Technical Notes and Correspondence

The Optimal Observability of Partially Observable Markov Decision Processes: Discrete State Space

Mohammad Rezaeian, Ba-Ngu Vo, and Jamie Scott Evans

Abstract—**We consider autonomous partially observable Markov decision processes where the control action influences the observation process only. Considering entropy as the cost incurred by the Markov information state process, the optimal observability problem is posed as a Markov decision scheduling problem that minimizes the infinite horizon cost. This scheduling problem is shown to be equivalent to minimization of an entropy measure, called estimation entropy which is related to the invariant measure of the information state.**

Index Terms—**Estimation entropy, observability, partially observable Markov decision processes, sensor scheduling.**

I. Introduction

This technical note focuses on partially observable Markov decision models where the Markov process is autonomous, i.e. the action does not influence its evolution. Instead the action influences observability of the Markov process. This is a general model for monitoring systems that have the capability to adjust their sensing devices for better observability of the scene. The aim is to determine an optimal policy for guiding, tuning or selecting sensors that achieves maximum observability of concurrent events in the scene. For reference we call this the *optimal observability problem*. Dynamically adjusting the measuring devices of the system based on the history of measurements in accordance with the optimal policy allows maximum information flow from the state of the system to the observer. A special case of this problem is sensor scheduling, for example waveform selection in radar systems, [1], [2].

Previous works on sensor scheduling [3], and sensor allocation [4] aimed at finding an optimal sequence of actions (sensor relocations or adjustments) that minimize certain cost functions. They differ from the optimal observability problem where an optimal policy is sought upon which we choose actions based on real measurements. Some previous works which have addressed the online sensor scheduling as the *selection* of one sensor at a time based on past measurements are [5]–[7]. In [5], the cost function comprises of estimation errors and measurement costs, whereas in [7] the observability of the state process is the key criterion for the dynamic sensor scheduling. Another result closely related to this technical note is the optimal control of POMDPs which is analyzed in [8] and has been extensively used in various applications (see, e.g., [9]). The basic difference between the optimal control and optimal observability will be highlighted in our discussion on the optimal observability problem. Information theoretic criteria have increasingly been used in the analysis of stability [10] and fundamental limits of disturbance rejection (controllability) of systems [11], [12]. This technical note introduces an information theoretic approach to the observability of systems.

In this technical note we use an information theoretic measure, estimation entropy [13], [14] to analyze the optimal observability problem. The minimum estimation entropy over different policies for a POMDP is equal to the infinite horizon cost of an optimal Markov Decision Process (MDP) with the cost function being the entropy of the belief state. To capture the basic relation between a POMDP and its corresponding MDP on the belief state in this problem, first we introduce H-processes. For an H-process we define the estimation entropy as a generalized concept of entropy rate. Both analytically and with an example, we show that the more estimation entropy for a POMDP the less observability of the state of a system. The corresponding MDP scheduling problem that finds the policy with minimum estimation entropy can be solved by iterative algorithms such as policy iteration, value iteration or their point-based versions [16].

To facilitate the discussions we use the following notation. The domain of a random variable X is denoted by $\mathbb{X}$ if it is a general space, or by $\mathcal{X}$ if it is a finite set, where in the latter case without loss of generality we assume that $\mathcal{X} = \{1, 2, \cdots, |\mathcal{X}|\}$. A discrete time stochastic process is denoted by $\mathbf{X} = \{X_n : n \in Z\}$. For a process $\mathbf{X}$, a sequence $X_0, X_1, \ldots X_n$ is denoted by X_0^n, whereas X^n refers to $X_{-\infty}^n$. A realization of random variable X is denoted by x, and the probability $Pr(X = x)$ is shown by $p(x)$ (similarly for conditional probabilities), whereas $\underline{p}(X)$ represents a row vector as the distribution of X, i.e. the k-th element of the vector $\underline{p}(X)$ is $Pr(X = k)$. For a random variable X defined on a set $\mathcal{X}$, we denote by $\nabla_{\mathcal{X}}$ the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$, where $\mathbb{R}(\mathbb{R}^+)$ is the set of (non-negative) real numbers. A specific element of a vector or matrix is referred to by its index in square brackets. The entropy of a discrete random variable X is denoted by $H(X) = \sum_{x \in \mathcal{X}} -Pr(x)\log(Pr(x))$, whereas $h : \nabla_{\mathcal{X}} \rightarrow \mathbb{R}^+$ represents the entropy function over $\nabla_{\mathcal{X}}$, i.e. $h(\underline{p}(X)) = H(X)$ for all possible random variables X on $\mathcal{X}$.

II. Optimal Observability Problem for POMDP

Similar to a hidden Markov process (HMP) [13], a POMDP is a Markov process $\{S_n\}_{n=-\infty}^{\infty}$, $S_n \in \mathcal{S}$ that is observed through a discrete memoryless channel with the observation process being $\{Z_n\}_{n=-\infty}^{\infty}$, $Z_n \in \mathcal{Z}$. We assume that both processes are stationary and time homogenous. In this case, a POMDP is defined by a transition probability matrix Q of the Markov process and the measurement (emission) matrix T of the memoryless channel. The elements of matrices $Q_{|\mathcal{S}|\times|\mathcal{S}|}$ and $T_{|\mathcal{S}|\times|\mathcal{Z}|}$ are the conditional probabilities, $Q[s, s'] = p(S_{n+1} = s'|S_n = s)$, $T[s, z] = p(Z_n = z|S_n = s)$. Unlike a HMP, the matrices Q and T of the POMDP are functions of a control action $a \in \mathbb{A}$, and we choose the action based on our past and current observations.

