2015 IEEE Conference on Control Applications (CCA)
Part of 2015 IEEE Multi-Conference on Systems and Control
September 21-23, 2015. Sydney, Australia

Optimal control and optimal sensor activation for Markov decision problems with costly observations

Rene K. Boel[1] and Jan H. van Schuppen[2‡]

Abstract

This paper considers partial observation Markov decision processes. Besides the classical control decisions influencing the transition probabilities of the Markov process, we also consider control actions that can activate the sensors to provide more or less accurate information about the system state, explicitly including the cost of activating sensors. We synthesize control laws that minimize a discounted operating cost of the system over an infinite interval of time, where the instantaneous cost function depends on the current state, the control influencing the transition probabilities, and the control actions activating the sensors. A general computationally efficient optimal solution for this problem is not known. Hence we design supoptimal controllers that only use knowledge of the value function for the full state information Markov decision problem. Our solution guarantees that the discounted cost of operating the plant increases only by a bounded amount with respect to the minimal cost for the full state information problem. A new concept of pinned conditional distributions of the state given the observed history of the plant is required in order to implement these control laws online.

Keywords:
POMDP, active sensor control, stochastic control, suboptimal feedback control, partial information control

1. Introduction

Feedback controllers for stochastic systems require that the effect of current decisions on future expected behaviour is properly taken into account. An optimal balance must be found between the current investment in control action and the reduction this action achieves for the long term cost of operation of the plant. In classical Markov decision processes (MDP) the control value u_t is selected, as a function $\mathfrak{u}_t(X_t)$ of the current state $X_t \in \mathbb{X}$ of the Markov process, so as to minimize some cost function (in this paper a discounted cost over an infinite time interval). If the instantaneous cost at each instant t is given by $c(X_t, u_t)$ the optimal control law is given by $\mathfrak{u}^*(x) = argmin_u(c(x,u) + E_u(V(X_{t+1}) \mid X_t = x, u(t) = u))$ where the value function $V(x)$ is obtained by solving the Hamilton-Jacobi-Bellman (HJB) equation. Dynamic programing (DP) properly takes into account all the future effects of the current control decision, finding the optimal balance between instantaneous cost increase and future cost reduction. This paper considers partial observation Markov decision problems (POMDP) where the current state is not completely known, and therefore the optimal control law $\mathfrak{u}^*(X_t)$ cannot be applied.

Usually sensor activation is not considered as part of the control decision. However the increased use of networked control systems, often with battery operated sensors, at remote locations incurring communication costs, forces the control loop designer to explicitly consider the sensor activation decisions. In this paper we explicitly quantify the effect and the cost of activating the sensors. We distinguish on the one hand the classical actuator control u_a indicating how strongly the actuator pushes the state in a desired direction, and on the other hand the control value u_s indicating how much power is expended in order to improve the accuracy of the observations. The actuator control finds an optimal compromise between the instantaneous cost increase due to making an expensive decision u_a, and the reduction of the expected future cost that this decision induces. Selecting an expensive sensor activation control value u_s can reduce the uncertainty about the next state, which allows a lower future cost of operating the plant by avoiding wrong decisions due to wrong state estimates. The value u_s must be selected so as to find an optimal trade-off between the instantaneous sensor

**This work was not supported by any organization

†[1]R. Boel is with SYSTeMS Research Group, Ghent University, B-9052 Zwijnaarde, Belgium rene.boel at ugent.be

‡[2]J. van Schuppen is with Van Schuppen Control Research, Gouden Leeuw 143, 1103 KB Amsterdam, The Netherlands jan.h.van.schuppen at xs4all.nl

activation cost and the expected future cost reduction it can achieve.

This sensor activation control problem only makes sense for the partial information case, for which the DP approach is often not practically feasible due to the curse of dimensionality. For these partially observed Markov decision problems (POMDP) the conditional distribution $\pi_t(.) = \mathscr{P}(X_t = . \mid \mathscr{H}_t)$ of the state, given the available information $\mathscr{H}_t$ at time t, plays the role of the state at time t : π_t is a Markov process, albeit with a much larger state space; the HJB equation now defines the value function $V_{part}(\pi_t)$, a much more complicated object than $V(x)$, and stationary optimal control laws $\mathfrak{u}_{\mathfrak{part}}{}^*(\pi)$. In practice $V_{part}(\pi_t)$ and $\mathfrak{u}_{\mathfrak{part}}{}^*(\pi)$ can almost never be calculated, and heuristic approximations are used for these control design problems. In this paper we develop a novel control synthesis strategy for this problem, allowing consideration of the sensor activation cost, while requiring only knowledge of the easily obtained value function $V(x)$ for the full state information DP. Provided certain inequalities are satisfied our partial information control laws guarantee a bounded increase in cost of the plant operation as compared to the full information case. Note that this paper does not provide any bounds on the cost increase with respect to the partial information cost $V_{part}(\pi_t)$ (which of course always is larger than $V(X_t)$).

