STRUCTURE OF OPTIMAL POLICIES IN ACTIVE SENSING

Aditya Mahajan

Department of Electrical Engineering
McGill University, Montreal, QC, Canada

ABSTRACT

We consider the optimal design of a sensing system in which a sensor can choose how and when to communicate to an estimator. The optimal choice of transmission and estimation policies is made difficult by the fact that the sensor and the estimator may use their entire history of observations. Traditionally, Markov decision theory is used to analyze such multi-stage decision problems. But, Markov decision theory assumes a single decision maker—an assumption that is not satisfied in an active sensing system that has two decision makers with different information. In this paper, we use the approach of Nayyar *et al* (2011) to investigate the system as a dynamic team. Using a series of structural results, we show that the optimal policy is easy to implement. We also obtain a dynamic programming decomposition to find optimal sensing and estimation policies.

Index Terms— sensor networks, estimation theory, dynamic teams, stochastic control

1. INTRODUCTION

1.1. Motivation

Energy consumption is an important factor in the cost of operating a sensor network. Sensors have limited battery. Once the battery runs out, either the battery or the sensor needs to be replaced, often manually, and at significant cost. Therefore, it is important to investigate techniques for reducing the energy consumption of the sensor network, even at the cost of slight poorer performance.

One way to conserve energy is to allow the sensor to go to sleep either for a pre-specified duration [1, 2], or for an arbitrary duration to be woken up by external signal [3]. A related approach is to allow the fusion center to control which sensors to active at each time step [4].

Another approach is to allow the sensor not to choose when to transmit or not. The simplest instance of such a scheme is sensor "censoring" considered in [5, 6] for decentralized hypothesis testing in a sensor network where each sensor takes one measurement and chooses to transmit its likelihood ratio or not.

In this paper, we consider a model with only one sensor and an estimator. To conserve energy the sensor can decide what and when to transmit. A similar model was considered in [7,8] when the underlying state is assumed to have a Gauss-Markov distribution and the estimation cost is quadratic. For that model it was shown that the optimal transmission is of a threshold type and the optimal estimation policy is similar to Kalman filtering. However, in many sensing applications the underlying state is not Gauss-Markov. One such example is sensing of environmental variables like temperature, rainfall, soil-moisture, etc. where the underlying state saturates after some level. For that reason, we instigate active sensing without assuming the underling process to be Gauss-Markov.

1.2. Model and Problem Formulation

Consider the sensing system in which a sensor observes the value of an environmental indicator (like temperature, soil moisture, rainfall, etc.) and communicates that to an estimator. Suppose that the environmental indicator evolves as a first-order time-homogeneous Markov chain $\{X_t,\ t = 1, \dots, T\}$, where X_t takes value a set $\mathcal{X}$. Assume that Markov chain starts at x_0 which is known to the sensor and the estimator. The sensor generates signals $\{Y_t,\ t = 1, \dots, T\}$, $Y_t \in \mathcal{Y}$, and sends them to the estimator. The signaling alphabet $\mathcal{Y}$ equals $\mathcal{X} \cup \{\mathfrak{b}\}$, where $\mathfrak{b}$ (called blank) indicates that the sensor did not transmit anything. Transmitting any symbol in $\mathcal{X}$ consumes p_* units of power, while transmitting a blank does not consume any power. That is, if the sensor signals $y \in \mathcal{Y}$, then it incurs a cost

$$p(y) = \begin{cases} p_*, & \text{if } y \in \mathcal{X}, \\ 0, & \text{if } y = \mathfrak{b}. \end{cases} \tag{1}$$

If the sensor transmits all its measurements (*i.e.*, set $Y_t = X_t$), then the estimator perfectly tracks the state X_t of the environmental indicator. This perfect tracking consumes p_* units of power per unit time. In some applications, power consumption at the sensor is more improtant than perfect reconstruction at the estimator. In such applications, the sensor may encode its observations in real-time as follows:

$$Y_t = f_t(X_{1:t}, Y_{1:t-1}). \tag{2}$$

Since transmitting blanks is free, the sensor saves power whenever a blank is transmitted.