Manuscript received January 13, 2009; revised August 27, 2009, March 08, 2010, and June 10, 2010; accepted July 06, 2010. Date of publication September 09, 2010; date of current version December 02, 2010. This work was supported by the Australian Research Council, Discovery Grant DP0878158. Recommended by Associate Editor C. Szepesvari.

M. Rezaeian and J. S. Evans are with the Department of Electrical and Electronic Engineering, University of Melbourne, Victoria, 3010, Australia (e-mail: rezaeian@unimelb.edu.au; jse@ee.unimelb.edu.au).

B.-N. Vo is with the School of Electrical Electronic and Computer Engineering, M018, The University of Western Australia, Crawley WA 6009 Perth, Australia (e-mail: ba-ngu.vo@uwa.edu.au).

Color versions of one or more of the figures in this technical note are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2010.2074231

As with a HMP, for a POMDP we can define two random vectors π_n and ρ_n on the simplexes $\nabla_{\mathcal{S}}, \nabla_{\mathcal{Z}}$, respectively, which are functions of Z^{n-1}, [13]

$$\pi_n(Z^{n-1}) = \underline{p}(S_n|Z^{n-1}) \tag{1}$$

$$\rho_n(Z^{n-1}) = \underline{p}(Z_n|Z^{n-1}). \tag{2}$$

The random vector π_n has elements $\pi_n[k] = p(S_n = k|Z^{n-1})$, and similarly for ρ_n. The random vector π_n as a function of the random sequence Z^{n-1} is called the *information-state* [8] since it encapsulates all the information about the state at time n through all past measurements.

A. A Unified Framework for Observable, Partially Observable and Hidden Markov Processes

Some key properties of the information state π_n and ρ_n that are shared in HMP and POMDP are preserved and explainable under only existence of a pair of mappings between these variables, irrespective of the map definitions. The existence of such mappings in fact implies that these processes can be described by an iterated function system, [13]. We define such a general framework as an H-process and prove its basic properties. Then we use H-process to characterize the observability of POMDPs in the sequel.

Definition 1: A pair of correlated processes $(\mathbf{Z}, \mathbf{S})$ with finite domain sets $\mathcal{Z}, \mathcal{S}$, respectively, is called an H-process if the sequences π_n and ρ_n defined in (1) and (2) are related by some mappings $\zeta : \nabla_{\mathcal{S}} \to \nabla_{\mathcal{Z}}$ and $\eta : \mathcal{Z} \times \nabla_{\mathcal{S}} \to \nabla_{\mathcal{S}}$

$$\rho_n = \zeta(\pi_n), \quad \pi_{n+1} = \eta(z_n, \pi_n). \tag{3}$$

We refer to $\mathbf{Z}$ as the observable component and $\mathbf{S}$ as the hidden component of the H-process. H-process can also refer to a single process, where the two components $\mathbf{Z}$ and $\mathbf{S}$ in the above definition are the same, hence $\rho_n = \pi_n$. In other words, a single process $\mathbf{Z}$ is an H-process if ρ_n defined in (2) recursively satisfies $\rho_{n+1} = \tilde{\eta}(z_n, \rho_n)$ for a given function $\tilde{\eta}$. Therefore H-process can also be seen as a generalization of (observable) Markov process. We note that a time homogenous process $\mathbf{Z}$ is Markov if $\rho_{n+1} = \tilde{\eta}(z_n)$ for a given function $\tilde{\eta}$.

An example of an H-process is the hidden Markov process, where the mapping ζ, η are

$$\zeta(\pi_n)[m] = \sum_k T[k, m]\pi_n[k],$$
$$\eta(z, \pi) = \frac{\pi D(z)Q}{\pi D(z)\underline{1}} \tag{4}$$

and $D(z)$ is a diagonal matrix with $d_{k,k}(z) = T[k, z]$, $k = 1, 2, \ldots, |\mathcal{S}|$. In contrast, for a POMDP, the relationship of the matrices $T(a)$ and $Q(a)$ on the control action implies

$$\rho_n = \zeta_{a_n}(\pi_n), \quad \pi_{n+1} = \eta_{a_n}(z_n, \pi_n) \tag{5}$$

where

$$\zeta_a(\pi) = \pi T(a), \quad \eta_a(z, \pi) \triangleq \frac{\pi D(z, a)Q(a)}{\pi D(z, a)\underline{1}} \tag{6}$$

and $D(z, a)$ is a diagonal matrix with $d_{k,k}(z) = T(a)[k, z]$, $k = 1, 2, \ldots, |\mathcal{S}|$. Hence, π_n, ρ_n are not only functions of Z^{n-1} but also depend on the sequence of actions a^n, and if this sequence is independent of the observations (open loop), then the pair $(\mathbf{Z}, \mathbf{S})$ is not an H-process.