The partial observation control problem has a simple (almost) analytical solution for linear systems with quadratic cost and Gaussian noise, the LQG problem, extended with explicit cost $C(u_s)$ for sensor activation control value u_s. The separation theorem still holds (see **?**, **?**). The HJB equation is then decomposed into 2 independent minimizations, one involving $u_{a,t}$ (with optimal value $-K.\hat{X}_t$ where K is as in the classical LQ problem), the other involving $tr(Q_{u_{s,t}}.P)$ (where P defines the quadratic cost of the linear regulator problem). Actuator and sensor activation control problems are completely separated. The conditional distribution π_t is Gaussian, with mean $\hat{x}_t$ and error covariance matrix $Q_{u_s}(t)$. The positive definite matrix $Q_{u_{s,t}}$, calculated by the deterministic Riccati equation of the Kalman filter, is independent of $u_a(\tau), \tau \leq t$ thanks to the fact that for linear systems the increment $d\hat{x}_t$ is uncorrelated with the past evolution of the observations, and uncorrelated implies independent for Gaussian random processes. As soon as the assumptions of linearity or of Gaussian noise are dropped, this separation is no longer true. The problems of selecting $u_a(t)$ and $u_s(t)$ then become tightly coupled, and the corresponding POMDP becomes computationally intractable.

In order to avoid technical details in the analysis we only consider here countable-state controlled Markov chains $X_t, t \in \mathbb{Z}$, with transition probabilities $\Pi(x \to x'; u_a)) = \mathscr{P}(X_{t+1} = x' \mid X_t = x, u_{a,t} = u_a)$, and with sensors modeled by the probability distribution $Q(y \mid x, x', v) = \mathscr{P}(Y_{t+1} = y \mid X_t = x, X_{t+1} = x', u_{s,t} = u_s)$ of the observed values Y_t as function of the evolution of the state. The bounded instantaneous cost $c(X_t, u_{a,t}, u_{s,t}) \leq c_M$ depends not only on the state X_t and the control value $u_{a,t}$, but also on a control value $u_{s,t}$ that selects how the sensors are operated and activated. The actuator control and the sensor activation decisions depend at each time t on the currrently available information $\mathscr{H}_t$, which remembers all the past observations $Y_{\tau \leq t}$. The values of $u_a(t)$ and $u_s(t)$ are selected so as to minimize the discounted cost over an infinite time horizon, knowing $\mathscr{H}_t$.

The idea behind the control law proposed in this paper is as follows. In classical Markov decision problems (MDP) the optimal (stationary Markov) control law $\mathfrak{u}^*(x) = argmin_u H(x,u)$ with $H(x,u) = c(x,u) + E_u(V(X_{t+1}) - V(x) \mid X_t = x, u_t = u)$ achieves $H(x, \mathfrak{u}^*(x)) = 0$ (of course in the full state information case the optimal choice for the sensor activation part u_s in $u = (u_a, u_s) = (u_a, \emptyset)$ is trivial). The classical DP proof actually also shows that if a control law $\mathfrak{u}^\delta$ can be selected so that $\forall x : H(x, \mathfrak{u}^\delta(x)) \leq \delta$, then the discounted cost (with discounting factor γ) when using $\mathfrak{u}^\delta$ will be at most $\delta/(1-\gamma)$ higher than the minimal expected future cost $V(X_t)$ achieved by $\mathfrak{u}^*(X_t)$. We propose to use $\mathfrak{u}^\delta$ as suboptimal control laws for the partial observation problem with costly sensor activation. While this bound may in practice not be very tight it does at least provide a guarantee that the system will not become unstable. This approach should be compared to some of the proofs of stability for MPC (see e.g. **?**).

The proposed control design method requires the knowledge of the value function $V(x)$ for the full state Markov decision problem, which is often easy to calculate numerically. Notice though that the δ-approximation of the HJB equation must be verified for each value x of the state. The conditioning in the inequality requires that we calculate, for each value $x \in \mathbb{X}$, the *pinned* conditional distribution $\pi_{x,t}(\omega)$, representing $\pi_t(\omega)$ restricted to the subset $\Omega_{x,t} = \{\omega \in \Omega : X_t(\omega) = x\}$ of the complete probability space Ω (ω describes all the random variables influencing system dynamics and sensors). Section 4 describes how this can be achieved.

The reason that in the DP approach an optimal control law is found by optimizing the HJB equation over one single time step is that the value funtion $V(x)$ correctly quantifies the expected future cost given that $X_t = x$. For the sensor activation control problem knowledge of $V_{part}(\pi)$ would be required in order to correctly quantify by how much the improvement in the

accuracy of value for π_{t+1} would reduce future costs after time $t+1$. This cost reduction should then be compared to the cost of the sensor activation $u_{s,t}$ at time t. The method proposed in this paper avoids the need for calculating $V_{part}(\pi)$. The long term effect of a more accurate conditional distribution π_{t+1} is approximately quantified by looking ahead over a time window $[t+1, t+2, \ldots, t+H]$ using control laws u_τ dependent on $\pi_\tau, \tau \in [t+1, t+2, \ldots, t+H]$, with final cost dependent on the conditional distribution of $V(X_{t+H})$. This approximates the future performance improvement in the same way that MPC controllers approximate the effect of current control actions.

