Joint 48th IEEE Conference on Decision and Control and
28th Chinese Control Conference
Shanghai, P.R. China, December 16-18, 2009

Power Constrained Dynamic Quantizer Design for Multisensor Estimation of HMMs with Unknown Parameters

Nader Ghasemi and Subhrakanti Dey

Abstract—**This paper addresses an estimation problem for hidden Markov models (HMMs) with unknown parameters, where the underlying Markov chain is observed by multiple sensors. The sensors communicate their binary-quantized measurements to a remote fusion centre over noisy fading wireless channels under an average sum transmit power constraint. The fusion centre minimizes the expected state estimation error based on received (possibly erroneous) quantized measurements to determine the optimal quantizer thresholds and transmit powers for the sensors, called the optimal policy, while obtaining strongly consistent parameter estimates using a recursive maximum likelihood (ML) estimation algorithm. The problem is formulated as an adaptive Markov decision process (MDP) problem. To determine an optimal policy, a stationary policy is adapted to the estimated values of the true parameters. The adaptive policy based on the maximum likelihood estimator is shown to be average optimal. A nonstationary value iteration scheme is employed to obtain adaptive optimal policies which has the advantage that the policies are obtained recursively without the need to solve the Bellman optimality equation at each time step. We provide some numerical examples to illustrate the analytical results.**

I. Introduction

In recent years there has been an enormous research effort dedicated to wireless sensor networks (WSNs) due to their wide range of current and potential applications. In detection/estimation applications involving such WSNs, the severe bandwidth constraints, limitations imposed by the fading wireless channels and the energy/power constraints of the small battery-powered sensors have thrown up a new set of challenges. To overcome these limitations, various estimation problems with *quantized* (binary or with a small number of bits) data have been studied, see, e.g., [1], [2].

In this paper, we focus on designing power-efficient binary quantizers for estimation of hidden Markov models (HMM) whose description depend on unknown parameters. The underlying Markov process is observed by multiple sensors which communicate binary-quantized measurements to a central entity, known as the fusion centre, over fading wireless channels modeled by finite-state Markov chains.

Hidden Markov models have long been considered as useful stochastic signal models in a broad range of areas, such as robotics, econometrics, biochemistry, and biology. There are many studies reported in the literature which address state estimation of HMMs with various types of observations and under different constraints (see, e.g., [3], [4], [5], [6]). In particular, a recent study [5] considered designing optimal binary quantizers for state estimation of a general HMM using an *unconstrained* Markov decision process (MDP) approach. Furthermore, in a most recent study [6], the authors addressed the same problem as in [5] using an alternative *constrained* MDP approach which is shown to be more efficient in terms of computations and memory requirements. In these studies, though, it is assumed that the HMM parameters are *known* to the state estimation algorithm.

In most real applications, however, parameters of the HMM are unknown to the state estimator. Compared to those studies [5], [6], in this paper, we explore the problem of *joint* state and parameter estimation of a general HMM with *unknown* parameters using multiple sequences of binary-quantized observations. The optimization problem is formulated as an adaptive Markov decision problem. We propose a coupled algorithm in which state and parameter estimation are performed *jointly* to compute optimal quantizer thresholds and optimal sensor transmit power allocation, called the optimal policy. The first component of this coupled algorithm is an MDP module, called the MDP controller, whose function is to obtain an optimal policy by performing state estimation so as to minimize state estimation error constrained on an average sum power budget across the sensors. The MDP controller uses a nonstationary value iteration (NVI) scheme in order to obtain *adaptive* optimal policies. The NVI scheme adapts the optimal policy to the current estimate of unknown model parameters received from the second component, referred to as the parameter estimator module. This is a learning module, which estimates the model parameters through interaction with both the MDP controller and the underlying dynamical system. The advantage of the NVI scheme is that the policies are obtained in an *iterative* manner without the need to solve the Bellman optimality equation at each time step, which in our case is highly computation-intensive.

Notations: Throughout the paper, $\mathbb{R}$ and $\mathbb{N}$ denote the sets of real numbers and positive integers, respectively. $\mathbb{P}_\phi$ denotes probability distribution, depending on a parameter (vector) ϕ, with respect to some σ-finite measure. In this paper, vector means a column vector and $'$ denotes the transpose notation.

II. Dynamical System Model

Consider a dynamical system whose state evolves according to a discrete-time finite-state homogeneous first-order stationary Markov process $\{X_k\}_{k=1}^{\infty}$ with state space $\mathbb{X}(\phi) = \{\tilde{x}_1(\phi), \cdots, \tilde{x}_n(\phi)\}$ and transition probability matrix $\mathbf{X}(\phi) = [x_{ij}(\phi)]$, where $x_{ij}(\phi) = \mathbb{P}_\phi(X_k = \tilde{x}_j | X_{k-1} = \tilde{x}_i)$ for $i, j = 1, \cdots, n$. The order $n \in \mathbb{N}$ of the Markov process

This work was supported in part by the Australian Research Council (ARC) under grant ARC DP 0985397.

N. Ghasemi and S. Dey are with the Department of Electrical and Electronic Engineering, The University of Melbourne, Parkville, Melbourne, VIC 3010, Australia. n.ghasemi@ee.unimelb.edu.au, sdey@unimelb.edu.au

978-1-4244-3872-3/09/$25.00 ©2009 IEEE

$\{X_k\}$ is fixed and known, whereas, the state values $\tilde{x}_i$ are unknown. The transition probability matrix $\mathbf{X}(\phi)$ and state space $\mathbb{X}(\phi)$ depend measurably on a parameter (vector) ϕ in a compact Euclidean space Φ. The "true" value of the parameter ϕ is denoted by $\phi^\circ \in \Phi$, and is assumed to be fixed but *unknown*. Note that for all $\phi \in \Phi$, we have $x_{ij}(\phi) \geq 0$ and $\sum_j x_{ij}(\phi) = 1$ for each i. The initial state probability vector of $\{X_k\}$ is denoted by $\pi = [\pi_i]$, where $\pi_i = \mathbb{P}(X_1 = \tilde{x}_i)$.

The Markov process $\{X_k\}$ is observed indirectly by noisy measurements $Y_{m,k} = X_k + w_{m,k},\ m = 1 \cdots M$, obtained from M sensors where M is fixed. Write $\mathbf{Y}_k = (Y_{1,k}, \cdots, Y_{M,k})'$ as the random vector of measurements obtained from the M number of sensors at time k. Also, let $\{\mathbf{Y}_k\}_{k=1}^{\infty}$ denote a vector of M random processes, with each process $\{Y_{m,k}\}_{k=1}^{\infty}$ being a sequence of conditionally independent random variables given a realization $\{x_k\}$ of $\{X_k\}$. Each random measurement $Y_{m,k}$ is characterized by a conditional density $f(.|x_k; \theta^m(\phi))$ with respect to the Lebesgue measure $\mho$ for $\theta^m : \Phi \mapsto \Theta$, where Θ is a Euclidean space. Write $\mathbf{w}_k = (w_{1,k}, \cdots, w_{M,k})'$ and let $\{\mathbf{w}_k\}_{k=1}^{\infty}$ be a vector of M independent noise processes, where each process $\{w_{m,k}\}_{k=1}^{\infty}$ is assumed to be an i.i.d. sequence of scalar real-valued innovations with known marginal distribution parameterized by a vector $\theta^m(\phi) \in \Theta$.

Due to severe bandwidth limitations in sensor networks, the measurements $\mathbf{Y}_k$ are then quantized according to a threshold-based binary quantization scheme, where the sequence $\{\mathbf{r}_k\}_{k=1}^{\infty} \stackrel{\triangle}{=} \{(r_{1,k}, \cdots, r_{M,k})\}_{k=1}^{\infty}$ denotes the sequence of quantization thresholds. Let $\mathbf{Y}_k^q = (Y_{1,k}^q, \cdots, Y_{M,k}^q)'$ represent the quantized data at time k, where $Y_{i,k}^q \in \{b_1, b_2\}$. The m-th sensor transmits its quantized output $Y_{m,k}^q$, with power level $p_{m,k}$ to a remote fusion centre over a discrete time flat fading channel. The transmission power for each sensor is chosen from a set $\mathbb{V}$ of finitely many discrete power levels, which is generally the case for most practical sensor systems. Let $\{\mathbf{p}_k\}_{k=1}^{\infty} \stackrel{\triangle}{=} \{(p_{1,k}, \cdots, p_{M,k})\}_{k=1}^{\infty}$ be the sequence of power levels and $\mathbf{Z}_k \stackrel{\triangle}{=} (Z_{1,k}, \cdots, Z_{M,k})'$ be the sensors' channel state vector at time k. We model each channel state process $\{Z_{m,k}\}_{k=1}^{\infty}$ as a stationary ergodic Markov chain[1] with state space $\mathbb{C} = \{\tilde{c}_1, \cdots, \tilde{c}_u\}$ and transition probability matrix $\mathbf{C}^m = [c_{ij}^m]$, where[2] $c_{ij}^m = \mathbb{P}(Z_{m,k} = \tilde{c}_j | Z_{m,k-1} = \tilde{c}_i)$, $1 \leq i, j \leq u$. Each channel state $\tilde{c}_i$ may represent a value of the channel gain. The initial state distribution of $\{Z_{m,k}\}$ is given by $\pi^m = [\pi_i^m]$, where $\pi_i^m = \mathbb{P}(Z_{m,1} = \tilde{c}_i)$.