For a POMDP, a control policy is a rule, defined by a function $w : \nabla_{\mathcal{S}} \to \mathcal{A}$ for choosing actions based on the belief π_n. Having a fixed policy w for selecting actions, i.e. $a_n = w(\pi_n)$, (5) reduces to $\rho_n = \zeta_w(\pi_n)$, $\pi_{n+1} = \eta_w(z_n, \pi_n)$, where ζ_w and η_w are defined by

$$\zeta_w(\pi) = \pi T(w(\pi)), \quad \eta_w(z, \pi) \triangleq \frac{\pi D(z, w(\pi))\, Q(w(\pi))}{\pi D(z, w(\pi))\, \underline{1}}. \tag{7}$$

Thus, for a POMDP with a fixed policy w (called a closed loop POMDP), the pair $(\mathbf{Z}, \mathbf{S})$ is an H-process defined by the mappings η_w and ζ_w. We note that the closed loop POMDP is not a HMP (which requires that ζ be a linear transformation as in (4)), but they share the basic properties of H-processes that we explain next. Also, according to relation of the H-process with complete observable stochastic system that we show later, the closed loop POMDP corresponds to a (complete observable) Markov decision process with a given decision policy.

1) Markov Information State: A key property of an H-process is that the information state π_n is a Markov chain on $\nabla_{\mathcal{S}}$. It can be seen from (3) that for an H-process knowing π_n, the only randomness in π_{n+1} is due to z_n, whose distribution is only a function of π_n, c.f. $\underline{p}(Z_n|Z^{n-1}) = \zeta(\pi_n(Z^{n-1}))$. Therefore knowing π_n, uniquely defines the distribution of π_{n+1}, independent of its prior history. Moreover, knowing π_n, according to (3) π_{n+1} can at most take $|\mathcal{Z}|$ different values, and if these values are distinct, then they have probabilities $(\zeta(\pi_n))[l], l = 1, 2 \ldots, |\mathcal{Z}|$, otherwise the overlapping values have the corresponding sum of probabilities, i.e. for any n

$$Pr\{\pi_{n+1} = x' | \pi_n = x\} = \sum_{l : x' = \eta(l, x)} \zeta(x)[l]. \tag{8}$$

In general, a time homogeneous Markov chain $\mathbf{X}$ on a general space $\mathbb{X}$ is defined by a transition probability kernel $P(x, B)$, $x \in \mathbb{X}$, $B \in \mathcal{B}(\mathbb{X})$, ($\mathcal{B}$ refers to the Borel sets), where $P(x, B) = Pr\{x_{n+1} \in B|x_n = x\}$. Extending (8) from singleton set $\{x'\}$ to any set B, the transition probability kernel $P(x, B)$ for the Markov chain π_n is

$$P(x, B) = \sum_{l=1}^{|\mathcal{Z}|} 1_B(\eta(l, x))\, \zeta(x)[l] \tag{9}$$

where $1_B(\cdot)$ is the indicator function of B. This kernel encapsulates both the state dynamic and the measurement characteristics. We show in Section III that the invariant measure of this kernel characterizes the observability of the hidden component of an H-process through its observable component.

For a closed loop POMDP with a policy w, the corresponding transition probability for the information state is

$$P_w(x, B) = \sum_{l=1}^{|\mathcal{Z}|} 1_B(\eta_w(l, x))\, \zeta_w(x)[l] \tag{10}$$

which may not necessarily have a unique invariant measure. Nonetheless, by extending the Kaijser's result [22], we can show that for a closed loop POMDP with a given policy w, if for every π the matrices $Q(w(\pi))$ and $T(w(\pi))$ have non-zero elements, then the transition probability (10) has a unique invariant measure.

Note that in an H-process if ρ_n and π_{n+1} depend also on an action variable a_n as in $\rho_n = \zeta_{a_n}(\pi_n)$, $\pi_{n+1} = \eta_{a_n}(z_n, \pi_n)$ through arbitrary functions η_a and ζ_a, then the Markov property of information state is still valid, but instead of (9) we have a Markov decision process with the transition kernel

$$P_a(x, B) = \sum_{l=1}^{|\mathcal{Z}|} 1_B(\eta_a(l, x))\, \zeta_a(x)[l]. \tag{11}$$

In this case the problem of finding a policy w for $a_n = w(\pi_n)$ that minimizes

$$J(x) = \lim_{n\to\infty} \sum_{t=0}^{n-1} \frac{1}{n} \mathbb{E}\left[c(\pi_t)|\pi_0 = x\right]$$

for a given x and cost function c is a (fully observable) Markov decision scheduling problem.