This paper explains the proposed approach first by a motivating example in section 2, considering a simple 2-state machine repair problem. Section 3 reviews the classical DP results used in this paper, for discounted cost problems. The proposed method for designing suboptimal controllers is explained in detail in section 4. Section 5 of this paper discusses a few possible applications - queueing systems and jump Markov LQG - of our control synthesis method where an optimal control law, and hence $V(x)$, is known for the full state information.

2. Motivating two-state example

In order to introduce the concepts of costly observations control and pinned conditional distributions we consider as a very simple example a two-state Markov chain representing a machine that can be either in the good state $X_t = 0$, or in the bad state $X_t = 1$. At each time t the transition probability from good to bad state is ρ; once in the bad state the system remains in the bad state until a repair action is carried out. The control agent selects to carry out a repair, at a cost r per repair, at successive times $T_n, n = 1, \ldots$ (i.e. $u_{a,t} = 0, t \neq T_n, u_{a,t} = 1$ otherwise). If no repair is carried out and the system state remains unchanged, if $u_{a,T_n} = 1$ the machine returns immediately to (or remains in) the good state $X_{t+1} = 0$. In the good state $X_t = 0$ the machine produces with probability $1-\lambda$ one good item per time unit, generating the observable output $Y_t = 0$, and no cost; if $X_t = 0$ then at time t the machine produces with probability λ a defective item, with output $Y_t = 1$ (due to a fault that may have nothing to do with the state of the machine); in the bad state the machine always produces a defective item, and $Y_t = 1$. Each defective item causes a cost of 1 unit. The control agent selects the repair times so as to minimize the discounted cost $J_\gamma(u) = E_u[\Sigma_{t=1}^{\infty} \gamma^t . Y_t + r.\Sigma_{n=1}^{\infty} \gamma^{T_n}]$.

The optimal control law is trivial if the agent knows the current state. The machine operation is a renewal process, with the machine starting in the good state $X_{T_n} = 0$ upon the n-th repair at time T_n, and entering the bad state $X_{T_{n+1}} = 1$ at $T_{n+1} > T_n$ (after a geometrically distributed time interval); if no repair were carried out at T_{n+1} defective items will be produced forever at a discounted future cost $1/(1-\gamma)$. If $1/(1-\gamma) < r$ it never is useful to repair the machine, $\mathfrak{u}_a(x) = 0, x = 0, 1$. If $1/(1-\gamma) > r$ it is obviously optimal to repair at time T_{n+1}, i.e. $\mathfrak{u_a}^*(0) = 0, \mathfrak{u_a}^*(1) = 1$ (delaying the repair by one time step would cause a cost $1 + \gamma.r > r$ if $1/(1-\gamma) > r$). The minimal cost, and the value function, can be calculated by an easy renewal argument: $V(0) = \frac{\lambda}{(1-\gamma)} + (r+1).\frac{\rho.\gamma.(1-\rho)}{1-\gamma+\gamma.\rho^2}$; adding the cost of one defective item and a repair, in state $X_t = 1$ gives $V(1) = \gamma.V(0) + r + 1$.

Unfortunately in practice the agent cannot detect the transition from good to bad state accurately and immediately since a defective item (and the corresponding observation $Y_t = 1$) can also be produced in the good state. The conditional probability $\pi_t = \mathscr{P}(X_t = 1 \mid Y_0, Y_1, \ldots, Y_t)$, of being in a bad state at time t, starting with an initial condition $\pi_0 = 0$, can be recursively calculated by Bayes' rule:

$$\pi_{t+1} = \quad 0 \qquad \text{if} \quad Y_{t+1} = 0 \quad (1)$$

$$\pi_{t+1} = \quad \frac{\rho + (1-\rho).\pi_t}{\rho + (1-\rho).\pi_t + \lambda.(1-\rho).(1-\pi_t)} \quad \text{if} \quad Y_{t+1} = 1$$

This recursion is denoted further on as $\pi_{t+1} = B(\pi_t, Y_{t+1})$.

The joint process (X_t, π_t) is also a Markov process. The upper part of fig. 1 represents the related Markov chain (X_t, Z_t) describing the behavior of the plant up to the time when a repair is carried out (arcs describing a repair action have been omitted: they lead from any state to (0,0)). Z_t counts how many defective items have been produced one after the other. Given that $\pi_0 = 0$ when the process returns to the state $(0,0)$, after a repair or after a good item has been produced, the value $Z_t = n$ uniquely defines $\pi_t = \pi(n)$.