The estimator generates an estimate $\hat{X}_t \in \mathcal{X}$ of the environmental indicator causally and in real-time as follows:

$$\hat{X}_t = g_t(Y_{1:t}) \tag{3}$$

The quality of estimation is determined by a distortion metric d on $\mathcal{X}$. The distortion at time t is given by $d(X_t, \hat{X}_t)$. Thus, it is possible to trade-off power consumed at the sensor with the distortion at the estimator. To quantify this trade-off, define the *cost* incurred at time t as

$$c(X_t, \hat{X}_t, Y_t) = p(Y_t) + \lambda d(X_t, \hat{X}_t)$$

where $\lambda > 0$ is a scaling factor.

The system operates for a horizon T. The collection $\mathbf{f} := (f_1, \dots, f_T)$ is called a *transmission policy* while the collection $\mathbf{g} := (g_1, \dots, g_T)$ is called an *estimation policy*. The performance of any policy $(\mathbf{f}, \mathbf{g})$ is given by the *expected cost*

$$J(\mathbf{f}, \mathbf{g}) := \mathbb{E}^{(\mathbf{f},\mathbf{g})}\Big[\sum_{t=1}^{T} p(Y_t) + \lambda d(X_t, \hat{X}_t) \Big] \qquad (4)$$

where the expectation is with respect to the joint measure induced on $\{(X_t, Y_t, \hat{X}_t), t = 1, \dots, T\}$ induced by the choice of the policy $(\mathbf{f}, \mathbf{g})$.

We are interested in the following optimization problem.

Problem 1 *Given the statistics of the Markov process $\{X_t, t = 1, \dots, T\}$, the transmission cost p_*, the distortion function d, and the scaling factor λ, pick a* causal policy $(\mathbf{f}, \mathbf{g})$ *of the form* (2) *and* (3) *that minimizes $J(\mathbf{f}, \mathbf{g})$ given by* (4).

2. STRUCTURE AND IMPLEMENTATION OF OPTIMAL POLICIES

2.1. Structure of optimal policies

The salient feature of the above model is that the data available at the sensor and the estimator is increasing with time. The complexity of storing and processing this data increases with time. We are interested in identifying a sufficient statistic that does not increase with time. Finding such a sufficient statistic is difficult for the following reason. The system is dynamic. So, not only should the sufficient statistic at time t be sufficient for the cost incurred at time t, it should also be sufficient to calculate the sufficient statistic at time $t+1$. Such a sufficient statistic is called an information state in stochastic control [9]. Normally, Markov decision theory is used to find such information states. However, Markov decision theory is restricted to systems with one decision maker, so it cannot be used directly in the above model which has two decision makers: the sensor and the encoder.

To circumvent these difficulties we use the framework developed in [10]. Using this framework we prove the the structure of optimal transmission and estimation policies in four stages. A brief proof outline is presented in Section 2.3.

Stage 1 The sensor may ignore the history of past observations and use a transmitting policy of the form

$$Y_t = f_t(X_t, Y_{1:t-1})$$

without any loss of optimality.

This stage removes part of the (time-increasing) data at the sensor. Even after that, the data at the sensor and the estimator is increasing with time.

Stage 2 The sensor and the estimator may restrict attention to policies that have the following structure without any loss of optimality.

1. The sensor either transmits its current observation or a blank. Formally, the transmission policy $\mathbf{f}$ is such that
$$\forall t, Y_t \in \{X_t, \mathfrak{b}\}$$
2. If the estimator receives a non-blank symbol, it chooses that symbol as its estimate. Formally, the estimation policy $\mathbf{g}$ is such that
$$\forall t, y_t \neq \mathfrak{b} \implies g(y_{1:t}) = y_t.$$

The second stage shows that causal real-time coding does not improve performance.

Stage 3 Let ΔX denote the space of probability distributions on $\mathcal{X}$. For any transmission policy $\mathbf{f}$ define $\Pi_t \in \Delta X$ as

$$\Pi_t(x) = \mathbb{P}(X_t = x \mid Y_{1:t-1}).$$

Π_t depends on the policy $\mathbf{f}$ only through $(f_1, \dots, f_{t-1})$. The sensor and the estimator may use Π_t as a sufficient statistic for $Y_{1:t-1}$ without any loss of optimality. Thus, using transmission and estimation policies of the form

$$Y_t = f_t(X_t, \Pi_t) \quad \text{and} \quad \hat{X}_t = g_t(Y_t, \Pi_t)$$

does not entail any loss of optimality.