Let $\mathbf{Y}_k^f = (Y_{1,k}^f, \cdots, Y_{M,k}^f)'$ be the vector of decoded binary symbols at the fusion centre, where $Y_{m,k}^f \in \{b_1, b_2\}.Y_{m,k}^f$ is described by the channel input-output transition probability $q_{ij}^m(\tilde{c}, \tilde{p}) \stackrel{\triangle}{=} \mathbb{P}(Y_{m,k}^f = b_j | Y_{m,k}^q = b_i, Z_{m,k} = \tilde{c}, p_{m,k} = \tilde{p})$, where $i, j \in \{1, 2\}$, $\tilde{c} \in \mathbb{C}$, $\tilde{p} \in \mathbb{V}$. The off-diagonal entries in the input-output transition matrix $\mathbf{Q}^m(\tilde{c}, \tilde{p}) \stackrel{\triangle}{=} [q_{ij}^m(\tilde{c}, \tilde{p})]$ are called crossover (error) probabilities. We assume that the sensors use a simple binary phase shift keying (BPSK) modulation scheme to transmit the binary quantized measurements over orthogonal additive white Gaussian noise (AWGN) channels[3]. The crossover probability can be computed[4] as $\varrho_k^m = Q(\sqrt{\gamma g_{m,k}^2 p_{m,k} \sigma_v^{-2} d_m^{-\varsigma}})$, where γ is a constant, $g_{m,k}$ is gain of the wireless channel, σ_v^2 is the variance of the Gaussian channel noise, d_m is the distance between the m-th sensor and the fusion centre, and ς is the path loss exponent of the wireless channel. For further details on computing the crossover probabilities, see [5].

We may now specify an HMM corresponding to the observation sequence $\{\mathbf{Y}_k^f\}_{k=1}^{\infty}$ by $\mathcal{H} = (\mathbf{X}(\phi), \mathbb{X}(\phi), \pi, \mathbf{\Psi}(\theta(\phi)))$, where $\theta(\phi) = (\theta^1(\phi), \cdots, \theta^M(\phi))'$, and $\mathbf{\Psi}$, the so-called state-to-observation probability[5] matrix, is a diagonal matrix with i-th diagonal entry $\psi_i(\mathbf{y}_k^f, \mathbf{z}_k, \mathbf{r}_k, \mathbf{p}_k; \theta(\phi))$, $i = 1, \cdots, n$ being the conditional probability mass function of $\mathbf{Y}_k^f$ with respect to the counting measure $\Im$ defined as $\mathbb{P}_\phi(\mathbf{Y}_k^f = \mathbf{y}_k^f | X_k = \tilde{x}_i, \mathbf{r}_k, \mathbf{Z}_k = \mathbf{z}_k, \mathbf{p}_k; \theta(\phi))$.

The task of the fusion centre is to find the optimal quantizer thresholds $\mathbf{r}_k$ and the optimal sensors transmit powers $\mathbf{p}_k$ while jointly estimating the state and the parameters of the underlying Markov chain $\{X_k\}$ with the objective being minimization of average state estimation error subject to an average sum power constraint across the sensors. If at each time instant k, the values of the optimal quantization thresholds $\mathbf{r}_k$ and optimal power levels $\mathbf{p}_k$ were exactly known, then estimating the parameters of $\mathcal{H}$ was rather a straightforward task and could have been done using existing estimation techniques[6]. On the other hand, if the parameters of $\mathcal{H}$ were exactly known, then optimal values of $\mathbf{r}_k$ and $\mathbf{p}_k$ could be determined using our state estimation algorithms presented in [5], or [6]. However, the crux of our present power-constrained quantization problem with unknown parameters is the lack of exact knowledge of the optimal $(\mathbf{r}_k, \mathbf{p}_k)$ to the parameter estimator and the lack of exact knowledge of the true parameter to the MDP controller.

Definition 2.1: Define the information state vector $\tilde{\alpha}_{k+1;\hat{\phi}^{(k)}}$ with i-th element $\tilde{\alpha}_{k+1;\hat{\phi}^{(k)}}(i)$, also known as normalized HMM filter density or normalized forward variable, being defined as $\mathbb{P}_\phi(X_{k+1} = \tilde{x}_i | \mathcal{D}_{k+1}, \mathcal{B}_{k+1}; \hat{\phi}^{(k)})$, where $\hat{\phi}^{(k)} = (\hat{\mathbf{X}}^{(k)}, \hat{\mathbb{X}}^{(k)}, \hat{\theta}^{(k)})$ denotes the sequence of estimates of model parameters up to time k, and $\mathcal{D}_k$, and $\mathcal{B}_k$ are the σ-fields generated by $(\mathbf{Y}_l^f, \mathbf{Z}_l)$, and $(\mathbf{r}_l, \mathbf{p}_l)$, for $l \leq k$, respectively. Also, define the filter state estimate as $\hat{X}_{k+1;\hat{\phi}^{(k)}} \stackrel{\triangle}{=} \mathbb{E}_\phi[X_{k+1} | \mathcal{D}_{k+1}, \mathcal{B}_{k+1}; \hat{\phi}^{(k)}]$.

[1] Note that finite-state Markov chain models have often been used in information theory literature to characterize wireless channels. The channel is typically modeled by appropriately partitioning the range of the received signal-to-noise ratio (SNR) into a set of intervals (states) using different partitioning criteria, see, e.g., [7] and references therein.

[2] To simplify our subsequent analysis, we assume that $c_{ij}^m > 0$ for all $1 \leq i, j \leq u$ and $1 \leq m \leq M$.

[3] Note that under certain standard symmetry assumptions on the modulation scheme and noise, the channel input-output transition probability matrix becomes a symmetric matrix (the channel is called a binary symmetric channel (BSC)), which is the case assumed in our analysis for simplicity.

[4] $Q(.)$ is the complementary standard normal cdf function.

[5] see [5] for details on deriving the state-to-observation probabilities.

[6] e.g., off-line schemes such as Baum-Welch algorithm or on-line such as [8].

The state space of the forward variable $\tilde{\alpha}_{k+1;\hat{\phi}^{(k)}}$ can be defined as the simplex $\mathbb{T}_{\tilde{\alpha}} = \{\tilde{\alpha} \in \mathbb{R}^n_+ \mid \langle \tilde{\alpha}, \mathbb{1}_n \rangle = 1\} \subset \mathbb{R}^n$, where $\mathbb{1}_n$ is the n-dimensional vector with all elements equal to one. For numerical tractability, we approximate the continuum information state $\tilde{\alpha}_{k+1;\hat{\phi}^{(k)}}$ by discretized vector $\alpha_{k+1;\hat{\phi}^{(k)}} \in \mathbb{T}$, where $\mathbb{T} \subset \mathbb{T}_{\tilde{\alpha}}$ is the state space of the discretized forward variable $\alpha_{k+1;\hat{\phi}^{(k)}}$. Note that after discretization, it is ensured that $\langle \alpha, \mathbb{1}_n \rangle = 1, \ \forall \alpha \in \mathbb{T}$ (see [3] for further details). Let $\mathbb{U}$ denote the range space of each variable $r_{m,k}$ which is, in general, equal to $\mathbb{R}$ for $\forall m \in \{1 \cdots M\}$[7].