It can be proven (see **?**), that a threshold policy is optimal for this partially observed Markov decision process (POMDP): repair as soon as $\pi_t > \pi_{th}$ (or equivalently when $Z_t > z_{th}$ where z_{th} is the smallest integer such that $\pi(z) \geq \pi_{th}$. The optimal value for the threshold can be calculated using a renewal type argument, albeit more complicated than in the full state information case. Using the optimal threshold policy each cycle starts in the state $(0,0)$ with $\pi_0 = 0$, and lasts until $Z_t \geq z_{th}$. The average discounted cost $J_{cycle}(z)$, during one cycle until the next repair, and using $z_{th} = z$, can be calculated using backward recursion (as in the classical absorption problem for Markov processes) $C(x,q) = E(\text{cost remainder of cycle} \mid (X_t, Z_t) = (x,q))$.

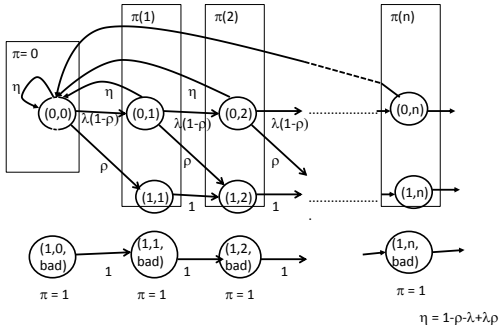

Figure 1. Markov model representing state of machine, the number of successive defective items, and the conditional distribution π_t

Clearly $C(x,z_{th}) = r+1, x = 0,1$ (since the cycle ends with a defective item and a repair). Backward recursion gives values for $C(x,z_{th}-j), j = 1,\ldots,z_{th}$ until $C(0,0) = J_{cycle}(z)$ is found. It is possible to calculate in the same way the probability distribution of the duration of a cycle. This allows one to calculate the discounted cost for each value of z, and using the fact that this cost has a unique minimum as a function of z, to find the optimal threshold z_{th} by repeating the calculation for $z = 1,2,\ldots$ until the discounted cost starts increasing.

Two types of error cause the partial information discounted cost to be higher than in the full state information case. A useless repair is carried out if π_t exceeds the threshold while the machine is still in the good state (if by accident z_{th} defective items are produced one after the other while the machine is in the good state). Or else defective items at a cost 1 are produced because the repair is not carried out while the machine is already in the bad state (because the optimal threshold z_{th} must be set high enough to avoid repair while the machine is in the good state). The cost increment due to those uncertainties thus depends on what are called type 1 and type 2 errors in hypothesis testing.

Note that for this simple POMDP the value function $V_{part}(\pi)$ can actually be calculated explicitly, using $J_{cycle}(z_{th})$. For more realistic examples though this value function in the POMDP case is very difficult to obtain. Consider the case where the repair is successful only with probability p_r, then after a repair the system restarts a new cycle in state $(0,0)$ with probability p_r, and in state $(1,0,bad)$ with probability $1-p_r$ (remaining in the bad state until the next repair, as indicated in the lower part of fig. 1). Proving optimality of a threshold policy and calculating an optimal threshold (if a threshold policy would be optimal) becomes a lot more difficult.

The system performance can be improved by activating some expensive sensor, in this case by inspecting the machine, at a cost h per inspection, at those points in time when the control agent selects $u_{s,t} = 1$ ($u_{s,t} = 0$ indicates no inspection). The observation $Y_{ins,T_{ins}}$ at inspection time T_{ins} is: if $X_{T_{ins}} = 0$ then $Y_{ins,T_{ins}} = 0$ (denoting good) with probability $p_{i,g}$, $Y_{ins,T_{ins}} = 1$ (denoting bad) otherwise; if $X_{T_{ins}} = 1$, then $Y_{ins,T_{ins}} = 1$ (bad) with probability $p_{i,b}$, good with probability $1-p_{i,b}$. Inspection improves the accuracy of the estimator π_t, and reduces the cost due to making wrong decisions, provided the cost of inspection is less than the future discounted cost reduction it achieves. The cost to be minimized is now

$$J_\gamma(\mathfrak{u}_a,\mathfrak{u}_s) = E_{\mathfrak{u}_a,\mathfrak{u}_s}[\sum_{t=1}^{\infty}\gamma^t.[Y_t + h.u_{s,t}] + r.\sum_{n=1}^{\infty}\gamma^{T_n}].$$

The control agent now must decide not only when to repair but also when inspect so as to minimize this discounted cost. Assuming an inspection is instantaneous the conditional distribution $\pi_{T_{ins}+}$ will be obtained by a Bayesian fusion of the a priori information $\pi_{T_{ins}-}$ with the information $Y_{ins,T_{ins}}$.