The third stage shows that we may *compress* $Y_{1:t-1}$ into Π_t. Thus, the sensor and the estimator do not need to store $Y_{1:t-1}$ (which is increasing with time). They can store Π_t instead which takes values in a time invariant space. However, in practice, the space required for storing Π_t is much larger than the space required for storing $Y_{1:t-1}$. If $\mathcal{X}$ is finite, Π_t is a real-vector; if $\mathcal{X}$ is continuous, then Π_t is a real-valued function. Nonetheless, combining the result of Stage 3 with that of Stage 2 proves an efficient implementation of the optimal policy.

Stage 4 Let $\tau = \tau_t(Y_{1:t-1})$ denote the last time before t when the sensor transmitted a non-blank, i.e.,

$$\tau = \max\{s < t : Y_s \neq \mathfrak{b}\}$$

If all the Y_s are blank, then we set $\tau = 0$. The sensor and estimator may use $(X_\tau, t - \tau)$ as a sufficient statistic for Π_t without any loss of optimality. In other words, using transmission and estimation policies of the form

$$X_t = f_t(X_t, X_\tau, t - \tau) \quad \text{and} \quad \hat{X}_t = g_t(Y_t, X_\tau, t - \tau)$$

does not entail any loss of optimality.

2.2. Implementation of optimal policies

The policies given in Stage 4 can be implemented as follows. For a given t and $t - \tau$, the sensor needs to store the array of

Algorithm 1: Sensor Policy

let $n = 0$ and $x_\tau = x_0$
for every t **do**
  let $n = n + 1$
  if $(x_t, x_\tau) \in A(t, n)$ **then**
    transmit x_t
    let $x_\tau = x_t, n = 0$
  else
    transmit $\mathfrak{b}$

Algorithm 2: Estimator Policy

let $n = 0$ and $x_\tau = x_0$
for every t **do**
  let $n = n + 1$
  if $y_t = \mathfrak{b}$ **then**
    estimate $B(t, n)[x_\tau]$
  else
    estimate y_t
    let $x_\tau = y_t, n = 0$

indices

$$A(t, t-\tau) = \{x, x' \in \mathcal{X} : f_t(x, x', t-\tau) = x\}$$

and for each $x' \in \mathcal{X}$, the estimator needs to store

$$B(t, t-\tau)[x'] = g_t(\mathfrak{b}, x', t-\tau).$$

Using this stored information, the sensor and estimator operate as shown in Algorithms 1 and 2.

For infinite horizon problems, the arrays A and B only depend on $t - \tau$ and not on t. Thus, instead of storing a sequence of time varying arrays, we only need to store one 2D array (A and B, respectively) at the sensor and the estimator. In many applications, we can bound the maximum time between successive non-blank transmission, i.e., upper bound $t - \tau$. Suppose this bound is N. Then the memory required to store A and B is at most $N \cdot |\mathcal{S}|^2$.

2.3. Outline of the proof

We briefly outline the important steps of the proof. To prove Stage 1, we show the following.

Lemma 1 *Define* $R_t = (X_t, Y_{1:t-1})$. *For any policy* $(\mathbf{f}, \mathbf{g})$*:*

1. *The process* R_t *is a controlled Markov process with control action* Y_t*, i.e.,*

$$\mathbb{P}(R_{t+1} \mid R_{1:t}, Y_{1:t}) = \mathbb{P}(R_{t+1} \mid R_t, Y_t)$$

2. *The expected conditional cost given* $(R_{1:t}, Y_{1:t})$ *depends only on* (R_t, Y_t)*, i.e.,*

$$\mathbb{E}[c(X_t, \hat{X}_t, Y_t) \mid R_{1:t}, Y_{1:t}] = \mathbb{E}[c(X_t, \hat{X}_t, Y_t) \mid R_t, Y_t].$$

The proof of the above lemma follows by showing conditional independence between appropriate random variables.

Arbitrarily fix the estimation policy $\mathbf{g}$ and consider the optimal design of the transmission policy. Markov decision theory [9] and Lemma 1 imply the result of Stage 1.

The proof of the result of Stage 2 proceeds by backward induction and an interchange argument. Fix a time t. Assume that the policy from time $t + 1$ up to T has the structure of Stage 2, while the policy from time 1 up to t has the structure of Stage 1. We can then construct an alternative policy that has the structure of Stage 2 from t up to T and performs as well as the original policy. (The details of this construction are omitted due to space limitations).

