Mathematical Theory of Networks and Systems (MTNS) 2008

Dynamic Programming with Non-Classical Information Structures

John Swigart and Sanjay Lall

Abstract

We consider the problem of Partially Observed Markov Decision Processes with a non-classical information structure. Under a particular constraint on the information structure, optimal decision policies can be found via a dynamic programming approach. We also consider state space systems with linear dynamics and quadratic cost objectives, and provide sufficient conditions on the information structure under which optimal control policies can be found analytically.

1 Introduction

In the study of decentralized stochastic control, the search for analytical, optimal control policies is, more often than not, an intractable problem. Within the group of feasible problems, only a handful of solutions have yet been found. However, the goal for most control engineers is to find linear control policies which are either optimal or suboptimal. The Partially Observed Markov Decision Process (POMDP), and more specifically, the classical LQG case (Linear dynamics, Quadratic cost, Gaussian noise) with full information are perhaps the most universally known systems which have been shown to be tractable.

It has been shown that the ability to find optimal analytical solutions is significantly affected by the information structure of the system [8]. Much research has been done to classify certain non-classical information structures for which optimal feedback controllers can be efficiently solved [4, 6]. In this paper, we attempt to augment this set of tractable information structures to include those information structures satisfying a particular constraint.

In this paper, we first consider a specific information structure, shown to produce linear optimal solutions in the LQ case [7, 9], and extend the dynamic

programming approach to find optimal solutions of the POMDP formulation of the problem. We go on to provide sufficient conditions on an information structure under which any LQ system can be analytically solved via dynamic programming. Lastly, while [7] and [9] showed the existence of optimal controllers in the LQ case, it wasn't until [3] that the computational complexity of such controllers was considered. We finish our analysis by showing that an assumption made on the maximal delay of information propagation allows us to bound the storage required for performing estimation.

2 Notation and System Model

We will employ, where possible, the notation used in [2, 1] as follows. In the remainder of this paper, we use subscripts to denote the time index. Thus, x_t denotes the state of the system at time t. We denote the sequence of variables $x_0, x_1, \ldots, x_t$ by $x_{0:t}$. We also define $A_{0\cdots t}$ as the product of the variables corresponding to times $0, \ldots, t$; that is $A_{0\cdots t} = A_0 A_1 \ldots A_t$. For a set $\mathcal{X}$, we interpret $\mathcal{X}^n$ as the n-fold Cartesian product of the set, so that $\mathcal{X}^n = \mathcal{X} \times \ldots \times \mathcal{X}$ n-times, where $\mathcal{X}^0 = \emptyset$. Also, for any finite set Y, we denote 2^Y as the set of all subsets of Y. Lastly, we denote $\mathbb{Z}^+$ as the set of non-negative integers.

A Markov decision process provides a model for sequential decision making in a stochastic environment. The decisions, or control actions, made at each time step affect the future evolution of the system. The goal of the decision maker is to choose actions such that the system trajectory optimizes some objective function over a finite horizon N.

For the purposes of this paper, we consider decisions made at discrete times $t \in \{0, \ldots, N\}$. At each time t, the system occupies a state, and we denote the set of all states by $\mathcal{X}$. Similarly, at each time step, the decision maker chooses an action from the set of all possible actions, denoted $\mathcal{U}$. In the partially observed Markov decision process (POMDP), these decisions are based on an imperfect knowledge of the system state, and the set of all possible observations as seen by the observer is denoted $\mathcal{Y}$.

First, for a finite set $\mathcal{X}$ let $M_{\mathcal{X}}$ be the set of distributions on $\mathcal{X}$; that is, the set of functions $f : \mathcal{X} \to [0, 1]$ such that $\sum_{x \in \mathcal{X}} f(x) = 1$. Our formal definition of a POMDP is as follows.