In order to calculate the increase in cost due to wrong state estimates causing wrong decisions, one has to calculate the pinned conditional distributions $\pi_{z,t}(\omega)$ for $\omega \in \Omega_{z,t} = \{\omega \in \Omega : X_t(\omega) = z\}, z = 0,1$. Immediately after a repair at time $t = T_n$, when $X_{T_n}(\omega) = 0$, the pinned conditional distribution becomes $\pi_{0,T_n+}(\omega) = 0, \omega \in \Omega_{0,T_n}$ using the fact that a repair never changes the good state of the machine to the bad state. If the machine is in the bad state X_{T_n} just prior to the repair, then the pinned conditional distribution immediately after the repair is $\pi_{1,T_n+}(\omega) = 1-p_r, \omega \in \Omega_{1,T_n}$ since the repair is successful only with probability p_r. In between inspection times T_{ins}, while no repair is carried out, $\pi_{x,t}, x = 0,1$ is updated as follows (in our model the Markov process X_t first selects an update to X_{t+1}, and only then does the observation process Y_t select a new random value): $\pi_{0,t+1} = B(\pi_{0,t}, Y_{t+1})$, since $X_{t+1} = 0$ can be reached only if $X_t = 0$, while $X_{t+1} = 1$ can be reached from $X_t = 0$ with likelihood $\rho.(1-\pi_t)$, and from $X_t = 1$ with likelihood π_t. Note that the likelihood of being in a state at the preceding time is according to the conditional distribution π_t, not according to the pinned distributions since the selection of the next transition, in Ω, is made before the output selection. Normalizing, and then updating the pinned conditional distributions, gives:

$$\pi_{1,t+1} = \frac{\rho.(1-\pi_t).B(\pi_{0,t},Y_{t+1}) + \pi_t.B(\pi_{1,t},Y_{t+1})}{\rho.(1-\pi_t)+\pi_t}$$

This update can be iterated until the next repair or the next inspection.

At the time of an inspection the pinned conditional distribution $\pi_{0,T_{ins}}$ is updated to $\pi_{0,T_{ins}+} = \frac{\pi_{0,T_{ins}}.(1-p_{i,b})}{p_{i,g}}$ if $Y_{ins,T_{ins}} = 0$; if $Y_{ins,T_{ins}} = 1$ then $\pi_{0,T_{ins}+} = \frac{\pi_{0,T_{ins}}.p_{i,b})}{1-p_{i,g}}$ This result follows from conditioning in the appropriate subset of Ω : combine the information $\pi_{0,T_{ins}}$ obtained by observing $Y_\tau, \tau \leq t$, with the inspection outcome, which in $\Omega_{0,T_{ins}}$ is $Y_{ins,T_{ins}} = 0$ with probability $p_{i,g}$, or $Y_{ins,T_{ins}} = 1$ with probability $1 - p_{i,g}$. The update of $\pi_{1,T_{ins}}$ is analogous, now using the fact that in $\Omega_{1,T_{ins}}$ the outcome of the inspection $Y_{ins,T_{ins}} = 1$ with probability $p_{i,b}$, $Y_{ins,T_{ins}} = 0$ with probability $1 - p_{i,b}$.

The discussion in this section does not actually define a good partial information control law $\mathfrak{u}_{a,t}$, it only describes what a good control law should depend on: the conditional distribution π_t, and the pinned conditional distributions $\pi_{z,t}, z \in \mathbb{X}$. In section 4 we will derive some methods for selecting good suboptimal control laws, and in subsection 5.1 we will show how to apply these suboptimal control laws to the machine repair problem.

3. Partially observed Markov decision processes

The example treated in section 2 was extremely simplified. In section 4 we analyze for a general Markov process $X_t \in \mathbb{X}$ the vector of conditional distributions $\pi_t(x) = \mathscr{P}(X_t = x \mid \mathscr{H}_t) \in \mathbb{R}_+^{\sharp\mathbb{X}}$, and of $\sharp\mathbb{X}$ pinned conditional distributions $\pi_{z,t}(x), x, z \in \mathbb{X}$ (to be defined below). Using π_t, $\pi_{z,t}$ we synthesize feedback control laws, depending only on the observed history, that guarantee an upper bound on the increase in operational cost due to the uncertainty about the current state. This method for synthesizing suboptimal partial information control laws, and the calculation of the bound on the resulting increase in cost, require only knowledge of the value function $V(x)$ for the full state dynamic programming problem, not the much more complicated value function $V_{part}(\pi)$ for the POMDP. In this section we review those results on Markov decision theory that are most relevant to our method.

This paragraph summarizes the full state feedback results for Markov decision problems for finite state spaces in discrete time (extensions to countable and continuous state spaces require additional conditions): consider the Markov process $X_t \in \mathbb{X} \subseteq \mathbb{Z}, \forall t \in \mathbb{Z}, X_0 = x_0 \in \mathbb{Z}$ with transition probabilities

$$\Pi(x \to x'; u_{a,t}) \tag{2}$$

that depend on the control values $u_{a,t}(X_t) \in U_a$. Control law $\mathfrak{u}_{a,t}(\mathscr{H}_t) \in \mathbb{U}_a$ is selected as a function of the past states $X_\tau, \tau \leq t$ so as to minimize the discounted cost

$$J_\gamma(u, X_0) = E_u\big(\Sigma_{t\in\mathbb{Z}}\gamma^t.[c(X_t) + g(\mathfrak{u}_{a,t})] \mid X_0\big) \tag{3}$$

where E_u represents the expectation for the controlled process (this may not be a Markov process if $\mathfrak{u}_{a,t}$ depends on more than the current state X_t).