2) Sufficient Statistic Property for Information State: For an H-process we can write

$$\underline{p}(Z_n|\pi_n, Z^{n-1}) = \underline{p}(Z_n|Z^{n-1}) = \rho_n = \zeta(\pi_n) \tag{12}$$

where the first equality is due to π_n being a function of Z^{n-1}. Since the right hand side of (12) is (only) a function of π_n (and it is a distribution on $\mathcal{Z}$), the left hand side must be equal to $\underline{p}(Z_n|\pi_n)$, i.e. we have shown $\underline{p}(Z_n|\pi_n, Z^{n-1}) = \underline{p}(Z_n|\pi_n) = \zeta(\pi_n)$. This shows that for estimating the upcoming observation (or its distribution), π_n is a sufficient statistic for all the past observations. By a similar argument we have

$$\underline{p}(S_n|\pi_n, Z^{n-1}) = \underline{p}(S_n|Z^{n-1}) = \pi_n = \underline{p}(S_n|\pi_n) \tag{13}$$

which shows that for estimating the state at time n, π_n is a sufficient statistic for the past measurement before n. This generalizes the sufficient statistic characterization of the information state that we had for the HMP to any H-process.

Moreover, the sufficient statistic property is valid for π_k replacing Z^{k-1} in estimating S_n for any $k < n$. This is by induction and

$$\begin{aligned}\pi_n(Z^{n-1}) &= \underline{p}(S_n|Z_{n-1}, Z^{n-2}) \\ &= \eta(Z_{n-1}, \pi_{n-1}) \\ &= \underline{p}(S_n|Z_{n-1}, \pi_{n-1})\end{aligned} \tag{14}$$

i.e. $\pi_{n-1}(Z^{n-2})$ is a sufficient statistic for Z^{n-2} and can replace Z^{n-2} as the dependent variables of $\pi_n(Z_{n-1}, \pi_{n-1})$. In particular, π_0 can replace $Z_{-\infty}^{-1}$, so we can also write (1) as

$$\pi_n\left(Z_0^{n-1}, \pi_0\right) = \underline{p}\left(S_n|Z_0^{n-1}, \pi_0\right), \quad \forall n > 0. \tag{15}$$

3) H-Process as a Stochastic System: A general stochastic system [19] is defined as

$$\begin{aligned}x_{n+1} &= f_n(x_n, u_n, d_n), \\ y_n &= h_n(x_n, v_n)\end{aligned} \tag{16}$$

for given functions $f_n(\cdot,\cdot,\cdot)$ and $h_n(\cdot,\cdot)$, where x_n, u_n and y_n are the state, input and output of the system, respectively, and d_n and v_n are independent state and measurement noise variables, respectively. Specializing this system to time invariant complete observable close loop with a stationary policy g, i.e. $y_n = x_n$, and $u_n = g(x_n)$, we have

$$x_{n+1} = f\left(x_n, g(x_n), d_n\right) \triangleq \eta(d_n, x_n). \tag{17}$$

In this special case, the only sources of randomness are the random variables $X_0, D_0, D_1, \cdots, D_n$. The joint distribution of these variables defines the stochastic system, and if they are independent, then the state process is Markovian [19, p 18]. We see from (17) that an H-process is such a stochastic system with state of $x_n = \pi_n$ and discrete noise component $d_n = z_n$, but the noise components are not independent. Their conditional distribution is a function of state, c.f. $P(Z_n|Z^{n-1}) = \rho_n(Z^{n-1}) = \zeta(\pi_n(Z^{n-1}))$. The Markovian nature of the state is however established through $\rho_n = \zeta(\pi_n)$. The dependent variables $Z_0, Z_1, \cdots, Z_n$ have the following joint distribution

$$\begin{aligned}P(z_0, z_1, \cdots, z_n) &= \rho_0[z_0]\rho_1[z_1]\cdots\rho_n[z_n] \\ &= \zeta(\pi_0)[z_0]\zeta(\pi_1)[z_1]\cdots\zeta(\pi_n)[z_n].\end{aligned} \tag{18}$$

While for an H-process $(\mathbf{Z}, \mathbf{S})$ the $\mathbf{S}$ component is partially observable, the H-process representation gives a complete observable model, in particular turning a POMDP into a MDP problem.

B. Observability of POMDP

We now discuss the observability of POMDPs via the complete observable Markov process π_n characterized in the H-process representation of POMDPs. A policy w that dictates the control action a_n based on belief π_n in general can be designed for two basic purposes, controllability or observability of a POMDP. In the controllability problem the aim is to control the state process to move it towards more favored states, but in the observability problem the state process is autonomous and the aim is to dynamically adjust the measuring apparatus to have the best observation of the state process. Both problems can be defined based on a specific cost function on the belief space[1] $\nabla_{\mathcal{S}}$ and the aim is to find the optimal policy w on this space that minimizes the infinite horizon (or discounted) expected cost over time. In the controllability problem, associated to each state k is a cost $\beta[k]$ where $\beta \in \mathbb{R}^{|\mathcal{S}|}$ is a fixed column vector. Hence, with a belief π on the probability distribution of states, the cost function on $\nabla_{\mathcal{S}}$ is $c(\pi) = \pi\beta$ which is a linear function. The value function, representing the prospective expected cost on $\nabla_{\mathcal{S}}$ is then a piecewise linear and convex function. Dynamic programming methods such as value iteration yield the optimal action to be taken at every point $\pi \in \nabla_{\mathcal{S}}$, i.e. the optimal policy. Although the problem is intractable for $\nabla_{\mathcal{S}}$ of high dimension, methods such as incremental pruning [21] or point-based value iteration [16] are tractable approximations.