Definition 1. *A partially observed Markov decision process (POMDP) is a tuple (A, C, g) where*

1. *A is a sequence $A_0, A_1, \ldots, A_N$, with $A_0 \in M_{\mathcal{X}}$, and for $t \geq 1$, we have $A_t : \mathcal{X} \times \mathcal{X} \times \mathcal{U} \to [0, 1]$, such that $A_t(\cdot, z, a) \in M_{\mathcal{X}}$ for all $z \in \mathcal{X}$, $a \in \mathcal{U}$.*

2. *C is a sequence $C_0, C_1, \ldots, C_{N-1}$, with $C_0 : \mathcal{Y} \times \mathcal{X} \to [0, 1]$, such that $C_0(\cdot, z) \in M_{\mathcal{Y}}$ for all $z \in \mathcal{X}$ and for $t \geq 1$, we have $C_t : \mathcal{Y} \times \mathcal{X}^{t+1} \times \mathcal{U} \to [0, 1]$, such that $C_t(\cdot, z, a) \in M_{\mathcal{Y}}$ for all $z \in \mathcal{X}^{t+1}, a \in \mathcal{U}$.*

3. *g is a sequence $g_0, g_1, \ldots, g_N$, with $g_t : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$.*

For the purposes of this paper, we are interested in controllers which make decisions based on some subset of the current and previous observation and decision variables. To this end, we define the information structure of a POMDP.

Definition 2. *For each $t \in \mathbb{Z}^+$, define the set of integers $Z_t = \{0, 1, \ldots, t\}$. The* information structure *of a POMDP is defined by (Y, U), where*

1. *Y is a sequence $Y_0, Y_1, \ldots$, such that $Y_t \in 2^{Z_t}$.*
2. *U is a sequence $U_0, U_1, \ldots$, such that $U_t \in 2^{Z_{t-1}}$.*

We can use this definition of the information structure to denote the information available to the controller at time t by

$$\eta_t = \{y_i, u_j \mid i \in Y_t, j \in U_t\} \tag{1}$$

where $y_i \in \mathcal{Y}$ and $u_j \in \mathcal{U}$ for all $i, j \in \mathbb{Z}^+$. We also define i_t to be a realization of η_t by $i_t = \{w_i, a_j \mid i \in Y_t, j \in U_t\}$.

The above information structure determines what observations and previous decisions are available to the decision maker at time t. Note that in the classical, full information case, we have $Y_t = Z_t$ and $U_t = Z_{t-1}$ for all $t \in \mathbb{Z}^+$. We can now define the POMDP policy which determines the actions to be taken at each time step.

Definition 3. *A POMDP policy is a sequence $K = (K_0, K_1, \ldots, K_{N-1})$, where $K_t : \mathcal{U} \times \mathcal{Y}^{|Y_t|} \times \mathcal{U}^{|U_t|} \to [0, 1]$ for all $t \in \mathbb{Z}^+$ such that $K_t(\cdot, z, a) \in M_{\mathcal{U}}$ for all $z \in \mathcal{Y}^{|Y_t|}, a \in \mathcal{U}^{|U_t|}$.*

Now, for any POMDP (A, C, g) and policy K, we define the state stochastic process $x_{0:N}$, the observation process $y_{0:N-1}$, and the action process $u_{0:N-1}$ by

$$\begin{aligned}&\text{Prob}(x_{0:t} = z_{0:t}, y_{0:t} = w_{0:t}, u_{0:t} = a_{0:t}) = A_0(z_0)C_0(w_0, z_0) \\ &\times \prod_{k=1}^{t} A_k(z_k, z_{k-1}, a_{k-1}) \prod_{k=1}^{t} C_k(w_k, z_{0:k}, a_{k-1}) \prod_{k=0}^{t} K_k(a_k, \{w_i, a_j \mid i \in Y_k, j \in U_k\})\end{aligned} \tag{2}$$

Lastly, for a given POMDP, the goal of our decision makers is to choose the policy K which minimizes the cost function over the finite horizon N, given by

$$J(K_{0:N-1}) = E\left(\sum_{t=0}^{N-1} g_t(x_t, u_t) + g_N(x_N)\right) \tag{3}$$

2.1 Temporal Skyline Information

We can now define our information structure of interest in this paper.

Definition 4. *We call the information structure in (1) a* temporal skyline information structure *(TS) if*

$$Y_{t-1} \subseteq Y_t \tag{4}$$

$$U_t = Z_{t-1} \tag{5}$$

is satisfied for all $t \in \mathbb{Z}^+$.

We also define, for each Y_t, the complementary set $Y_t^\perp = Z_t \setminus Y_t$. In other words, the TS structure can be viewed as a partially nested structure [4], where the decision maker at time t knows at least all of the information that the decision maker at time $t-1$ knew. We note the distinction, though, between TS structures and information structures which are simple delays. In particular, there is no queuing of information in a TS structure which is typical of most delay systems.