Thus, starting with a policy that has the structure of Stage 1, we can proceed iteratively in a backward manner and construct an alternative policy that has the structure of Stage 2, but performs as well as the original policy. This construction implies the result of Stage 2.

To prove the result of Stage 3, we use the framework developed in [10] and consider the system from a point of view of a coordinator that observes the common data $Y_{1:t-1}$ available at both the sensor and the estimator. This coordinator chooses maps $\varphi : \mathcal{X} \mapsto \mathcal{Y}$ and $\gamma : \mathcal{Y} \mapsto \mathcal{X}$ that are used by the sensor and the estimator to select their actions (Y_t and $\hat{X}_t$, respectively) as a function of their *local* data (X_t and Y_t respectively). Thus, first the coordinator chooses the maps (φ_t, γ_t) using a coordination policy $\mathbf{h} = (h_1, \dots, h_T)$ as:

$$(\varphi_t, \gamma_t) = h_t(Y_{1:t-1})$$

and then the sensor and the estimator choose their actions as

$$Y_t = \varphi_t(X_t), \quad \hat{X}_t = \gamma_t(Y_t).$$

Mahajan *et al* [10] shows that this *coordinated system* is equivalent to the original system. Furthermore, the coordinated system is a centralized (single-agent) partially observed system. So, we can use the results from POMDPs (partially observable Markov decision processes) to find the structure of the optimal policies and translate that result back to the original model. These steps give the result of Stage 3.

Stage 4 follows from the fact that if we are using policies of the form of Stage 2, then $\Pi_t(x_t)$ depends on $Y_{1:t-1}$ through only $Y_{\tau:t-1}$. By definition $Y_\tau = X_\tau$ and $Y_s = \mathfrak{b}$, $s = \tau + 1, \dots, t - 1$. This yields the result of Stage 4.

3. DYNAMIC PROGRAMMING DECOMPOSITION

We cannot directly obtain a dynamic programming decomposition for the above model because it has two decision makers, the sensor and the estimator. Markov decision theory assume one decision maker. Nonetheless, we can use the framework of [10] and consider the system from the point of view of a coordinator (as described in Section 2.3). The coordinated system is a centralized partially observed system. So, we can use

Markov decision theory to obtain a dynamic programming decomposition to find an optimal coordinated policy and then translate that policy to the original system.

Using the above approach, we get the following:

Proposition 1 *Define a sequence $\{V_t\}$ of function recursively as follows:*

$$V_T(\Pi_T) = \min_{(\varphi_T, \gamma_T)} \Big\{ \mathbb{E}[c(X_T, \hat{X}_T, Y_T) \mid \Pi_T] \Big\}$$

and for $t = T-1, \dots, 1$

$$V_t(\Pi_t) = \min_{(\varphi_t, \gamma_t)} \Big\{ \mathbb{E}[c(X_t, \hat{X}_t, Y_t) + V_{t+1}(\Pi_{t+1}) \mid \Pi_t] \Big\}$$

where φ_t and γ_t are maps defined in Section 2.3. Let $h_t^*(\pi) = (\varphi_t^*, \gamma_t^*)$ denote the arg min *at time t when $\Pi_t = \pi$. Then, $\mathbf{h}^*$ is an optimal coordination policy. Furthermore,*

$$f_t^*(x_t, \pi_t) = \varphi_t^*(\pi_t)(x_t), \quad g_t^*(y_t, \pi_t) = \gamma_t^*(\pi_t)(y_t).$$

is an optimal transmission and estimation policy for the original system.

Although the above dynamic program sequentially decomposes the search of an optimal policy, it is not practical because at each step we have to solve a functional optimization problem (search the best (φ_t, γ_t)). However, we can exploit the structure of the optimal policy obtained in Stage 2 to further simplify the dynamic program.

