal Proc.*, vol. 47, pp. 2035–2040, July 1999.

[11] J. M. Bernardo and A. Smith, *Bayesian Theory*. Wiley, 2001.

[12] A. R. Cassandra, M. Littman, and N. Zhang, "Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes," in *Proc. of the 13th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, 1997, pp. 54–61.

[13] V. Krishnamurthy, "Decentralized emission management for low probability of intercept sensor platforms in network centric warfare – a multi-armed bandit approach," *IEEE Trans. Aerospace and Electronic Systems*, vol. 41, no. 1, pp. 133–152, Jan. 2005.

[14] V. Krishnamurthy and R. J. Evans, "Hidden Markov model multiarm bandits: A methodology for beam scheduling in multitarget tracking," *IEEE Trans. Signal Processing*, vol. 49, no. 12, pp. 2893–2908, Dec. 2001.

[15] D. A. Castañon, "Approximate dynamic programming for sensor management," in *Proc. 36th IEEE Conference on Decision and Control*, 1997, pp. 1202–1207.

[16] ——, "Stochastic control bounds on sensor network performance," submitted to IEEE Conference on Decision and Control, 2005.

[17] V. Krishnamurthy, "Algorithms for optimal scheduling and management of hidden Markov model sensors," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1382–1397, June 2002.

[18] R. Evans, V. Krishnamurthy, G. Nair, and L. Sciacca, "Networked sensor management and data rate control for tracking maneuvering targets," *IEEE Trans. Signal Processing*, vol. 53, no. 6, pp. 1979–1991, June 2005.

[19] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling, "Learning policies for partially observable environments: Scaling up," in *Proc. of the 12th International Conference on Machine Learning (ICML)*, 1995, pp. 362–370.

[20] W. S. Lovejoy, "Computationally feasible bounds for partially observed Markov decision processes," *Operations Research*, vol. 39, pp. 162–175, 1991.

[21] P. Runkle, L. Carin, L. Couchman, T. Yoder, and J. Bucaro, "Multi-aspect identification of submerged elastic targets via wave-based matching pursuits and hidden Markov models," *J. Acoustical Soc. Am.*, vol. 106, pp. 605–616, Aug. 1999.

[22] M. McClure and L. Carin, "Matched pursuits with a wave-based dictionary," *IEEE Trans. Signal Proc.*, vol. 45, pp. 2912–2927, Dec. 1997.

[23] R. M. Gray, "Vector quantization," *IEEE ASSP Magazine*, pp. 4–29, Apr. 1984.

[24] M. J. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," in *Bayesian Statistics 7*. Oxford University Press, 2003.

[25] Y. Ephraim and N. Mehrav, "Hidden Markov processes," *IEEE Trans. Information Theory*, vol. 48, no. 6, pp. 1518–1569, June 2002.

[26] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. Information Theory*, vol. 13, pp. 260–269, Apr. 1967.

[27] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[28] R. I. Brafman and M. Tennenholz, "R-max – a general polynomial time algorithm for near-optimal reinforcement learning," *J. Machine Learning Research*, vol. 3, pp. 213–231, 2002.