3 POMDP Optimization

We are interested in finding a control policy K which minimizes the cost in (3) for a given POMDP with a TS information structure. To lighten notation we define, for each $t = 0, 1, \ldots, N-1$, the function $P_t = A_{0\cdots t}C_{0\cdots t}$, with $P_N = A_N P_{N-1}$. Then, the joint probability $\text{Prob}(z_{0:t}, w_{0:t}, a_{0:t})$ from (2) can be written succinctly as $P_t K_{0\cdots t}$. Also, for a POMDP with a TS information structure, we define the function D_t by

$$D_t(i_t) = \sum_{\substack{z_{0:t} \\ Y_t^\perp}} P_t$$

where the summation over $Y_t^\perp$ means that we sum over the variables $\{w_i \mid i \in Y_t^\perp\}$. We now define the following functions in a recursive manner.

Definition 5. *Given a POMDP* (A, C, g) *with a TS information structure* (Y, U)*, we define the value function* V_t *recursively as follows. Let*

$$V_N(a_{N-1}, i_{N-1}) = \frac{1}{D_{N-1}(i_{N-1})} \sum_{\substack{z_{0:N} \\ Y_{N-1}^\perp}} g_N(z_N) P_N \tag{6}$$

and for $0 \le t \le N-1$*, we define*

$$V_t(i_t) = \min_{a_t} \frac{1}{D_t(i_t)} \sum_{\substack{z_{0:t+1} \\ Y_t^\perp}} (g_t + V_{t+1}) P_{t+1} \tag{7}$$

Definition 6. *Given a POMDP* (A, C, g) *with a TS information structure* (Y, U)*, we define the cost-to-go* Q_t *for all* $0 \le t \le N-1$ *to be*

$$Q_t(a_t, i_t) = \frac{1}{D_t(i_t)} \sum_{\substack{z_{0:t+1} \\ Y_t^\perp}} (g_t + V_{t+1}) P_{t+1} \tag{8}$$

Hence, we see that Q_t is the expected future cost from time t onwards, conditioned on the information i_t available to the decision maker at time t and the decision a_t.

Having established the structure of this system in the previous section, we can now state the first result of this paper.

Theorem 7. *Given a POMDP (A, C, g), with a TS information structure (Y, U), let the functions $Q_0, \ldots, Q_{N-1}$ be defined as in (8). Then the policy K is optimal if for each t, K_t is a minimizer for*

$$\min_{K_t} \sum_{a_t} K_t(a_t, i_t) Q_t(a_t, i_t)$$

Moreover, there exists a deterministic optimal K_t given by

$$K_t(a_t, i_t) = \begin{cases} 1 & a_t = \mu_t(i_t) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

where $Q_t(\mu_t(i_t), i_t) \leq Q_t(a_t, i_t)$ for all i_t.

Proof. We give an outline of the proof of Theorem 7 by explicit computation of the dynamic programming algorithm. To begin, we can express the optimal total cost as

$$J_{opt}(K_{0:N-1}) = \min_{K_{0:N-1}} \sum_{\substack{z_{0:N} \\ a_{0:N-1} \\ w_{0:N-1}}} \left(g_0(z_0, a_0) + \ldots + g_N(z_N)\right) P_N K_{0 \cdots N-1}$$

Using the TS structure of the Y_t and (6), we decompose the cost as

$$J_{opt}(K) = \min_{K_{0:N-1}} \sum_{\substack{a_{0:N-2} \\ Y_{N-2}}} K_{0 \cdots N-2} \sum_{\substack{a_{N-1} \\ Y_{N-1} \setminus Y_{N-2}}} K_{N-1} \sum_{\substack{z_{0:N} \\ Y_{N-1}^{\perp}}} \left(\sum_{t=0}^{N-1} g_t + V_N \right) P_N \tag{10}$$

where again we abbreviate summations over observation variables $\{w_i \mid i \in Y_t\}$ by summation over Y_t. This decomposition in (10) is the property which allows us to proceed in the standard POMDP dynamic programming approach. Noting that the denominator of Q_t is independent of a_t in (8), our cost becomes

$$J_{opt}(K) = \min_{K_{0:N-1}} \sum_{\substack{a_{0:N-2} \\ Y_{N-2}}} K_{0:N-2} \sum_{\substack{z_{0:N-1} \\ Y_{N-2}^{\perp}}} P_{N-1} \left(\sum_{t=0}^{N-2} g_t + \sum_{a_{N-1}} K_{N-1} Q_{N-1} \right)$$