The structure derived in Stage 2 imposes the following restriction on the maps φ_t and γ_t.

$$\forall t, \quad \varphi_t(x_t) \in \{x_t, \mathfrak{b}\} \quad \text{and} \quad y_t \neq \mathfrak{b} \implies \gamma_t(y_t) = y_t.$$

Thus, the map φ_t is equivalent to the *silence set* S_t defined as

$$S_t = \{x \in \mathcal{X} : \varphi_t(x) = \mathfrak{b}\}$$

and the map γ_t is equivalent to the estimate $\hat{x}_t^* = g_t(\mathfrak{b})$ chosen when a blank is received. Using, this equivalence, we can simplify the result of Proposition 1 as follows:

Proposition 2 *Define a sequence $\{V_t\}$ of function recursively as follows:*

$$V_T(\Pi_T) = \min_{(S_T, \hat{x}_T^*)} \Big\{ \mathbb{E}[c(X_T, \hat{X}_T, Y_T) \mid \Pi_T] \Big\} \tag{5}$$

and for $t = T-1, \dots, 1$

$$V_t(\Pi_t) = \min_{(S_t, \hat{x}_t)} \Big\{ \mathbb{E}[c(X_t, \hat{X}_t, Y_t) + V_{t+1}(\Pi_{t+1}) \mid \Pi_t] \Big\} \tag{6}$$

Let $h_t^*(\pi) = (S_t^*, \hat{x}_t^*)$ denote the arg min *at time t when $\Pi_t = \pi$. Then, $\mathbf{h}^*$ is an optimal coordination policy. Furthermore,*

$$f_t^*(x_t, \pi_t) = \begin{cases} \mathfrak{b}, & \text{if } x_t \in S_t^*(\pi_t) \\ x_t, & \text{otherwise;} \end{cases}$$

$$g_t^*(y_t, \pi_t) = \begin{cases} \hat{x}_t^*(\pi_t), & \text{if } y_t = \mathfrak{b}, \\ y_t, & \text{otherwise;} \end{cases}$$

is an optimal transmission and estimation policy for the original system.

In the above proposition, the expected instantaneous cost is given by

$$\mathbb{E}\Big[c(X_t, \hat{X}_t, Y_t) \mid \Pi_t, S_t, \hat{x}_t^*\Big] = p_* \cdot (|\mathcal{X}| - |S_t|) + \lambda \cdot \sum_{x \in S_t} d(x, \hat{x}_t^*) \cdot \Pi_t(x) / \Pi_t(S_t).$$

The above dynamic program can be solved using numerical methods for solving POMDPs [11, 12].

4. REFERENCES

[1] Q. Cao, T. Abdelzaher, T. He, and J. Stankovic, "Towards optimal sleep scheduling in sensor networks for rare-event detection," in *Proc. IEEE IPSN*, 2005.

[2] A. Fuemmeler and V. V. Veeravalli, "Smart sleeping policies for energy efficient tracking in sensor networks," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 2091–2101, May 2008.

[3] G. Atia, V. V. Veeravali, and J. A. Fuemmeler, "Sensor scheduling for energy-efficient target tracking in sensor networks," *submitted to IEEE Trans. Signal Process.*, Aug. 2010.

[4] W. Wu and A. Arapostathis, "Optimal sensor querying: General Markovian and LQG models with controlled observations," *IEEE Trans. Autom. Control*, vol. 53, no. 6, pp. 1392–1405, Jul. 2008.

[5] C. Rago, P. Willett, and Y. Bar-Shalom, "Censoring sensors: A low-communication rate scheme for distributed detection," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 2, pp. 554–568, Apr. 1996.

[6] S. Appadwedula, V. V. Veeravalli, and D. L. Jones, "Decentralized detection with censoring sensors," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1362–1373, Apr. 2008.

[7] O. C. Imer and T. Başar, "Optimal estimation with limited measurements," in *Proc. IEEE CDC/ECC*, 2005.

[8] G. M. Lipsa and N. Martins, "Remote state estimation with communication costs for first-order LTI systems," *IEEE Trans. Autom. Control*, Aug. 2011.

[9] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation Identification and Adaptive Control*. Prentice Hall, 1986.

[10] A. Nayyar, A. Mahajan, and D. Teneketzis, "Dynamic programming for multi-controller stochastic control with partial information sharing: A common-information approach," *submitted to the IEEE Trans. Autom. Control*, 2011.

[11] A. Cassandra, M. L. Littman, and N. L. Zhang, "Incremental pruning: A simple, fast, exact method for partially observable Markov decision processes," in *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, 1997.

[12] M. T. J. Spaan and N. Vlassis, "Perseus: Randomized point-based value iteration for POMDPs," *Journal of Artificial Intelligence Research*, vol. 24, pp. 195–220, 2005.