The focus of this technical note is the optimal observability of POMDPs. Since the state process is autonomous, the matrix Q does not depend on the action a, and it is only T that is a function of a. Moreover, the cost function cannot be a linear function on $\nabla_{\mathcal{S}}$. The (positive) cost function needs to be designed to penalize the belief π_n as it moves away from the vertices of $\nabla_{\mathcal{S}}$. At the vertices, the belief about the state is complete certainty, thus incurs zero (minimum) cost. More specifically, we consider the following definition of the observability problem in this technical note.

- *The Optimal Observability Problem* for an autonomous POMDP is the optimization problem over policies w that minimizes the average cost

$$J(x) = \lim_{n\to\infty} \sum_{t=0}^{n-1} \frac{1}{n} \mathbb{E}\left[c(\pi_t)|\pi_0 = x\right]$$

for any initial belief x, where the cost function is the entropy function, $c(\pi) = h(\pi)$, and π_t is the belief Markov process with kernel P_w in (10).

In the optimal observability problem, the optimal policy ensures that as the state process evolves autonomously, the belief π_n in its expectation hops only between the regions close to the vertices, so the average ambiguity about state (cost) is minimized. This implies the maximum observability of hidden process from the observed process in the long run.

[1] In a more general setting the cost function can be defined on $\nabla_{\mathcal{S}} \times \mathbb{A}$. This is when we want to avoid, as much as possible, some specific actions (for other reasons like costly measurement) although they are good in affecting the trajectory of the intended process.

In [20], the problem of finding an optimal policy for a POMDP for which the control action only influences the measurement has been considered with a cost function $c(\pi) = 1 - \pi\pi^T$ (T stands for transpose). This also has zero cost at vertices. By piecewise linear approximation of the cost function and discounted cost criterion, the problem can be turned into an optimization similar to the controllability problem, where an approximate and computationally intensive solution can be obtained using the value iteration algorithm. However this cost function does not relate to a measure of information, hence it does not provide an information theoretic measure of observability. Shannon entropy is a natural choice when uncertainty or inversely, observability needs to be measured. We relate the observability problem with the entropy criteria to the minimization of a limiting entropy measure that we define in the next section.

III. The Estimation Entropy

The entropy rate of a process $\mathbf{Z}$ [18], denoted by $\hat{H}_Z$ is the limit of Cesaro mean of the i-sequence $H(Z_i|Z_0^{i-1}) = H(Z_0^i) - H(Z_0^{i-1})$, i.e.

$$\hat{H}_Z = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H\left(Z_i|Z_0^{i-1}\right) = \lim_{n\to\infty} \frac{1}{n} H\left(Z_0^n\right). \tag{19}$$

Extending the above definition of entropy rate for a single process to a pair of processes, we define estimation entropy as a limiting entropy measure for an H-process.

Definition 2: The estimation entropy of an H-process $(\mathbf{Z}, \mathbf{S})$ is

$$\hat{H}_{S/Z} \triangleq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H\left(S_i|Z_0^{i-1}\right). \tag{20}$$

We note that if for an H-process $(\mathbf{Z}, \mathbf{S})$ the function ζ is one to one and invertible, then the observable component $\mathbf{Z}$ is also an H-process with $\tilde{\eta}(z_n, \rho_n) = \zeta(\eta(z_n, \zeta^{-1}(\rho_n)))$. Moreover, the estimation entropy of a single H-process reduces to its entropy rate, c.f. $\hat{H}_{Z/Z} = \hat{H}_Z$. Therefore estimation entropy, in particular for H-processes can be seen as a generalization of entropy rate from a single process to a pair of processes.

Since convergence of a sequence α_n implies the convergence of its Cesaro mean (i.e. $1/n \sum_{i=1}^n \alpha_i$) to the same limit [18, Theorem 4.2.3], for an H-process

$$\hat{H}_{S/Z} = \lim_{n\to\infty} H\left(S_n|Z_0^{n-1}\right) \tag{21}$$

when the limits in (21) exist. Similarly, for the entropy rate we have $\hat{H}_Z = \lim_{n\to\infty} H(Z_n|Z_0^{n-1})$, when the limit exists, [18]. However, the non-existence of these limits does not mean that the entropy rate or estimation entropy do not exist. One sufficient condition for the existence of the limit in (21) is the stationarity of the H-process. This is because in the stationary case the n-sequences of $H(S_n|Z_0^{n-1})$ is non-increasing and positive. Therefore for a stationary H-process the estimation entropy can also be written as (21).

From (21) (or (20)) we see that the estimation entropy for an H-process is the limit of (running average of) residual uncertainty about the hidden component under the knowledge of all past observed process, thus it inversely measures the observability of the hidden process. However, we also show that under ergodicity conditions the estimation entropy is the long run average entropy of the belief process. To this end, we need to define and use the following operator for a given transition probability kernel P on ∇ [17]

$$Pf(x) \triangleq \int_{\nabla} f(y)P(x, dy) \tag{22}$$

for any real-valued bounded measurable function f. The n times repetition of this operator on a function f is denoted by $P^n f(x)$, and it is equal to $P^n f(x) = \mathbb{E}[f(x_n)|x_0 = x]$ when x_n evolves by P as a Markov process.