For this discounted cost problem there exists a stationary Markovian optimal control law $\mathfrak{u}_a^*(x) : \mathbb{X} \to U_a$ that depends only on the current state X_t, independent of time, that is at least as good as any control law in the set $\mathbb{U}_a$. If such a stationary Markovian control law $\mathfrak{u}_a(X_t)$ is implemented then the controlled process is a Markov process with transition probabilities $\Pi(x \to x'; \mathfrak{u}_{a,t}(x))$. The lowest achievable cost, given the initial state $X_0 = x$ is the value function

$$V_\gamma(x) = inf_{\mathbb{U}_a} J_\gamma(\mathfrak{u}, x) = J_\gamma(\mathfrak{u}_a^*, x) \tag{4}$$

HJB Optimality Theorem (see e.g. **?**, **?**) The value function $V_\gamma(x)$ is the unique positive bounded solution to the equation:

$$\begin{aligned} \forall x \in \mathbb{X} : V_\gamma(x) = \quad & inf_{u_a \in U_a}\{c(x) + g(u_a) + \\ & +\gamma.\Sigma_{y\in\mathbb{X}} V_\gamma(y).\Pi(x \to y; u_a)\} \end{aligned} \tag{5}$$

If $\forall x \in \mathbb{X}$ a value $u_a^* \in U$ can be selected such that

$$V_\gamma(x) = c(x) + g(u_a^*(x)) + \gamma.\Sigma_{y\in\mathbb{X}} V_\gamma(y).\Pi(x \to y; u_a^*) \tag{6}$$

then applying the stationary Markovian control law $\mathfrak{u}_a^*(X_t) = u_a^*$ at each time t when the current state is X_t minimizes the future discounted cost of the system. The same optimal control law is obtained if at each time t the control value $u_{a,t}^*$ is selected as the first component of the sequence of control laws $\mathfrak{u}_{a,j}^* \in \mathbb{U}, j = 0, \ldots, H-1$ that minimizes the control laws (not control values) over the prediction window $[t, t+H-1]$:

$$\begin{aligned} V_\gamma(x) = \quad & inf_{\mathfrak{u}_{a,t+j}\in\mathbb{U}_a, j=0,\ldots,H-1} E_{u_{a,j}, j=0,\ldots,H-1} \\ & [\Sigma_{j=0}^{H-1} \quad c(X_{t+j}) + g(u_{a,t+j}) + \gamma^H.V_\gamma(X_{t+H}) \mid X_t = x] \end{aligned} \tag{7}$$

If $\forall x \in \mathbb{X}$ a value $u_a^\delta(x) \in U$ can be selected s. t.

$$\begin{aligned} V_\gamma(x) \leq \quad & c(x) + g(u_a^\delta(x)) + \\ & +\gamma.E_{u_a^\delta(x)}[V_\gamma(X_{t+1}) \mid X_t = x] + \delta_t \end{aligned} \tag{8}$$

where δ_t is a possibly random time series, then applying the control law $\mathfrak{u}_a^\delta(x)$ at each time t when the current state is $X_t = x$, guarantees that the future discounted cost of operating the system achieves a cost that is at most $V_\gamma(X_t) + E_{u_a^\delta}\Sigma_{t\in\mathscr{Z}}\gamma^t.\delta_t$. If a deterministic upper bound $\forall t : \delta \geq \delta_t$ is known then this discounted cost is at most

$V_\gamma(X_t) + \frac{\delta}{1-\gamma}$ (see e.g. **?**, pp.152-155). These control laws $u_a^\delta(x)$ are called δ-suboptimal controllers.

Moreover if $\forall x \in \mathbb{X}$ (8) is replaced by the inequality "left hand side of (7)" $\leq$ "right hand side of (7)" + δ," then the discounted cost when applying control law $\mathfrak{u}_a^\delta(X_t)$ is at most $\delta/(1-\gamma)$ higher than $V_\gamma(X_t)$. $\diamond$

Note that in (7) the first control selection $u_{a,t}$ requires only the choice of a control value $\in U_a$, (since the current state $X_t = x$ is supposed to be known), but at each later time, for $\mathfrak{u}_{a,t+j} \in \mathbb{U}_a, j =, \ldots, H-1$, one must select a control law. Indeed (7) is valid only provided one maintains the closed loop control over the window $[t, t+H-1]$. Classical textbooks only prove the computationally much more efficient case $H = 1$. While the minimization in (5) is over the value space U_a, the minimization in (7) is over an H-dimensional space of control laws, a vector in $U_a^{H,\sharp\mathbb{X}}$. However as explained in the introduction we will need (7) later on for horizons $H > 1$ because the effect of a sensor activation decision is not quantified properly for $H = 1$. This result should be compared to **?**.