III. The MDP Controller

A. Notation and Preliminaries

Let $\mathbb{S}$ and $\mathbb{A}$ be Borel subsets of complete separable metric spaces endowed with the Borel σ-algebra $\mathcal{S}$ and $\mathcal{A}$ respectively. The cartesian product $\mathbb{K} \triangleq \mathbb{S} \times \mathbb{A}$ is endowed with the corresponding product metric topology and the product σ-algebra $\mathcal{S} \times \mathcal{A}$. Let $(\Omega, \mathcal{F})$ be the measurable space consisting of a sample space Ω as the product space $\{\mathbb{K}\}^\infty$ and the corresponding product σ-field $\mathcal{F}$ of subsets of Ω. Let $\{\mathbb{Q}(\mathbb{S}|\mathbb{K};\phi) : \phi \in \Phi\}$ denote the space of all stochastic kernels on $\mathbb{S}$ given $\mathbb{K}$, indexed by a parameter (vector) $\phi \in \Phi$, i.e., for a given parameter value $\phi \in \Phi$, a conditional probability measure $p \in \mathbb{Q}(\mathbb{S}|\mathbb{K};\phi)$ is a function such that (i) for any Borel set $S \in \mathcal{S}$, the mapping $p(S|.;\phi) : \mathbb{K} \mapsto [0,1]$ is a Borel measurable function on $\mathbb{S}$; and (ii) for each $\kappa \in \mathbb{K}$, $p(.|\kappa;\phi)$ is a probability measure on the Borel σ-algebra of $\mathbb{S}$. Furthermore, let $\{\mathbb{G}(.;\phi) : \phi \in \Phi\}$ be the space of all continuous measurable functions, where for any $g \in \mathbb{G}$, $\{g(.;\phi) : \mathbb{K} \mapsto \mathbb{R}, \ \phi \in \Phi\}$ is a family of real-valued functions indexed by a parameter ϕ.

Definition 3.1: Define $(\mathbb{S}, \mathbb{A}, g(.;\phi), p(.|.;\phi))$ to be an *adaptive MDP* depending on an unknown parameter $\phi \in \Phi$, where $\mathbb{S}$ and $\mathbb{A}$ are called state and action spaces, respectively. $\mathbb{S}$ is assumed to be a nonempty countable (possibly infinite) set endowed with the discrete topology[8]. The action space $\mathbb{A}$ is assumed to be a nonempty Borel space. Further, $g(.;\phi) \in \mathbb{G}$, the so-called immediate (or per-stage) cost function, is a measurable function on $\mathbb{K} \times \Phi$ and $p(.|.;\phi) \in \mathbb{Q}$ is the transition law of the MDP. $\mathbb{K}$ is the set of admissible state-action pairs defined as $\mathbb{K} \triangleq \{(s,a) | s \in \mathbb{S}, a \in \mathbb{A}(s)\}$ which is a topological subspace of $\mathbb{S} \times \mathbb{A}$. Henceforth, for simplicity, we may use shorter notations $g(\phi)$ and $p(\phi)$.

B. Adaptive Markov Decision Process Model

In this section, we formulate our quantization problem as an adaptive infinite-horizon average cost MDP problem. Let $\mathcal{M}_\phi = (\mathbb{S}, \mathbb{A}, g(\phi), p(\phi))$ be an adaptive MDP depending on a parameter vector $\phi \in \Phi$ as defined in Definition 3.1, where $\mathbb{S} = \mathbb{T} \times \mathbb{C}^M$, $\mathbb{A} = \mathbb{U}^M \times \mathbb{V}^M$ are corresponding state and action spaces, respectively. The immediate cost function $g(\phi)$ is defined by $g(\alpha_{k;\phi}, \mathbf{z}_k, \mathbf{r}_k, \mathbf{p}_k; \phi) \triangleq \mathcal{E}(\alpha_{k;\phi}) + \beta \mathcal{Q}(\mathbf{p}_k)$, which is a weighted combination of two cost functions: conditional expected state estimation error $\mathcal{E}(\alpha_{k;\phi}) \triangleq \mathbb{E}_\phi[\ |X_k - \hat{X}_{k;\phi}|^2 \ \big|\mathcal{D}_k, \mathcal{B}_k; \phi] = \sum_{i=1}^n [\tilde{x}_i(\phi) - \sum_{j=1}^n \tilde{x}_j(\phi) \alpha_{k;\phi}(j)]^2 \alpha_{k;\phi}(i)$, and total power consumption across the sensors $\mathcal{Q}(\mathbf{p}_k) \triangleq \sum_{m=1}^M p_{m,k}, \ \ p_{m,k} \in \mathbb{V}$. The weighting factor $\beta \geq 0$ is a trade-off parameter which assumes the role of a Lagrangian multiplier in constrained optimization.

For a given parameter value $\phi \in \Phi$, if the MDP $\mathcal{M}_\phi$ is in state $s = (\alpha, \mathbf{z}) \in \mathbb{S}$ and action $a = (\mathbf{r}, \mathbf{p}) \in \mathbb{A}(s)$ is taken, then the observation $\mathbf{y}^f$ will be received at the fusion centre and the MDP state changes to $\acute{s} = (\acute{\alpha}, \acute{\mathbf{z}}) \in \mathbb{S}$ according to the transition probability distribution $p(\acute{s}|s, a; \phi)$ which is computed by $p_z(\acute{\mathbf{z}} \mid \mathbf{z}) \ \big\langle \mathbf{\Psi}(\mathbf{y}^f, \acute{\mathbf{z}}, \mathbf{r}, \mathbf{p}; \theta(\phi)) \mathbf{X}'(\phi) \alpha, \ \mathbb{1}_n \big\rangle$, where for $\mathbf{z} = (\tilde{c}_{i_1}, \cdots, \tilde{c}_{i_M})'$ and $\acute{\mathbf{z}} = (\tilde{c}_{j_1}, \cdots, \tilde{c}_{j_M})'$, $p_z(\acute{\mathbf{z}} \mid \mathbf{z})$ is the product of M channel transition probabilities computed by $\prod_{m=1}^M c^m_{i_m j_m}$ for $i_m, j_m \in \{1, \cdots, u\}$. It is straightforward to show that the value of the forward variable $\acute{\alpha}$ in the next MDP state $\acute{s}$ is obtained by recursion $\big[\langle \bar{\alpha}, \ \mathbb{1}_n \rangle^{-1} \ \bar{\alpha}\big]_{round}$, where $[.]_{round} : \mathbb{T}_{\tilde{\alpha}} \mapsto \mathbb{T}$ is the discretization operator for the information state as described in [3], and $\bar{\alpha}$ is the unnormalized forward variable with respect to the Lebesgue measure $\mho$ computed by $\mathbf{\Psi}(\mathbf{y}^f, \acute{\mathbf{z}}, \mathbf{r}, \mathbf{p}; \theta(\phi)) \ \mathbf{X}'(\phi) \alpha$.

For each (fixed) value of $\phi \in \Phi$, we specify an objective function $J^\lambda_\phi(\mathring{s})$, expressed as the long-term average expected cost per time step, or simply the average cost defined by

$$J^\lambda_\phi(\mathring{s}) \triangleq \limsup_{N \to \infty}$$

$$N^{-1} \sum_{k=1}^N \mathbb{E}^\lambda_{\mathring{s},\phi} \Big[g(\alpha_{k;\phi}, \mathbf{z}_k, \mathbf{r}_k, \mathbf{p}_k; \phi) \big| \alpha_1 = \mathring{\alpha}, \mathbf{Z}_0 = \mathring{\mathbf{z}} \Big] \quad (1)$$

where $\mathring{s} = (\mathring{\alpha}, \mathring{\mathbf{z}})$ is the initial condition, $\lambda = (\mathbf{r}, \mathbf{p}) \in \mathbf{\Lambda}$ is a policy, and $\mathbf{\Lambda}$ is the non-empty space of all admissible randomized history-dependent policies. For each fixed value of $\phi \in \Phi$, $\mathbb{E}^\lambda_{\mathring{s},\phi}$ denotes the expectation with respect to $\mathbb{P}^\lambda_{\mathring{s},\phi}$ which is the unique probability measure on $(\Omega, \mathcal{F})$ for a given policy $\lambda \in \mathbf{\Lambda}$, and initial state $\mathring{s} \in \mathbb{S}$. The function $J^\lambda_\phi(\mathring{s})$ is a performance metric for our quantization problem measuring the performance when a given policy λ is used and the system starts with the initial condition $\mathring{s}$.