Before presenting the relation between the estimation entropy and the invariant measure of P, first we obtain a relation between conditional entropy and the above operator of P.

Lemma 1: For any H-process $H(S_n|Z_0^{n-1}, \pi_0 = x) = P^n h(x)$, where P is defined by (9) and h is the entropy function.

Proof: From definition of conditional entropy and (15)

$$\begin{aligned} H\left(S_n|Z_0^{n-1}, \pi_0 = x\right) &= \sum_{\mathbf{z}} p\left(Z_0^{n-1} = \mathbf{z}|\pi_0 = x\right) \\ &\quad \times h\left(\underline{p}(S_n|Z_0^{n-1} = \mathbf{z}, \pi_0 = x\right) \\ &= \sum_{\mathbf{z}} p\left(Z_0^{n-1} = \mathbf{z}|\pi_0 = x\right) \\ &\quad \times h\left(\pi_n(\mathbf{z}, x)\right) \\ &= \mathbb{E}\left[h(\pi_n)|\pi_0 = x\right] \\ &= P^n h(x). \end{aligned} \tag{23}$$

■

Theorem 1: For an H-process, if P has unique invariant measure μ, then

$$\hat{H}_{S/Z} = \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} P^t h = \int_{\nabla_{\mathcal{S}}} h d\mu. \tag{24}$$

Proof: The second equality is the direct result of the mean ergodic theorem [17]. Moreover according to this theorem the middle expression is a constant independent of x. Assuming the limit of $P^n h(n \to \infty)$ exists, it will be equal to that constant, and we have

$$\begin{aligned} \lim_{n\to\infty} P^n h &= \lim_{n\to\infty} H\left(S_n|Z_0^{n-1}, \pi_0 = x\right) \\ &= \lim_{n\to\infty} H\left(S_n|Z_0^{n-1}\right) \end{aligned} \tag{25}$$

where the first equality is from Lemma 1, and the second one is due to the fact that if for a set of random variables X, Y, Z, the quantity $H(Y|Z, X = x)$ is invariant with x, then $H(Y|Z) = H(Y|Z, X) = H(Y|Z, X = x)$. The right hand side of (25) is equal to $\hat{H}_{S/Z}$ when the limit exists. ■

For an H-process corresponding to a closed loop POMDP with a control policy w, $\hat{H}_{S/Z}$, kernel P and its invariant measure are all functions of policy w, and if the invariant measure of P_w, denoted by μ^w exists, then from Theorem 1

$$\hat{H}_{S/Z}(w) = \int_{\nabla_{\mathcal{S}}} h d\mu^w.$$

Since the function h is fixed over $\nabla_{\mathcal{S}}$, the invariant measure μ^w defines the estimation entropy and hence it characterizes the observability of the system under policy w. Hence the analysis of observability of a closed loop POMDP boils down to finding the invariant measure of the kernel P under the control policy.

On the other hand, since $P^n h(x) = \mathbb{E}[h(\pi_n)|\pi_0 = x]$, from Theorem 1 we see that under ergodicity condition (existence of a unique invariant measure), the average cost criterion for the optimal observability problem is the estimation entropy

$$\begin{aligned} J(x) &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{E}\left[h(\pi_t) | \pi_0 = x\right] \\ &= \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} P^t \\ &= \hat{H}_{S/Z}. \end{aligned} \tag{26}$$

Moreover, this average cost is independent of the initial belief π_0. This also shows the key property of the estimation entropy as the infinite horizon average uncertainty about the state (or a hidden) process, so it inversely represents the observability of the state process.

For a closed loop POMDP, under ergodicity of P_w, we write (26) as

$$\hat{H}_{S/Z}(w) = J(w, x), \quad \forall x. \tag{27}$$

The optimal observability problem therefore is minimization of either sides of (27) over policies w, so it can be viewed either as the minimization of $\hat{H}_{S/Z}$ as a function of w or as it was noted before the average cost MDP scheduling problem of $J(w, x)$ for any x. This MDP is defined by the set of conditional probabilities P_a in (11) where ζ_a and η_a are defined by (6). We briefly explain this approach in the next section.

IV. Optimum Observation Policy

Here we formalize the POMDP observability problem with a finite control set as an average cost MDP scheduling problem. Such an observability problem is uniquely defined by three integers (M, L, A), an $M \times M$ primitive probability matrix Q, and a set of $M \times L$ probability matrices T_a, $a = 1, 2, \ldots, A$. The corresponding MDP problem is defined as follows:

- The Markov decision process evolves on the state space $\mathbb{X}$, where $\mathbb{X}$ is the probability simplex in $\mathbb{R}^M$.
- The admissible action set for any $x \in \mathbb{X}$ is $\mathcal{A} = \{1, 2, \cdots, A\}$.
- The set of conditional probability distributions $P_a(x, B)$, $a \in \mathcal{A}$, is defined by (11).
- For any stationary policy $w : \mathbb{X} \to \mathcal{A}$ the state process $X(w) = \{X_t(w) : t \in \mathrm{Z}_+\}$ is a Markov chain with conditional probability $P_w(x, B) = P_{w(x)}(x, B)$.
- The average cost of a policy w for a given initial condition x is

  $$J(w, x) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=0}^{N-1} \mathbb{E}\left[c\left(X_t(w)\right) | X_0 = x\right]$$

  where the cost function $c : \mathbb{X} \to [0, \log(M)]$ is the entropy function defined by

  $$c(x) = h(x) \triangleq -\sum_m x[m] \log x[m].$$

Objective: Find the optimal policy w^* where $J(w^*, x) \leq J(w, x)$ for all polices w and any initial state x.