Making use of a standard result from dynamic programming, this expression is equivalent to

$$J_{opt}(K) = \min_{K_{0:N-2}} \sum_{\substack{a_{0:N-2} \\ Y_{N-2}}} K_{0:N-2} \sum_{\substack{z_{0:N-1} \\ Y_{N-2}^{\perp}}} P_{N-1} \left(\sum_{t=0}^{N-2} g_t + \min_{K_{N-1}} \sum_{a_{N-1}} K_{N-1} Q_{N-1} \right)$$

As a result, an optimal, deterministic policy for K_{N-1} is obtained by minimizing $\sum_{a_{N-1}} K_{N-1}Q_{N-1}$ as in (9). Using (7), our optimal cost can now be expressed in a form similar to its initial form.

$$J_{opt}(K) = \min_{K_{0:N-2}} \sum_{\substack{a_{0:N-2} \\ Y_{N-2}}} K_{0:N-2} \sum_{\substack{z_0,\ldots,z_{N-1} \\ Y^{\perp}_{N-2}}} (g_0 + \ldots + g_{N-2} + V_{N-1})P_{N-1}$$

Using induction, and applying the above argument at each iteration, we arrive at the value function iteration in (7), where at each step in the iteration an optimal, deterministic K_t is found by minimizing $\sum_{a_t} K_t Q_t$. ∎

4 LQG Optimization

The LQG formulation is a special case of the POMDP. However, as it is a very common formulation for problems, either by design or by linearization of complex dynamical systems, we shall analyze the specific approach taken in this case.

Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, and $C \in \mathbb{R}^{p\times n}$. The setup is the standard LQG formulation:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t + w_t \\ y_t &= Cx_t + v_t \\ x_0 \sim N(\tilde{x}_0, \Sigma_0) \qquad w_t &\sim N(0, \Sigma_w) \qquad v_t \sim N(0, \Sigma_v) \end{aligned}$$

where x_0, w_t, v_t are IID Gaussian random variables. This of course is a POMDP on the infinite state space $\mathbb{R}^n$. Our goal is to choose control policies $\gamma_t : \eta_t \to \mathbb{R}^m$ for $t = 0, \ldots, N-1$ to minimize the following standard finite-horizon, quadratic cost function.

$$J(\gamma_{0:N-1}) = \int \left(\sum_{i=0}^{N-1} x_i^T Q x_i + u_i^T R u_i + x_N^T Q_f x_N \right) \text{Prob}(\mu)\, d\mu. \tag{11}$$

where $Q, Q_f \in \mathbb{R}^{n\times n}$ are symmetric, positive semidefinite and $R \in \mathbb{R}^{m\times m}$ is symmetric, positive definite, and $\mu = (x_{0:t}, y_{0:t-1}, u_{0:t-1})$.

When optimizing the control policy for the cost in (11), we must deal with the probability distribution of the state x_t conditioned on the information η_t available to the decision maker at each time step t. In a manner equivalent to our work in section 2, we can break up the joint pdf of all states, observations, and decisions using conditional probabilities and the IID nature of the noise. Hence, we rewrite the large joint pdf as a series of smaller conditional probability distributions, as follows.

$$\text{Prob}\left(x_{0:t}, y_{0:t-1}, u_{0:t-1}\right) = A_{0\cdots t} C_{0\cdots t-1} K_{0\cdots t-1}$$

where we have the conditional probability distributions defined in section 2.

$$\begin{aligned} A_i\left(x_i, x_{i-1}, u_{i-1}\right) &= \text{Prob}\left(x_i | x_{i-1}, u_{i-1}\right) = f_N(x_i - Ax_{i-1} - Bu_{i-1}, \Sigma_w) \\ C_i\left(y_i, x_i\right) &= \text{Prob}\left(y_i | x_i\right) = f_N(y_i - Cx_i, \Sigma_v) \\ K_i\left(u_i, \eta_{i-1}\right) &= \text{Prob}\left(u_i | \eta_{i-1}\right) \end{aligned}$$

Both A_i and C_i simply represent the Gaussian pdfs for w_i and v_i, respectively, where $f_N(\mu, \Sigma)$ denotes the normal distribution function for $N(\mu, \Sigma)$. K_i now represents the decision policies, expressed as a conditional pdf, though we restrict our policies to be deterministic.