Our quantization problem may then be expressed as an adaptive stochastic control problem defined as following. Determine an average-optimal policy $\lambda^*_{\phi^\circ}$ and its corresponding average optimal cost $J^*_{\phi^\circ}$ as the solution to the following optimization problem

$$(\mathbf{P}) : \quad \inf_{\lambda \in \mathbf{\Lambda}} \ J^\lambda_\phi(s), \ \text{ for } \forall s \in \mathbb{S},$$

where J^λ_ϕ is the function defined in (1). It is clear that if we had the exact knowledge of the true parameter ϕ°, then the solution to the problem $(\mathbf{P})$ would reduce to finding optimal policies to the average cost MDP problem defined by $J^*_{\phi^\circ}(s) \triangleq \inf_{\lambda \in \mathbf{\Lambda}} \ J^\lambda_{\phi^\circ}(s)$, for $\forall s \in \mathbb{S}$, which is the problem associated with $\mathcal{M}_{\phi^\circ}$ and has been studied earlier in [5]. However, since the true parameter is unknown, the

[7] Note that the theory in Section III holds, in general, for $\mathbb{U} = \mathbb{R}$ with the usual topology. However, in further analysis, to simplify the implementation of our value iteration algorithm, we restrict the action space $\mathbb{U}$ to a finite set of discrete values in $\mathbb{R}$.

[8] that is, the topology consisting of all open subsets of $\mathbb{S}$.

idea to approach this problem is to compute a sequence of estimates $\{\hat{\phi}_k\}_{k=1}^{\infty}$ of the true parameter ϕ°, and show that if $\hat{\phi}_k$ converges to ϕ° $\mathbb{P}_{s,\phi^\circ}^{\lambda}$–a.s. as $k \to \infty$ then policies (suitably[9]) adapted to the *approximating MDP sequence* $\mathcal{M}_{\hat{\phi}_k} = (\mathbb{S}, \mathbb{A}, g(\hat{\phi}_k), p(\hat{\phi}_k))$ are average-optimal for the true MDP $\mathcal{M}_{\phi^\circ}$ [9]. We discuss this approach in the following section. We introduce the following assumptions.

Assumptions 3.1: For the adaptive MDP $\mathcal{M}_\phi$ the following hold:

A1) Each state $s \in \mathbb{S}$ is associated with a nonempty measurable compact set $\mathbb{A}(s) \subseteq \mathbb{A}$ of admissible actions when the MDP is in state s.

A2) The immediate cost $g(s, a; \phi)$ is a continuous function of $a \in \mathbb{A}(s)$ for $\forall s \in \mathbb{S}$ uniformly in $\phi \in \Phi$.

A3) For some constant G, the immediate cost function g satisfies $|g(s, a; \phi)| \leq G < \infty$ uniformly in ϕ.

C. ϕ-Optimality Equation

It has been shown (cf. Theorem 5.5.3 [10]) that the search domain for optimal policies in the stochastic control problem $(\mathbf{P})$ may be restricted only to the space of Markov policies instead of the general domain $\mathbf{\Lambda}$ of randomized history-dependent policies. Let $\mathbb{F}$ denote the space of all deterministic Markov decision rules defined as measurable functions $\lambda_\phi(s) : \mathbb{S}\times\Phi \mapsto \mathbb{A}$ such that $\lambda_\phi(s) \in \mathbb{A}(s)$ for every $s \in \mathbb{S}$ and $\phi \in \Phi$. Assume that for each $\phi \in \Phi$, the action $a_k = (\mathbf{r}_k, \mathbf{p}_k)$ at each time step k is determined by a stationary deterministic Markov policy $\lambda_\phi^\infty = \{\lambda_\phi, \lambda_\phi, \cdots\}$, where $a_k = \lambda_\phi(s_k)$. Henceforth, for brevity, λ_ϕ^∞ may be denoted by λ_ϕ. We introduce the following assumption.

Assumption 3.2: For any $a = (\mathbf{r}, \mathbf{p}) \in \mathbb{A}$, $\mathbf{y}^f \in \{b_1, b_2\}^M$, $\mathbf{z} \in \mathbb{C}^M$, the matrix $\mathbf{\Psi}(\mathbf{y}^f, \mathbf{z}, \mathbf{r}, \mathbf{p}; \theta(\phi))\mathbf{X}'(\phi)$ is primitive and non-singular uniformly in ϕ.

Remark 3.1: Notice that Assumption 3.2 holds under very mild conditions for the noise process $\{\mathbf{w}_k\}_{k=1}^{\infty}$. For further details see [6].

For each $\phi \in \Phi$, the ϕ-optimality equations (ϕ-OEs), also known as the Bellman equations[10], associated with the adaptive average cost MDP problem $(\mathbf{P})$ may be expressed as

$$\min_{\lambda_\phi \in \mathbb{F}} \Big[\sum_{\mathbf{y}^f, \acute{\mathbf{z}}} J_\phi^{\lambda_\phi}(\acute{s}) p(\acute{s}|s, a; \phi) - J_\phi^{\lambda_\phi}(s) \Big] = 0 \tag{2}$$

$$J_\phi^{\lambda_\phi}(s) + v(s; \phi) = \min_{\lambda_\phi \in \mathbb{F}} \Big[g(s, a; \phi) + \sum_{\mathbf{y}^f, \acute{\mathbf{z}}} v(\acute{s}; \phi) p(\acute{s}|s, a; \phi) \Big] \tag{3}$$

where $a = \lambda_\phi(s) \in \mathbb{A}(s)$, $\acute{s} = (\acute{\alpha}(\mathbf{y}^f), \acute{\mathbf{z}})$, $J_\phi^{\lambda_\phi}$ is the average per-stage cost in steady state which we are after its optimal value J_ϕ^*. The term $\sum_{\mathbf{y}^f, \acute{\mathbf{z}}} v(\acute{s}; \phi)\ p(\acute{s}|s, a; \phi)$ is called cost-to-go function in which $v \in B(\mathbb{S} \times \Phi)$ with $B(.)$ being the Banach space of real-valued bounded measurable functions v with the uniform norm $\parallel v \parallel \overset{\triangle}{=} \sup_{s\in\mathbb{S}, \phi\in\Phi} \mid v(s; \phi) \mid$. The function $v(.; \phi)$, referred to as the differential cost, is defined as the expected total difference between per-stage cost $g(.; \phi)$ and the stationary cost $J_\phi^{\lambda_\phi}$. We introduce the following assumption:

Assumption 3.3: The cost-to-go function is a continuous function of $a = (\mathbf{r}, \mathbf{p}) \in \mathbb{A}(s)$ for every $s = (\alpha, \mathbf{z}) \in \mathbb{S}$, $\phi \in \Phi$, and $v \in B(\mathbb{S} \times \Phi)$.

In order to establish the existence of solutions to the ϕ-OEs for every $\phi \in \Phi$, we need to provide the following ergodicity condition.

Lemma 3.1: Suppose that Assumption 3.2 holds. Then, there exists an MDP state $\bar{s} \in \mathbb{S}$ and a positive number e_0 such that the inequality $p(\bar{s}|s, a; \phi) \geq e_0$ holds uniformly in ϕ for $\forall s \in \mathbb{S}$, and $\forall a \in \mathbb{A}$.

Proof: The proof is straightforward and for brevity the detail is omitted here. ■

Using Lemma 3.1, we establish the following ergodicity result.

Lemma 3.2: For any stationary deterministic Markov policy $\lambda_\phi \in \mathbb{F}$, there exists a unique invariant probability measure Γ_{λ_ϕ} of the Markov state process $\{s_k\}_{k=1}^{\infty}$ induced by the stationary policy λ_ϕ. Γ_{λ_ϕ} is defined as the unique probability measure on $\mathbb{S}$ which for $a = \lambda_\phi(s) \in \mathbb{A}(s)$ satisfies $\Gamma_{\lambda_\phi}(j) = \sum_{s\in\mathbb{S}} p(j|s, a; \phi)\Gamma_{\lambda_\phi}(s),\ \forall j \in \mathbb{S}$.

Proof: The argument may be verified by the implications in part (a) of Lemma 3.3 in [9]. ■

From Lemma 3.2, for a given $\phi \in \Phi$, Markov chains induced by every stationary policy $\lambda_\phi \in \mathbb{F}$ have a single irreducible class. In fact, under Assumption 3.2, the average cost MDP problem $(\mathbf{P})$ associated with $\mathcal{M}_\phi$ forms a recurrent (ergodic) MDP. This means that the transition matrix corresponding to every deterministic stationary policy consists of a single recurrent class and no transient state. This is because using Lemma 3.2 it can be shown[11] that the cost associated with every deterministic stationary Markov policy λ_ϕ is uniform in s, that is, $J_\phi^{\lambda_\phi}(s) = j_\phi^{\lambda_\phi} \in B(\Phi)$ for $\forall s \in \mathbb{S}$.