As an average cost MDP scheduling problem the objective can be achieved by the policy iteration algorithm (PIA). The solution w^* will then be the optimum observation policy. However since $\mathbb{X}$ is not finite, the PIA may not converge. The convergence of this algorithm can be analyzed using methods from [15] which is beyond the scope of this technical note. Alternative approximation methods such as point-based value iteration [16] can be used for convergence and computational tractability.

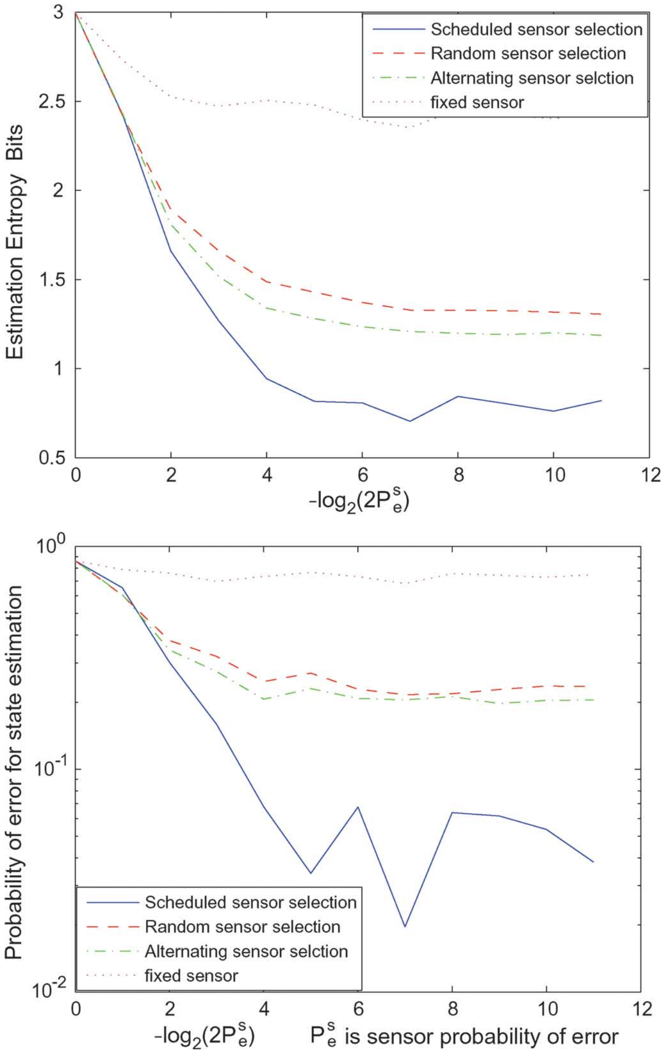

Fig. 1. Comparison of achieved estimation entropy and the probability of error in state estimation for various sensor selection policies as functions of sensor reliability.

A. Numerical Example

We consider a system with 8 states evenly spaced on a circle. Corresponding to each state there is a binary sensor that, when selected, indicates if the system is in that state or not, with the probability of erroneous indication being P_e^s. So for sensor i, conditioned that the state is i, the output will be '1' with the probability of $1 - P_e^s$, and '0' with the probability of P_e^s, and conditioned on state not being i, the opposite will happen. At any given time only one sensor can be selected for the observation of state. These binary sensors mimic a sensor network setup for detecting targets passing by the sensors, where for each sensor the probability of false alarm is P_e^s and it is equal to one minus the probability of target detection. We consider a Markov state process with the dynamic that it stays in the same state with probability of 0.9 but can go to both neighboring states with probability of 0.05 each.

The policy iteration algorithm has been adapted for this optimal observability problem to obtain a policy that we call it the scheduled sensor selection policy. We have compared this sub-optimal policy with 3 other heuristic policies in Fig. 1, namely, random sensor selection, alternating through all sensors, and only using one sensor. The comparison is in terms of the achieved estimation entropy as well as probability of the error for MAP estimation of the state, both at various sensor error probabilities. The results show the efficiency of the scheduled policy in reducing the estimation entropy and achieving less probability of error as compared to heuristic policies. As the figures show, this efficiency increases at more accurate sensor measurements, or less sensor probability of error. The two graphs also show the correspondence of the estimation entropy with the achievable probability of error in state estimation. Since estimation entropy inversely relates to the observability of the state process, this correspondence suggests an interpretation of optimal observability as achieving the least probability of error in state estimation for a given set of sensors.