Having previously shown in section 3 that a dynamic programming approach can be applied to POMDPs with a TS structure, it comes as no surprise that we can construct a dynamic programming algorithm for the LQG system with TS structure. However, in doing so, we arrive at one of the main results of this paper.

We first define the familiar Riccati recursion for discrete time systems.

Definition 8. *We define the matrices $P_0, \ldots, P_N$ in the following recursive fashion. Let $P_N = Q_f$, and for all $1 \leq t \leq N$, let*

$$P_{t-1} = Q + A^T P_t A - A^T P_t B \left(R + B^T P_t B\right)^{-1} B^T P_t A \tag{12}$$

Further define the real numbers $s_0, \ldots, s_N$ as follows. Let $s_N = 0$, and for all $1 \leq t \leq N$, let $s_{t-1} = s_t + \operatorname{trace}(P_t \Sigma_w) + \operatorname{trace}((Q + A^T P_t A - P_{t-1})\Sigma_{t-1})$.

Using these definitions, the following result provides a sufficient condition on the information structure of a LQG problem which allows us to find an optimal control policy.

Theorem 9. *Given a LQG system with any information structure (Y, U), suppose there exists matrices $\Sigma_0, \ldots, \Sigma_{N-1}$ such that, for all $t \in \{0, 1, \ldots, N-1\}$, and for all $\eta_0, \ldots, \eta_{N-1}$,*

$$cov\left(x_t | \eta_t\right) = \Sigma_t. \tag{13}$$

Then, the policy $\gamma = (\gamma_0, \ldots, \gamma_{N-1})$ defined by

$$\gamma_t(\eta_t) = -\left(R + B^T P_{t+1} B\right)^{-1} B^T P_{t+1} A \hat{x}_t(\eta_t)$$

is optimal, where $\hat{x}_t(\eta_t) = E\left(x_t | \eta_t\right)$, and P_t is given by (12).

Proof. We can show that the conditions of Theorem 9 are sufficient to find an optimal policy by explicit computation of the dynamic programming algorithm.

Firstly we use the probability distribution of the state transition to write x_N in terms of x_{N-1} and u_{N-1}, as follows:

$$J(\gamma_{0:N-1}) = \int \left(M_{N-2} + \begin{bmatrix} x_{N-1} \\ u_{N-1} \end{bmatrix}^T \begin{bmatrix} Q + A^T Q_f A & A^T Q_f B \\ B^T Q_f A & R + B^T Q_f B \end{bmatrix} \begin{bmatrix} x_{N-1} \\ u_{N-1} \end{bmatrix} \right) \times \operatorname{Prob}(\mu)\, d\mu + \operatorname{trace}(Q_f \Sigma_w)$$

where we've defined, for convenience

$$M_t = \sum_{i=0}^{t} x_i^T Q x_i + u_i^T R u_i$$

for each $0 \leq t \leq N-1$. Like before, $\text{Prob}(\mu)$ is simply a placeholder for the joint probability distribution of all states, observations, and actions. Next, we substitute in the decision policy for γ_{N-1} and use a perturbation $h\Gamma(\eta_{N-1})$ on the policy, with h sufficiently small, to optimize the cost over that decision.

$$\begin{aligned} J(\gamma_{0:N-2}, \gamma_{N-1} + h\Gamma) = \int \Bigg(M_{N-2} + \begin{bmatrix} x_{N-1} \\ \gamma_{N-1} + h\Gamma \end{bmatrix}^T \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \begin{bmatrix} x_{N-1} \\ \gamma_{N-1} + h\Gamma \end{bmatrix} \Bigg) \\ \times \text{Prob}(\mu)\, d\mu + \text{trace}(Q_f \Sigma_w) \\ = J(\gamma) + 2h \int \Gamma^T \left(S_{12}^T x_{N-1} + S_{22} \gamma_{N-1} \right) \text{Prob}(x_{N-1}, \eta_{N-1})\, dx_{N-1} d\eta_{N-1} + O(h^2) \end{aligned} \tag{14}$$

where

$$\begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} = \begin{bmatrix} Q + A^T Q_f A & A^T Q_f B \\ B^T Q_f A & R + B^T Q_f B \end{bmatrix}$$