Note that (8) requires knowledge of the exact value function $V_\gamma(x)$ for finding a suboptimal controller with bounded increase in cost compared to the optimal controller. It is not sufficient to use an approximation to the value function.

In practice it is usually not possible to apply the optimal feedback control law $u_a^*(X_t)$ because the state X_t cannot be measured accurately. The control agent only receives at time t noisy observation $Y_t \in \mathbb{Y}$ depending on the state $X_{\tau \leq t}$. Specifically the output $Y_{t+1} = y \in Y$ is generated according to the probability distribution $Q(y; x, x', u_{s,t})$. The control agent selects at each time t an actuator control value $u_{a,t} \in U_a$ and a sensor activation decision $u_{s,t} \in U_s$. Selecting the sensor activation control value $u_{s,t}$ causes an instantaneous cost $h(u_{s,t})$, and determines how much information $Q(y; x, x', u_{s,t})$ provides about x and x' (we allow the sensor output to depend on previous and current state in order to be able to observe transitions, e.g. an arrival in a queue corresponding to $x' - x = 1$). Of course true in all applications a more expensive choice $u_{s,t}$ generates output Y_t that provides more information on X_t, allowing better control decisions in the future.

In the classical setup for the partial information Markov decision problem (POMDP) the control agent remembers at each time t the history

$$\mathscr{H}_t = \{H_{init}, Y_0, u_{a,0}, u_{s,0}, \ldots, Y_{t-1}, u_{a,t-1}, u_{s,t-1}, Y_t\}.$$

Actuator and sensor control values are selected as functions of $\mathscr{H}_t$ so as to minmize the discounted cost:

$$J_\gamma(u,x) = E_u\{\Sigma_{t\in\mathbb{Z}} \gamma^t.[c(X_t) + g(u_t) + h(v_t)] \mid H_{init}, Y_0\} \tag{9}$$

It is well-known that these POMDP (**?**,**?**) can be treated as classical MDPs by replacing the state X_t by the conditional distribution $\pi_t = P(X_t = x \mid \mathscr{H}_t)$ of X_t given the observations available up to time t. The transition probabilities of $\pi_t = P(X_t = x \mid \mathscr{H}_t)$, given the control values $u_{a,t} = u_a, u_{s,t} = u_s$ are defined by the recursive *Bayes' algorithm:*
Calculate $\pi_{t+1}^- = \mathscr{P}(X_{t+1} = x' \mid \mathscr{H}_t) = \Sigma_{x\in\mathbb{X}} \pi_t(x).\Pi(x \to x'; u_a)$;
apply Bayes' rule
$\tilde{\pi}_{t+1}(x') = \Sigma_{x\in\mathbb{X}} \pi_t(x).\Pi(x \to x'; u_a).Q(y; x, x', u_s)$;
normalize $\pi_{t+1}(x') = \frac{\tilde{\pi}_{t+1}(x')}{\Sigma_{x'\in\mathbb{X}} \tilde{\pi}_{t+1}(x')} = B(\pi_t, \mathfrak{u}_a(\pi_t), \mathfrak{u}_s(\pi_t))(x')$

π_t is a Markov process: the distribution of Y_{t+1}, and of π_{t+1} only depend on π_t. The HJB theorem remains valid when adding an explicit cost $h(u_{s,t})$ for the sensor activations - see **?** for a similar problem formulation. Hence there exists a stationary Markovian optimal control laws $u_a^*(\pi_t), u_s^*(\pi_t)$ mapping the space of conditional distributions to the set U_a of transition control values, resp. the set U_s of observation control values. The disounted cost to be minimized is

$$J_\gamma^{\text{partial}}(\mathfrak{u}_a, \mathfrak{u}_s), \pi_0) \tag{10}$$
$$= E_{u,v}\big(\Sigma_{t\in\mathbb{Z}} \gamma^t.[\Sigma_{x\in\mathbb{X}} c(x).\pi_t(x) + g(u_{a,t}) + h(u_{s,t})] \mid \pi_0\big)$$

$V_\gamma^{\text{part}}(\pi) = inf_{\mathfrak{u}_a,\mathfrak{u}_s} J_\gamma^{\text{partial}}(\mathfrak{u}_a, \mathfrak{u}_s, \pi)$ satisfies:

$$V_\gamma^{\text{part}}(\pi) = \inf_{u_a\in U_a} \{\Sigma_{x\in\mathbb{X}} c(x).\pi(x) + g(u_a) \tag{11}$$
$$+ \inf_{u_s\in U_s} [h(u_s) + \gamma.\Sigma_{\pi'\in\Pi} V_\gamma^{\text{part}}(\pi').B(\pi_t, u_a, u_s)(\pi')]\}$$

where the set Π of possible values of the π_{t+1}, reachable from π_t can be calculated using Bayes' algorithm.

If a mapping $u_a^*(\pi), u_s^*(\pi)$ exists that achieves the minimum in (11) then this defines the stationary optimal control laws both for the actuator and for the sensor activation control. Unfortunately it is rarely possible in practice to solve this functional BHJ equation for $V_\gamma^{part}(\pi)$, and to obtain an optimal control law using (10)-(11).