V. Conclusion

In this technical note, we showed the fundamental role of estimation entropy as a new information theoretic measure in defining the observability for a system. We showed that the optimal closed loop control of the measuring devices of a system for providing the most flow of information from the state of the system to the observer is ultimately achieved by minimizing the estimation entropy over different control policies. The estimation entropy was related to the integral of the entropy function over the belief space, where the measure for this space is considered to be the invariant measure of a transition probability kernel. This probability kernel is well defined for the H-process corresponding to a closed loop POMDP, and under mild conditions the existence of its unique invariant measure can be verified. Here the H-process was defined to capture the core properties that are shared between the controlled and uncontrolled hidden processes in one hand, and to represent the closed loop POMDP's by a fully observable Markov process on the other hand.

Future extension of this work is possible for general state space models. For continuous alphabet the belief space is infinite dimensional. Although the belief transition probability kernel can be formulated, finding the invariant measure on this space is not practical. An alternative approach for the analysis of the observability of such systems is via the possible relation between the estimation entropy and the entropy of error in the minimum entropy filter defined in [23].

References

[1] D. J. Kershaw and R. J. Evans, "Waveform selective probabilistic data association," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, no. 4, pp. 1180–1188, Apr. 1997.

[2] B. La Scala, M. Rezaeian, and B. Moran, "Optimal adaptive waveform scheduling for target tracking," in *Proc. Int. Conf. Inform. Fusion*, Philadelphia, PA, Jul. 2005, pp. 552–557.

[3] S. Singh, N. Kantas, A. Doucet, B.-N. Vo, and R. J. Evans, "Simulation-based optimal sensor scheduling with application to observer trajectory planning," *Automatica*, vol. 43, no. 5, pp. 817–830.

[4] Y. Takeuchi, M. Sowa, and K. Horikawa, "An information theoretic scheme for sensor allocation of linear least-squares estimation," in *Proc. 41st SICE Annu. Conf.*, 2002, pp. 539–544.

[5] V. Krishnamurthy and D. Djonin, "Structured threshold policies for dynamic sensor scheduling—A partially observed Markov decision process approach," *IEEE Trans. Signal Processing*, vol. 55, no. 10, pp. 4938–4957, Oct. 2007.

[6] J. S. Evans and V. Krishnamurthy, "Optimal sensor scheduling for hidden Markov model state estimation," *Int. J. Control*, vol. 74, no. 18, pp. 737–742, Dec. 2001.

[7] M. Rezaeian, "Sensor scheduling for optimal observability using estimation entropy," in *Proc. 5th IEEE Int. Conf. Pervasive Comput. Commun.*, New York, Mar. 2007, pp. 307–312.

[8] R. D. Smallwood and E. J. Sondik, "Optimal control of partially observed Markov processes over a finite horizon," *Oper. Res.*, vol. 21, pp. 1071–1088, 1973.

[9] D. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Boston, MA: Athena Scientific, 2001, vol. 1.

[10] P. Minero, M. Franceschetti, S. Dey, and G. N. Nair, "Data rate theorem for stabilization over time-varying feedback channels," *IEEE Trans. Autom. Control*, vol. 54, no. 2, pp. 243–255, Feb. 2009.

[11] S. Yu and P. G. Mehta, "Fundamental performance limitations via entropy estimates with hidden Markov models," in *Proc. 46th IEEE Conf. Decision Control*, New Orleans, LA, Dec. 12–14, 2007, pp. 3982–3988.

[12] N. C. Martins, M. A. Dahleh, and J. C. Doyle, "Fundamental limitations of disturbance attenuation in the presence of side information," *IEEE Trans. Autom. Control*, vol. 52, no. 1, pp. 2523–2529, Jan. 2007.

[13] M. Rezaeian, Hidden Markov Process: A New Representation, Entropy Rate and Estimation Entropy preprint [Online]. Available: http://arxiv.org/abs/cs/0606114

[14] M. Rezaeian, "Estimation entropy and its operational characteristics in information acquisition systems," in *Proc. 11th Int. Conf. Inform. Fusion*, Jul. 2008, pp. 1–5.

[15] S. Meyn, "The policy iteration algorithm for average reward Markov decision processes with general state space," *IEEE Trans. Autom. Control*, vol. 42, no. 12, pp. 1663–1679, Dec. 1997.

[16] T. Smith and R. Simmons, "Point-based POMDP algorithms: Improved analysis and implementation," in *Proc. 21st Conf. Uncertainty Artif. Intell.*, 2005, [CD ROM].

[17] O. Hernandez and J. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*. New York: Springer-Verlag, 1999.

[18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[19] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*. Englewood Cliffs, NJ: Prentice Hall, 1986.

[20] V. Krishnamurthy, "Algorithms for optimal scheduling and management of hidden Markov model sensors," *IEEE Trans. Signal Processing*, vol. 50, no. 6, pp. 1382–1396, Jun. 2002.

[21] A. R. Cassandra, M. L. Littman, and N. L. Zhang, "Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes," in *Proc. 13th Annu. Conf. Uncertainty Artif. Intell.*, 1997, [CD ROM].

[22] T. Kaijser, "A limit theorem for partially observed Markov chains," *Annals Prob.*, vol. 3, pp. 677–696, 1975.

[23] L. Guo and H. Wang, "Minimum entropy filtering for multivariate stochastic systems with non-gaussian noises," *IEEE Trans. Autom. Control*, vol. 51, no. 4, pp. 695–700, Apr. 2006.