Since, under Assumption 3.2, the structure of every Markov chain induced by a deterministic stationary policy is classified as a recurrent Markov chain, we may characterize optimal policies and their corresponding average costs using only the single ϕ-optimality equation (3). The reason for this is that the ϕ-optimality equation (2) holds for every $J_\phi^{\lambda_\phi}(s)$ which is uniform in s and thus it always holds for every $\lambda_\phi \in \mathbb{F}$ and as such is uninformative. Therefore, for each $\phi \in \Phi$, we may characterize the optimal policy and the associated optimal cost only through a single ϕ-OE as following:

$$j_\phi^{\lambda_\phi} + v(s; \phi) = \min_{\lambda_\phi \in \mathbb{F}} \Big[g(s, a; \phi) + \sum_{\mathbf{y}^f, \acute{\mathbf{z}}} v(\acute{s}; \phi) p(\acute{s}|s, a; \phi) \Big] \tag{4}$$

Now, we establish existence of solutions to the ϕ-optimality equation (4) which is used in the following theorem. Under Assumptions 3.1 and 3.3 and using the ergodicity result in

[9] a suitably adapted policy refers to an NVI adaptive policy to be introduced in the next section.

[10] cf. [10, Section 8.4.]

[11] cf. part (b) of Lemma 3.3 in [9].

Lemma 3.1, it can be shown[12] that there exists a solution $\{j^*_\phi, v^*(.;\phi)\}$ to the ϕ-optimality equation (4), where $j^*_\phi \in B(\Phi)$ is a real-valued bounded measurable function on Φ and $v^*(.;\phi) : \mathbb{S} \times \Phi \mapsto \mathbb{R}$ is a real-valued bounded measurable function on $\mathbb{S}$ for each $\phi \in \Phi$.

Theorem 3.3: Suppose that Assumptions 3.1 and 3.3 hold, and there exist functions $j^*_\phi \in B(\Phi)$, and $v^*(.;\phi) \in B(\mathbb{S} \times \Phi)$ in the Banach spaces of real-valued bounded measurable functions which satisfy the ϕ-OE (4). Assume there is a stationary deterministic Markov decision rule $\lambda^*_\phi \in \mathbb{F}$ which minimizes the right-hand side of the ϕ-OE (4), i.e., for each $\phi \in \Phi$ and $s \in \mathbb{S}$

$$j^*_\phi + v^*(s;\phi) = g(s, \lambda^*_\phi(s);\phi) + \sum_{\mathbf{y}^f, \acute{\mathbf{z}}} v^*(\acute{s};\phi) p(\acute{s}|s, \lambda^*_\phi(s);\phi),$$

where $\lambda^*_\phi(s) \in \mathbb{A}(s)$. Then, the stationary policy λ^*_ϕ is average-optimal for the MDP $\mathcal{M}_\phi$, that is, action $a_k = (\mathbf{r}_k, \mathbf{p}_k) = \lambda^*_\phi(s_k)$ at time $k \geq 2$ determined by the stationary policy $\lambda^*_\phi = \{\lambda^*_\phi, \lambda^*_\phi, \cdots\}$ minimizes the cost J_ϕ defined in (1) and the value of the optimal cost is j^*_ϕ.

Remark 3.2: This theorem is essentially the ϕ-analogue (parameterized version) of the existence theorem[13] for optimal policies in average cost unichain models.

D. Nonstationary Value Iteration

In this section, we develop the formulation for approximating MDP models $\mathcal{M}_{\hat{\phi}_k}$ and introduce a nonstationary value iteration (NVI) scheme and corresponding NVI adaptive Markov policies and show that under appropriate assumptions these adaptive policies are average-optimal for the limit (true) MDP $\mathcal{M}_{\phi^\circ}$. The approach followed in this section is inspired by results on approximations and adaptive policies for average cost MDPs presented in [9].

Let $\{\hat{\phi}_k\}_{k=1}^\infty$ be a sequence in Φ converging to the true parameter ϕ° according to the following definition, where H_k refers to the vector space of admissible histories up to time k for $k \geq 0$, where $H_0 \triangleq \mathbb{S}$ and $H_k \triangleq \mathbb{K}^k \times \mathbb{S} = \mathbb{K} \times H_{k-1}$ for $k \geq 1$.

Definition 3.2: A sequence $\{\hat{\phi}_k\}_{k=1}^\infty$ of measurable functions $\hat{\phi}_k : H_k \mapsto \Phi$ is defined to be a sequence of strongly consistent estimators of the true parameter ϕ° such that $\lim_{k\to\infty} \hat{\phi}_k = \phi^\circ$ $\mathbb{P}^\lambda_{s,\phi^\circ}$–a.s. is satisfied uniformly in λ for every $s \in \mathbb{S}$

Remark 3.3: Note that there are several methods to estimate parameters of the HMM $\mathcal{H}$ in sense of the Definition 3.2. However, at this point, to maintain readability of the manuscript, it is simply assumed that the strongly consistent estimator $\{\hat{\phi}_k\}_{k=1}^\infty$ is available. This task is performed by the recursive maximum likelihood (ML) parameter estimator module which is discussed in Section IV.

Let $\mathcal{M}_{\hat{\phi}_k} = (\mathbb{S}, \mathbb{A}, g(\hat{\phi}_k), p(\hat{\phi}_k))$, for $k \in \{0, 1, \cdots\}$, be a sequence of MDPs, called the *approximating MDP sequence*, and $\hat{\phi}_k$ be a sequence of estimates in Φ converging to the true parameter ϕ° according to the Definition 3.2. It is clear that each of the approximating MDPs $\mathcal{M}_{\hat{\phi}_k}$ satisfies Assumptions 3.1, 3.2, and 3.3. Therefore, the ergodicity result in Lemma 3.2 holds for each $\mathcal{M}_{\hat{\phi}_k}$. The approximating MDP sequence $\{\mathcal{M}_{\hat{\phi}_k}\}_{k=0}^\infty$ converges to the true MDP $\mathcal{M}_{\phi^\circ}$ in the following sense:

Convergence criterion:
For any $\phi^\circ \in \Phi$ and any sequence of parameter estimates $\{\hat{\phi}_k\}_{k=1}^\infty$ in Φ such that $\hat{\phi}_k$ converges to ϕ° $\mathbb{P}^\lambda_{s,\phi^\circ}$–a.s. as $k \to \infty$, the following sequences

$$\zeta_k(\phi^\circ) \triangleq \sup_{s\in\mathbb{S}, a\in\mathbb{A}(s)} |g(s,a;\hat{\phi}_k) - g(s,a;\phi^\circ)|$$

$$\rho_k(\phi^\circ) \triangleq \sup_{s\in\mathbb{S}, a\in\mathbb{A}(s)} \parallel p(.|s,a;\hat{\phi}_k) - p(.|s,a;\phi^\circ) \parallel_{tv}$$

satisfy $\lim_{k\to\infty} \zeta_k(\phi^\circ) = 0$ and $\lim_{k\to\infty} \rho_k(\phi^\circ) = 0$, where $\parallel . \parallel_{tv}$ denotes the total variation norm for finite signed measures.

Remark 3.4: The above convergence criterion is in fact a continuity condition of the per-stage cost function $g(\phi)$ and the MDP transition kernel $p(\phi)$ in the parameter $\phi \in \Phi$ uniformly on $\mathbb{K}$. This criterion is basically used in order to prove Lipschitz continuity of the NVI function[14] $\bar{v}_k(.;\phi)$, which is introduced in Definition 3.3, in ϕ uniformly on $\mathbb{S}$. However, it is straightforward to show[15] that instead it is sufficient if $g(\phi)$ and $p(\phi)$ satisfy a regularity condition (Lipschitz continuity) in ϕ uniformly on $\mathbb{K}$. This is stated in the following assumption.

Assumption 3.4: There are constants L_1 and L_2 such that the following inequalities is satisfied uniformly in $\kappa = (s,a) \in \mathbb{K}$ for every $\phi, \acute{\phi} \in \Phi$.

$$\begin{aligned} |g(\kappa;\phi) - g(\kappa;\acute{\phi})| &\leq L_1 \bar{d}(\phi,\acute{\phi}), \\ \parallel p(.|\kappa;\phi) - p(.|\kappa;\acute{\phi}) \parallel_{tv} &\leq L_2 \bar{d}(\phi,\acute{\phi}), \end{aligned}$$

where $\bar{d}$ is the metric on the parameter space Φ.