Since $\gamma_{N-1}(\eta_{N-1})$ and $\Gamma(\eta_{N-1})$ are not functions of x_{N-1}, we can perform the integration over x_{N-1} in the first order perturbation term of (14) to show

$$\begin{aligned} & J(\gamma_{0:N-2}, \gamma_{N-1} + h\Gamma(\eta_{N-1})) \\ & = J(\gamma) + 2h \int \Gamma^T \Bigg(\int \left(S_{12}^T x_{N-1} + S_{22} \gamma_{N-1} \right) \text{Prob}(x_{N-1} | \eta_{N-1})\, dx_{N-1} \Bigg) \\ & \qquad \times \text{Prob}(\eta_{N-1})\, d\eta_{N-1} + O(h^2) \\ & = J(\gamma) + 2h \int \Gamma^T \left(S_{12}^T E\left(x_{N-1} | \eta_{N-1} \right) + S_{22} \gamma_{N-1} \right) \text{Prob}(\eta_{N-1})\, d\eta_{N-1} + O(h^2) \end{aligned} \tag{15}$$

Setting the first order perturbation of the cost in (15) to zero, we see that an optimal policy for γ_{N-1} is given by

$$\begin{aligned} \gamma_{N-1} &= -S_{22}^{-1} S_{12}^T E\left(x_{N-1} | \eta_{N-1} \right) \\ &= -\left(R + B^T Q_f B \right)^{-1} B^T Q_f A E\left(x_{N-1} | \eta_{N-1} \right) \\ &= F_{N-1} \hat{x}_{N-1} \end{aligned}$$

where $\hat{x}_{N-1}$ is the belief state of the system. Note, we haven't made any assumptions on the structure of Y_{N-1} or U_{N-1} up to this point. This result for γ_{N-1} is a product of the LQ formulation of the basic problem and holds for any information structure.

Now, substitute γ_{N-1} back into the cost function.

$$\begin{aligned} J = & \int M_{N-2}\, \text{Prob}(\mu)\, d\mu + \text{trace}(Q_f \Sigma_w) \\ & + \int \Bigg(\begin{bmatrix} x_{N-1} \\ F_{N-1} \hat{x}_{N-1} \end{bmatrix}^T \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix} \begin{bmatrix} x_{N-1} \\ F_{N-1} \hat{x}_{N-1} \end{bmatrix} \text{Prob}(x_{N-1}, \eta_{N-1}) \Bigg) dx_{N-1} d\eta_{N-1} \end{aligned}$$

Take the expectation of x_{N-1} conditioned on η_{N-1} in the last term.

$$J = \int M_{N-2} \operatorname{Prob}(\mu)\, d\mu + \operatorname{trace}(Q_f \Sigma_w) + \int \left(\hat{x}_{N-1}^T P_{N-1} \hat{x}_{N-1} + \operatorname{trace}(S_{11} \Sigma_{N-1}) \right) \operatorname{Prob}(\eta_{N-1})\, d\eta_{N-1}$$

where

$$\begin{aligned} P_{N-1} &= S_{11} + S_{12} F_{N-1} + F_{N-1}^T S_{12}^T + F_{N-1}^T S_{22} F_{N-1} \\ &= Q + A^T Q_f A - A^T Q_f B \left(R + B^T Q_f B \right)^{-1} B^T Q_f A \end{aligned}$$

satisfies the Riccati recursion in (12). We then make use of the identity $E\left(x^T D x\right) = \hat{x}^T D \hat{x} + \operatorname{trace}(D \operatorname{cov}(x))$ to give

$$J = \int M_{N-2} \operatorname{Prob}(\mu)\, d\mu + \operatorname{trace}(Q_f \Sigma_w) + \int \left(x_{N-1}^T P_{N-1} x_{N-1} + \operatorname{trace}\left((S_{11} - P_{N-1}) \Sigma_{N-1} \right) \right) \operatorname{Prob}(x_{N-1}, \eta_{N-1})\, dx_{N-1} d\eta_{N-1}$$

Now, we must make our first restriction on the information structure. By assumption, the covariance matrix, Σ_{N-1}, is independent of η_{N-1}. Hence, it can be taken out of the integration as we proceed to the next iteration. As a result, we have returned to the same form in which we started this iteration.

$$J = \int \left(\sum_{i=0}^{N-2} x_i^T Q x_i^T + u_i^T R u_i + x_{N-1}^T P_{N-1} x_{N-1} \right) \operatorname{Prob}(\mu)\, d\mu + s_{N-1}$$

where s_{N-1} is as specified in Definition 8.

Thus, by induction, by applying the above arguments at each iteration, we see that the controller gains can be determined by the following dynamic program:

$$\begin{aligned} F_t &= -\left(R + B^T P_{t+1} B \right)^{-1} B^T P_{t+1} A \\ \gamma_t(\eta_t) &= F_t \hat{x}_t(\eta_t) \end{aligned}$$

with P_t defined in (12). Moreover, the optimal cost is $J = s_0$. ∎

Interestingly, this is the same dynamic program as the classical full information case. The difference in the resulting controllers is purely a result of the state estimation $\hat{x}_t$ being conditioned on less information than the classical case. Also, it is important to note that the control system has separated into a simple controller gain and an estimator, and that the policy is linear in the state estimate.

Using the conditional probability notation defined above, the state conditional probability distributions can be expressed by the following fractional series of pdfs.

$$\operatorname{Prob}(x_t | \eta_t) = \frac{\operatorname{Prob}(x_t, \eta_t)}{\operatorname{Prob}(\eta_t)} = \frac{\sum_{\substack{x_{0:t-1} \\ Y_t^\perp}} A_{0\cdots t} C_{0\cdots t} K_{0\cdots t-1}}{\sum_{\substack{x_{0:t} \\ Y_t^\perp}} A_{0\cdots t} C_{0\cdots t} K_{0\cdots t-1}} \tag{16}$$

When the system has a TS information structure, we can simplify (16), using the following lemma.

Lemma 10. *Given a LQG system with a TS information structure, there exists matrices $\Sigma_0, \ldots, \Sigma_{N-1}$ such that, for all $t \in \{0, 1, \ldots, N-1\}$, and for all $\eta_0, \ldots, \eta_{N-1}$ we have cov $(x_t|\eta_t) = \Sigma_t$.*

Proof. An immediate consequence of (4) is that $Y_t^{\perp} \cap Y_i = \emptyset$ for all $i \leq t$. As a result, since the policies $K_0, \cdots, K_{t-1}$ are independent of $Y_t^{\perp}$, they can be pulled out of the summations in (16) and canceled, leaving the following expression for the conditional probability distribution.

$$\text{Prob}\,(x_t|\eta_t) = \frac{1}{D_t(\eta_t)} \sum_{\substack{x_{0:t-1} \\ Y_t^{\perp}}} P_t \tag{17}$$

Hence, we have eliminated the K_i from the conditional probability distribution. It is clear from the definitions of P_t and D_t that (17) is a convolution of the Gaussian distributions $A_0, \ldots, A_t$ and $C_0, \ldots, C_t$, normalized by the denominator D_t. Since each A_i and C_i are normal distributions with fixed covariances, the resulting convolution in (17) is a Gaussian distribution, whose covariance Σ_t is also fixed. ■

Corollary 11. *Given a LQG system with a TS information structure, an optimal policy can be found by the algorithm of Theorem 9.*

It is important to note that Corollary 11 guarantees that the TS structure considered previously, and as a result any delayed system with perfect memory, can be solved via the dynamic programming approach constructed in Theorem 9.

5 Estimator

As noted above, the control process is separable in the TS case, and linear in the state estimate. In the full information POMDP problem, we have the well-known Kalman filter estimator with the following recursive form.