Remark 3.5: Using the Mean Value Theorem, it is trivial to show that Assumption 3.4 holds for the functions $g(\phi)$ and $p(\phi)$ due to the fact that both functions are differentiable a.e., and their first derivatives are upper-bounded $\mathbb{P}_\phi$–a.e., that is, bounded by an essential upperbound.

We define nonstationary value iteration (NVI) functions $\bar{v}_k(s;\hat{\phi}_{k-1})$ recursively as follows.

Definition 3.3: Let $\bar{v}_0(.)$ be an arbitrary function defined in $B(\mathbb{S} \times \Phi)$, e.g., $\bar{v}_0(.) \triangleq 0$, and for every $s \in \mathbb{S}$, and $k \geq 0$

$$\begin{aligned} \bar{v}_{k+1}(s;\hat{\phi}_k) \triangleq \min_{a\in\mathbb{A}(s)} \Big[& g(s,a;\hat{\phi}_k) + \\ & \sum_{\mathbf{y}^f, \acute{\mathbf{z}}} \bar{v}_k(\acute{s};\hat{\phi}_{k-1}) p(\acute{s}|s,a;\hat{\phi}_k) \Big] \end{aligned} \quad (5)$$

It is clear from Definition 3.3 that the NVI functions $\bar{v}_k(s;\hat{\phi}_{k-1})$ are obtained in an iterative manner starting from an arbitrary initial function $\bar{v}_0(.)$ without the need to solve

[12] cf. [9, Corollary 3.6].

[13] cf. [10, Theorem 8.4.4].

[14] cf. [9, Proposition 5.6].

[15] the proof is similar to that of [9, Theorem 4.8]. For brevity, the detail is omitted here.

the Bellman ϕ-optimality equation (4) at each time step k, which in our case is computationally intensive. This advantage makes the NVI scheme directly applicable to our problem. The NVI adaptive policy corresponding to the NVI functions $\bar{v}_k(.;\hat{\phi}_{k-1})$ is defined as following.

Definition 3.4: Let $\bar{\lambda} = \{\bar{\lambda}_k\}_{k=0}^{\infty}$, called the NVI adaptive policy, be a sequence of deterministic Markov decision rules, where for each $k \geq 0$, $\bar{\lambda}_k(.;\hat{\phi}_k) \in \mathbb{F}$ is a measurable function such that the action a obtained by $a = \bar{\lambda}_k(s;\hat{\phi}_k) \in \mathbb{A}(s)$ minimizes the right hand side of the ϕ-optimality equation (5) for every $s \in \mathbb{S}$. It is clear that the initial action at time $k=0$ is determined by $\bar{\lambda}_0(s;\hat{\phi}_0) = \arg\ \min_{a\in\mathbb{A}(s)} g(s,a;\hat{\phi}_0)$.

The following theorem establishes the average optimality of the NVI adaptive policy $\bar{\lambda}$ for the true MDP $\mathcal{M}_{\phi^\circ}$.

Theorem 3.4: Suppose that Assumptions 3.1–3.4 hold. Let $\{\hat{\phi}_k\}_{k=0}^{\infty}$ be any sequence of measurable functions in Φ converging to the true parameter ϕ° $\mathbb{P}_{s,\phi^\circ}^{\lambda}$–a.s. . Also, let $\bar{\lambda}_{\phi^\circ} = \{\bar{\lambda}_k\}_{k=0}^{\infty}$ be an adaptive policy as defined in Definition 3.4, where $\bar{\lambda}_k(s_k;\hat{\phi}_k(h_k)) \in \mathbb{F}$ for every $h_k \in H_k$. Then $\bar{\lambda}_{\phi^\circ}$ is an average-optimal policy for the true MDP $\mathcal{M}_{\phi^\circ}$.

Remark 3.6: Theorem 3.4 is the ϕ-analogue of [9, Theorem 6.6] for average cost MDPs. The proof is similar to that of [9, Corollary 7.8, pp. 80] and the detail is omitted here.

IV. Recursive ML Parameter Estimator

In this section, we develop the formulation for a recursive (on-line) expectation maximization (EM) algorithm to estimate the parameters of the HMM $\mathcal{H} = (\mathbf{X}(\phi), \mathbb{X}(\phi), \pi, \mathbf{\Psi}(\theta(\phi)))$. The proposed method is an adaptation of the on-line estimation algorithm based on relative entropy information measure presented in [8]. As such, we only present the variations necessary to our problem and all further details are omitted.

Define $\hat{\phi}_k \triangleq (\mathbf{X}(\hat{\phi}_k), \mathbb{X}(\hat{\phi}_k), \hat{\theta}_k(\hat{\phi}_k)) \in \Phi$ as the estimate of model parameters at time $k \geq 0$. Let $\mathbf{O}_k^K \in \mathcal{O}_k^K$ denote the observable (incomplete) data at the fusion centre from time instant k up to time K, where $\mathcal{O}_k^K$ is the σ-field generated by $(\mathbf{Y}_l^f, \mathbf{Z}_l, \mathbf{r}_l, \mathbf{p}_l)$, for $k \leq l \leq K$. For simplicity, we denote $\mathbf{O}_1^k$ by $\mathbf{O}^k$, and $\mathbf{O}_k^k$ by $\mathbf{O}_k$. Henceforth, using this notation, for brevity the state-to-observation probability distribution $\psi_i(\mathbf{y}_k^f, \mathbf{z}_k, \mathbf{r}_k, \mathbf{p}_k; \theta(\phi))$, $i = 1, \cdots, n$ may be denoted by $\psi_i(\mathbf{O}_k;\theta(\phi))$. Let $\varphi^k = \{\hat{\phi}_t\}_{t=0}^k$ denote the sequence of model estimates till time k based on the observations $\mathbf{O}^k$. Also, denote the sequence of unobservable Markov chain states till time k by $X^k = \{X_t\}_{t=1}^k$. In the following, $l_c(.)$ denotes a probability measure on $(\Omega, \mathcal{F})$ with respect to some σ-finite measure. It is shown in [8] that the M-step of the on-line EM algorithm maximizes the relative entropy information measure which is equivalent to maximizing $\mathcal{J}(\phi) \triangleq \mathbb{E}_{\phi^\circ}\big[\log l_c(\mathbf{O}^k;\phi)\big]$, where $l_c(\mathbf{O}^k;\phi)$ is the marginal likelihood function of the observable (incomplete) data parameterized by ϕ. The M-step may be expressed as following:

$$\hat{\phi}_k \ = \arg\ \max_{\phi\in\Phi}\ \ \bar{Q}_k(\mathbf{O}^k, \varphi^{k-1};\phi), \quad k \geq 1, \qquad (6)$$

$$\text{subject to:}\ \ \sum_{j=1}^{n} x_{ij}(\phi) = 1, \quad \forall i = 1, \cdots, n$$

$$x_{ij}(\phi) \geq 0, \quad \forall i,j = 1, \cdots, n$$

where $\bar{Q}_k(.)$ for $k \geq 1$ is computed in the E-step as following:
E**-step:**

$$\bar{Q}_k(\mathbf{O}^k, \varphi^{k-1};\phi) \triangleq \mathbb{E}_{X^k}\Big[\log l_c(\mathbf{O}^k, X^k;\phi)|\mathbf{O}^k, \varphi^{k-1}\Big],$$

where $l_c(\mathbf{O}^k, X^k;\phi)$ is the likelihood function of the complete data if X^k were fully observable.

Remark 4.1: Note that in the optimization problem (6), depending on the distribution of the sensors' observations further constraints on the elements of the parameter vector ϕ might be required. As an example, for zero-mean Gaussian measurement noise processes $\{\mathbf{w}_k\}_{k=1}^{\infty}$, the standard deviation parameter $\theta^m(\phi)$ in the conditional density $f(.|x_k;\theta^m(\phi))$ must be strictly positive[16].

Lemma 4.1: From Theorem 3.4 suppose at each time step $k \geq 1$, action $(\mathbf{r}_k, \mathbf{p}_k)$ is determined according to a *deterministic* NVI adaptive policy $\bar{\lambda}_k(.;\hat{\phi}_k) \in \mathbb{F}$. Further assume that the trajectory $\mathbf{O}^{k+1} \in \mathcal{O}^{k+1}$ has been observed. Then, for $k \geq 0$, the function $\bar{Q}_{k+1}(.)$ may be evaluated as following:

$$\bar{Q}_{k+1}(\mathbf{O}^{k+1}, \varphi^k;\phi) = \sum_{t=1}^{k+1} \chi_{t|k+1}(\phi) + \sum_{t=1}^{k+1} \bar{\chi}_{t|k+1}(\phi, \mathbf{O}_t)$$

$$+ \sum_{t=1}^{k} \log p_z(\mathbf{z}_{t+1} \mid \mathbf{z}_t) + \sum_{m=1}^{M}\sum_{i=1}^{u} \bar{\delta}(z_{m,1} - \tilde{c}_i) \log \pi_i^m, \quad (7)$$

where $\bar{\delta}(.)$ is the Kronecker delta function, and the functions $\chi_{t|k+1}(.)$ and $\bar{\chi}_{t|k+1}(.)$ are evaluated as following

$$\chi_{t|k+1}(\phi) = \sum_{i=1}^{n}\sum_{j=1}^{n} \bar{\xi}_{t-1|k+1}(i,j) \log s_{ij}^2(\phi)$$

$$\bar{\chi}_{t|k+1}(\phi, \mathbf{O}_t) = \sum_{i=1}^{n} \bar{\gamma}_{t|k+1}(i) \log \psi_i(\mathbf{O}_t;\vartheta(\phi)), \qquad (8)$$

where $\bar{\xi}_{t|k}(i,j) = \mathbb{P}_\phi(X_t = \tilde{x}_i, X_{t+1} = \tilde{x}_j|\mathbf{O}^k, \varphi^{k-1})$ and $\bar{\gamma}_{t|k}(i) = \mathbb{P}_\phi(X_t = \tilde{x}_i|\mathbf{O}^k, \varphi^{k-1})$ are the standard variables used to implement the forward-backward procedure in order to evaluate $\bar{Q}_k(.)$ in the EM algorithm. These variables can be computed recursively using standard recursions which can be found in [8], and are not presented here due to space limitations.

Remark 4.2: From (7) we may write the following recursion for the function $\bar{Q}_{k+1}(.)$:

$$\bar{Q}_{k+1}(\mathbf{O}^{k+1}, \varphi^k;\phi) = \bar{Q}_k(\mathbf{O}^k, \varphi^{k-1};\phi) + \log p_z(\mathbf{z}_{k+1} \mid \mathbf{z}_k)$$

$$+ \chi_{k+1|k+1}(\phi) + \bar{\chi}_{k+1|k+1}(\phi, \mathbf{O}_{k+1}) \ , \ k \geq 1 \qquad (9)$$

[16] Note that for tractability purposes, in our simulations, we shall deal with unconstrained optimization, where we assume a new parameterization to ensure positiveness of the parameter elements by considering square roots as $\phi = (\mathbf{S}(\phi), \mathbb{X}(\phi), \vartheta(\phi))$, where $\mathbf{S}(\phi) = [s_{ij}(\phi)]$ with $s_{ij} = \sqrt{x_{ij}}$, and $\vartheta^m = \sqrt{\theta^m}$.

We now present the following stochastic approximation algorithm which recursively adjusts the parameter vector ϕ by finding the (local) maximum of the objective function $\bar{Q}_k(.)$ at each time step. Approximately, under appropriate regularity conditions, the M-step of the on-line EM algorithm can be written as the following recursion:

$$M\textbf{-step}:\ \ \hat{\phi}_{k+1} = \hat{\phi}_k + \epsilon_{k+1}(\phi) S(\hat{\phi}_k, \mathbf{O}_{k+1}),\ \ k \geq 0 \tag{10}$$

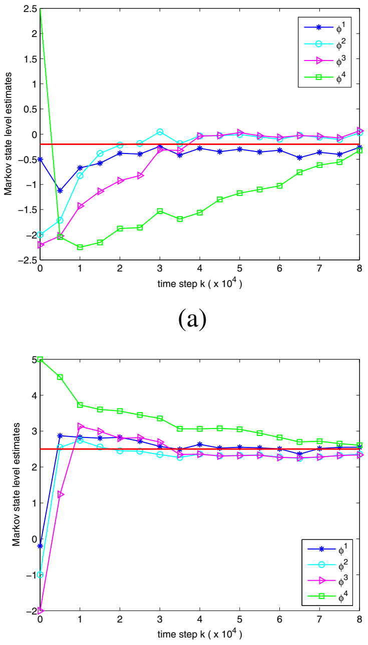
where $\epsilon_{k+1}(\phi)$ is a sequence of decreasing small scalar gains which satisfy certain well-known conditions[17] and $S(\hat{\phi}_k, \mathbf{O}_{k+1})$ is the score function defined by

$$S(\hat{\phi}_k, \mathbf{O}_{k+1}) \triangleq \nabla_\phi\ \bar{Q}_{k+1}(\mathbf{O}^{k+1}, \varphi^k; \phi)\Big|_{\phi=\hat{\phi}_k} \tag{11}$$

Convergence analysis of the stochastic approximation algorithm given by (10) can be carried out using the mean ODE approach (see [11], [12]) under some regularity assumptions on the observation probability distributions along with an additional geometric ergodicity condition on an extended Markov chain [13]. A detailed convergence analysis is presented in [14].

V. Performance Evaluation

In this section, we first illustrate the performance of the on-line EM parameter estimation algorithm. Unless otherwise mentioned, the variables which are assumed fixed throughout the following experiments are as follows: the step size in discretizing the information state $\tilde{\alpha}_{k+1;\hat{\phi}^{(k)}}$ is 0.01; the path loss exponent of the wireless channel is $\varsigma = 2$ for deployment of the sensors in an open rural area; the constant coefficient γ for computing crossover probabilities is $\gamma = 2$. For these simulations, we generated random sequences of 80000 observations obtained by two sensors measuring a two-state Markov chain $\{X_k\}_{k=1}^{\infty}$ with state space $\mathbb{X}(\phi^\circ) = \{-0.2, 2.5\}$ and transition kernel $\mathbf{X}(\phi^\circ) = \begin{bmatrix} 0.94 & 0.06 \\ 0.28 & 0.72 \end{bmatrix}$. The measurement noises $\{w_{m,k}\}_{k=1}^{\infty}$ of the sensors are assumed to be zero-mean white Gaussian noise processes with the noise variance vector $\theta(\phi^\circ)^2 = \sigma^2(\phi^\circ) = (0.5, 0.3)'$. The sensors are located at different distances from the fusion centre with distance vector $d = (80.0, 180.0)'$, where the figures are given in meters. The wireless channels from the sensors to the fusion centre are assumed to be independent and each channel is modeled by a two state Markov chain with state space $\mathbb{C} = \{\tilde{c}_1, \tilde{c}_2\}$. The channel states $\tilde{c}_1$ and $\tilde{c}_2$ represent the corresponding channel gains $g_1^2 = 3 \times 10^{-8}$ and $g_2^2 = 2 \times 10^{-9}$ respectively. The channels are assumed to be asymmetric, that is, having different fading statistics with the transition probability matrices given by $\mathbf{C}^1 = \begin{bmatrix} 0.66 & 0.34 \\ 0.61 & 0.39 \end{bmatrix}$, $\mathbf{C}^2 = \begin{bmatrix} 0.79 & 0.21 \\ 0.32 & 0.68 \end{bmatrix}$. The noise power of the wireless channel for every sensor is $\sigma_v^2 = 3 \times 10^{-14}\ W$. The power levels for each sensor is chosen from the action space $\mathbb{V} = \{65, 30, 10\}$, with the figures being in mW. The action space of the quantization thresholds for each sensor is given by the finite set $\mathbb{U} = \{-0.8, 0.7, 1.8, 3.5\}$. The tradeoff parameter is set to $\beta = 0$. We picked $\epsilon_k(\phi) = 1/(\upsilon_1 + k)^{\upsilon_2(\phi)}$, where υ_2 takes different values depending on the parameter type[18]. The typical values chosen in our simulations are $\upsilon_1 = 50, \upsilon_2(\mathbf{X}) = 0.7, \upsilon_2(\mathbb{X}) = 0.25, \upsilon_2(\theta) = 0.62$.



(a)

(b)

Fig. 1. Convergence of the state level parameters (a) $\tilde{x}_1(\hat{\phi}_k)$ and (b) $\tilde{x}_2(\hat{\phi}_k)$ for various initial conditions. The true values $\tilde{x}_1(\phi^\circ) = -0.2$ and $\tilde{x}_2(\phi^\circ) = 2.5$ are marked by the red lines.