Definition 12. *Given a POMDP (A, C, g) with a full information structure, define the following probability distribution functions.*

$$p_{0|0} = \frac{P_0}{\sum\limits_{x_0} P_0} \qquad p_{t|t} = \frac{\sum\limits_{x_t} A_{t+1} C_{t+1} p_{t-1|t-1}}{\sum\limits_{x_t, x_{t+1}} A_{t+1} C_{t+1} p_{t-1|t-1}} \text{ for } t \geq 1 \tag{18}$$

In the framework of our above discussion, this nice form for the full information Kalman filter is a consequence of the fact that $Y_t^{\perp} = \emptyset$ at every time step t. This

can be easily seen when we take a closer look at the state conditional probability distribution used in (17). In our more general case however, it's possible that we are not updating our estimate with recent observations, but with old observations that were not previously available. In such a case, we cannot achieve the nice relationship between $p_{t|t}$ and $p_{t-1|t-1}$. Instead, we make another restriction, that all observations become available after some specified time. In order to formalize this, we specify the following assumption.

Assumption 13. *There exists an interval k, such that for all $t \in \mathbb{Z}^+$, we have $i \in Y_t$ for all $i \leq t-k$.*

Theorem 14. *Given a POMDP (A, C, g) with a TS information structure (Y, U) which satisfies Assumption 13, the probability distribution of the belief state $q_{t|t} = \text{Prob}(x_t|\eta_t)$ can be computed, for all $t \geq k$ by*

$$q_{t|t} = \frac{\sum\limits_{\substack{x_{t-k:t-1} \\ Y_t^\perp}} A_{t-k+1\cdots t} C_{t-k+1\cdots t} p_{t-k|t-k}}{\sum\limits_{\substack{x_{t-k:t} \\ Y_t^\perp}} A_{t-k+1\cdots t} C_{t-k+1\cdots t} p_{t-k|t-k}}$$

using the recursion on $p_{t|t}$ defined in (18).

Proof. This result follows directly from the definitions of $q_{t|t}$ in (17) and $p_{t-k|t-k}$ in (18). ∎

In the LQG framework, $p_{t-k|t-k}$ is a Gaussian pdf, which can be completely specified by its mean and covariance. As a result, we can bound the storage size required to compute $q_{t|t}$. Whereas remembering every y_i would require unlimited memory in the steady state, we have reduced the memory required to simply being, at worst, the size of k observations and decisions plus the mean and covariance of $p_{t-k|t-k}$.

6 Conclusions

We showed that for a partially observed Markov decision process with a non-classical temporal skyline information structure that optimal decision policies could be found via a dynamic programming approach. The key here was the subset constraint for observations available to the controller at successive time steps. In the LQ framework we provided sufficient conditions for an information structure under which a dynamic programming solution could be constructed to analytically find optimal decision policies. It is worth noting here that we have demonstrated only one possible method for finding these optimal policies. As with the classical full information structure, there exist other approaches for finding optimal policies, such as spectral factorization or augmentation of the state space. Lastly, we also showed that a simple maximal delay constraint on the information structure allowed us to bound the computation required to perform the state estimation at each time step.

Bibliography

[1] S. Adlakha, S. Lall, and A. Goldsmith. Information state for Markov decision processes with network delays. *Submitted to Proceedings of the IEEE Conference on Decision and Control*, 2008.

[2] S. Adlakha, R. Madan, S. Lall, and A. Goldsmith. Optimal control of distributed Markov decision processes with network delays. *Proceedings of the IEEE Conference on Decision and Control*, 2007.

[3] G. Casalino, F. Davoli, R. Minciardi, P. P. Puliafito, and R. Zoppoli. Partially nested information structures with a common past. *IEEE Transactions on Automatic Control*, 29(9):846–850, 1984.

[4] Y-C. Ho and K. C. Chu. Team decision theory and information structures in optimal control problems – Part I. *IEEE Transactions on Automatic Control*, 17(1):15–22, 1972.

[5] R. Radner. Team decision problems. *Annals of mathematical statistics*, 33:857–881, 1962.

[6] M. Rotkowitz and S. Lall. Decentralized control information structures preserved under feedback. *Proceedings of the IEEE Conference on Decision and Control*, pages 569–575, 2002.

[7] N. Sandell, Jr. and M. Athans. Solution of some nonclassical LQG stochastic decision problems. *IEEE Transactions on Automatic Control*, 19(2):108–116, 1974.

[8] H. S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal of Control*, 6(1):131–147, 1968.

[9] T. Yoshikawa. Dynamic programming approach to decentralized stochastic control problems. *IEEE Transactions on Automatic Control*, pages 796–797, 1975.