For the Markov chain state levels $\tilde{x}_1(\phi)$ and $\tilde{x}_2(\phi)$, we have examined four different initial estimates in the range around $\pm 3\sigma$ away from the true values $\tilde{x}_1(\phi^\circ)$, and $\tilde{x}_2(\phi^\circ)$, where $\sigma \triangleq \max_m\{\sigma_m(\phi^\circ)\}$. The initial state level estimates are $\phi^1 = (\tilde{x}_1(\phi^1), \tilde{x}_2(\phi^1))' = (-0.5, -0.2)'$, $\phi^2 = (-2.0, -1.0)'$, $\phi^3 = (-2.2, -2.0)'$, and $\phi^4 = (2.5, 5.0)'$. The remaining elements of these initial parameter estimates are the same and given by $(x_{11}, x_{12}, x_{21}, x_{22}, \theta^1, \theta^2) = (0.75, 0.25, 0.4, 0.6, 6.25, 0.25)$.

Fig. 1 shows the effect of the initial estimates on the convergence of the state level parameters $\tilde{x}_1(\hat{\phi}_k)$ and $\tilde{x}_2(\hat{\phi}_k)$ respectively. The estimates are averaged over each 5×10^3 time steps. Convergence is achieved in all cases, though, it is slower only in cases where initial estimates of more than one parameter elements are $\pm 3\sigma$ or further apart from the true values as in ϕ^4. Even in cases where initial estimate of a particular parameter element is close or equal to the true value of some other element, we still achieve a relatively fast convergence, as in ϕ^1.

Fig. 2 shows the convergence of the NVI adaptive cost $j_{\hat{\phi}_k}^{\bar{\lambda}_k}$

[17] the conditions are $\epsilon_k(\phi) \geq 0$, $\sum_k \epsilon_k(\phi) = +\infty$, $\sum_k \epsilon_k^2(\phi) < \infty$, see [11] for further details.

[18] υ_1 may also be chosen based on the parameter type to further improve the convergence rate.

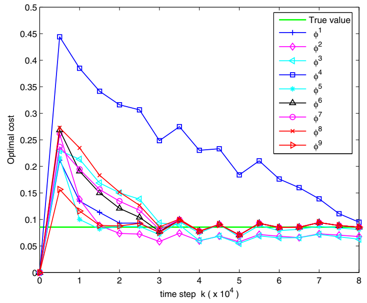


Fig. 2. Convergence of the NVI cost $j_{\hat{\phi}_k}^{\bar{\lambda}_k}$ under different initial conditions. The green line represents true optimal cost $j^*_{\phi^\circ}$.

for various initial conditions. It is clear that ϕ^4 has slower convergence regarding the NVI cost because convergence of its parameters is also achieved in a slower rate as can be seen in Fig. 1. Nevertheless, under all the given initial conditions, the NVI cost does converge to the true optimal cost $j^*_{\phi^\circ}$ obtained by the relative value iteration algorithm (with the true parameter values) presented in [4]. Note that after the transient period, the relative error between the NVI cost and $j^*_{\phi^\circ}$ reduces to less than 4×10^{-3} for $k \geq 50000$.

Fig. 3 shows the optimal state estimation error for various average sum power values across the sensors (obtained by varying the trade-off factor β) for the NVI adaptive policy $\bar{\lambda}_k$ as $k \rightarrow \infty$ and the optimal policy $\lambda^*_{\phi^\circ}$ (obtained by the relative value iteration algorithm of [4] for the true parameter values). As the available average sum power becomes low, clearly the quality of the parameter estimates becomes poorer. As a result, the difference between the average state estimation error computed by the NVI policy and the one computed based on the true optimal policy becomes larger. The solution to this is (if possible) to first compute the parameter estimates using the highest power levels at the sensors in order to learn the model parameters more accurately. This could be thought of as a training phase. Then, based on these estimated parameters, we may perform state estimation to obtain the optimal quantization thresholds and power allocation using the relative value iteration algorithm of [5] for a given average power constraint (see also [4]). The performance of this scheme is shown by the red curve as 'Estimated Policy' in Fig. 3. As expected, the performance of this scheme is closer to that of the optimal policy $\lambda^*_{\phi^\circ}$, particularly when $\beta \rightarrow \infty$.

VI. Summary and Conclusions

In this paper, we have presented a novel method for designing optimal binary quantizers for state and parameter estimation of hidden Markov models using observations obtained by multiple sensors. A coupled recursive ML based parameter estimation algorithm and a nonstationary value iteration (NVI) based adaptive MDP algorithm is proposed at the fusion centre for minimizing the average expected state estimation error under an average sum power constraint across the sensors. Convergence of the parameter estimates and the asymptotic performance of the NVI algorithm are illustrated via extensive numerical studies.

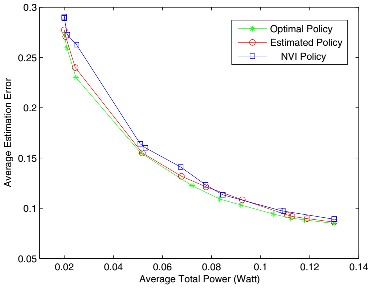


Fig. 3. Optimal and NVI Error/Power curves.

References

[1] E. J. Msechu, S. I. Roumeliotis, A. Ribeiro, and G. B. Giannakis, "Decentralized Quantized Kalman Filtering With Scalable Communication Cost," *IEEE Transactions on Signal Processing*, vol. 56, no. 8, pp. 3727–3741, August 2008.

[2] K. You, L. Xie, S. Sun, and W. Xiao, "Multiple-Level Quantized Innovation Kalman Filter," in *Proceedings of the 17th IFAC World Congress*, vol. 17, Seoul, Korea, July 2008.

[3] M. Huang and S. Dey, "Dynamic Quantizer Design for Hidden Markov State Estimation Via Multiple Sensors With Fusion Center Feedback," *IEEE Transactions on Signal Processing*, vol. 54, no. 8, pp. 2887–2896, August 2006.

[4] ——, "Dynamic Quantization for Multisensor Estimation Over Bandlimited Fading Channels," *IEEE Transactions on Signal Processing*, vol. 55, no. 9, pp. 4696–4702, September 2007.

[5] N. Ghasemi and S. Dey, "Power-Efficient Dynamic Quantization for Multisensor HMM State Estimation over Fading Channels," in *Proceedings of the IEEE International Symposium on Communications, Control and Signal Processing (ISCCSP'08)*, 12–14 March 2008, pp. 1553–1558.

[6] ——, "A Constrained MDP Approach to Dynamic Quantizer Design for HMM State Estimation," *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 1203–1209, March 2009.

[7] M. Hassan, M. M. Krunz, and I. Matta, "Markov-based channel characterization for tractable performance analysis in wireless packet networks," *IEEE Transactions on Wireless Communications*, vol. 3, no. 3, pp. 821–831, May 2004.

[8] V. Krishnamurthy and J. B. Moore, "On-Line Estimation of Hidden Markov Model Parameters Based on the Kullback-Leibler Information Measure," *IEEE Transactions on Signal Processing*, vol. 41, no. 8, pp. 2557–2573, August 1993.

[9] O. Hernández-Lerma, *Adaptive Markov Control Processes*, F. John, J. E. Marsden, and L. Sirovich, Eds. New York, NY, USA: Springer-Verlag New York, Inc., 1989.

[10] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY: John Wiley & Sons, 1994.

[11] A. Benveniste, M. Métivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*, A. V. Balakrishnan, I. Karatzas, and M. Yor, Eds. vol. 22, in Applications of Mathematics, Berlin Heidelberg: Springer-Verlag, 1990.

[12] H. J. Kushner and G. G. Yin, *Stochastic approximation and recursive algorithms and applications*, 2nd ed., B. Rozovskii and M. Yor, Eds. vol. 35, in Applications of Mathematics, New York, NY, USA: Springer-Verlag New York, Inc., 2003.

[13] V. Krishnamurthy and G. G. Yin, "Recursive Algorithms for Estimation of Hidden Markov Models and Autoregressive Models with Markov Regime," *IEEE Transactions on Information Theory*, vol. 48, no. 2, pp. 458–476, February 2002.

[14] N. Ghasemi and S. Dey, "Dynamic Quantization and Power Allocation for Multisensor Estimation of Hidden Markov Models," *submitted to Automatica*, 2009.