4. Suboptimal history-adapted controllers with performance bound

Combining equations (7) and (8), including the sensor activation cost $h(u_{s,t})$, one can obtain a suboptimal solution with bounded increment in cost for the partial observation control compared to full state DP by using

$$V_\gamma(x) \leq E_{u_j, v_j, j=0,\ldots,H-1}[\Sigma_{j=0}^{H-1} \gamma^j.[c(X_{t+j}) + g(u_{a,j}) \tag{12}$$
$$+ h(u_{s,j})] + \gamma^H.V_\gamma(X_{t+H}) + \delta_t \mid X_t = x]$$

The value function $V_\gamma(x)$ in (12) satisfies equation (5) (and also (7)). The inequality (12) must be verified for each value $x \in \mathbb{X}$. The control agent however does not know the state $X_t = x$, so that $(u^\delta_{a,t}, u^\delta_{s,t})$ must be selected not as functions of the state X_t but as functions of the history $\mathscr{H}_t$. In order to calculate the bounds δ_t for each value of x one needs to calculate the probabilistic behaviour of π_t in a subset $\Omega_{x,t} = \{\omega : X_t(\omega) = x\} \subset \Omega = \{X_\tau, Y_\tau, \tau \in \mathbb{Z}\}$ of the probability space. The control agent can thus use the pinned conditional distribution $\pi_{x,t}(z)$, that evaluates the conditional distribution of $X_t = z$ given $\mathscr{H}_t$ for $\omega \in \Omega_{x,t}$. These pinned conditional distributions $\pi_{x,t}(z)$ allow the calculation of the difference between $V_\gamma(x)$ and the right hand side of (12):

$$\delta_t(x, u_{a,j}, u_{s,j}, j = 0, \ldots, H-1, \mathscr{H}_t) = \tag{13}$$
$$E_{u_{a,j}, u_{s,j}, j=0,\ldots,H-1}\{ \sum_{j=0}^{H-1} \gamma^j . [c(X_{t+j}) + g(u_{a,j}) + h(u_{s,j})] + \gamma^H . V_\gamma(X_{t+H}) - V_\gamma(x) \mid \pi_t, \pi_{x,t}\}$$

Note that $\pi_{x,t}(z)$ uses only as information $\mathscr{H}_t$, and does not *know* that $X_t = x$ at time t.

In order to recursively calculate $\pi_{x,t}$ observe that $\Omega_{x,t} = \cup_{x' \in \mathbb{X}} (\Omega_{x',t-1} \cap \{\omega : X_t(\omega) = x\})$. Note that $\omega \in \Omega_{x,t}$ also describes the random generation of $Y_t(\omega)$ according to $Q(y; x, x', u_s)$. For values of $\omega \subset \Omega_{x',t-1}$ the evolution from $\pi_{x',t-1}(z')$ to $\pi^+_{x',t-1}(z)$ follows the rules of the Bayesian recursive update: $\Sigma_{z' \in \mathbb{X}} \pi_{x',t-1}(z) . \Pi(z' \to z; u_a) . Q(y; z', z, u_s)$; and finally normalize. In order to calculate $\pi_{x,t}(z)$ one needs to consider only those values of ω that are in $\Omega_{x,t}$, i.e. consider for all the precursors $X_{t-1}(\omega) = x', x' \in \mathbb{X}$, $\omega \in \Omega_{x',t-1} \cap \Omega_{x,t}$ the value $\pi^+_{x',t-1}(z)(\omega)$. The weight of each of these precursors is proportional to the probability that the state at time $t-1$ is $X_{t-1} = x'$, given the prior information $\mathscr{H}_{t-1}$, is $\pi_{t-1}(x')$. In order to describe the update for each value $\pi_{x',t-1}(z), z \in \mathbb{X}$ one has to combine these weighted likelihoods $\pi^+_{x',t-1}(z) . \pi_{t-1}(x')$, taking into account that the likelihood of going from $X_{t-1} = x'$ to $X_t = x$ while observing $Y_t = y$ is $\Pi(x' \to x; u_{a,t-1}) . Q(y; x, x', u_{s,t-1})$. Thus

$$\tilde{\pi}_{x,t+1}(z) = \Sigma_{x' \in \mathbb{X}} (\phi(x')) \tag{14}$$
$$\phi(x') = \pi_t(x') . \Pi(x' \to x; u_{a,t}) . Q(y; x, x', u_{s,t}) . \pi^+_{x',t}(z)$$

followed by a normalization.

The pinned conditional probability distribution of $X_{t+j}, j = 0, \ldots, H-1$, defined only for $\omega \in \Omega_{x,t}$ (arbitrarily assigned a value $\emptyset$ for $\omega \notin \Omega_{x,t}$) must be calculated using as initial distribution at time t the pinned conditional distribution $\pi_{x,t}(z)$. Since the actual observations Y_{t+j} during the prediction window are not known at the time when the control values $(u_{a,t}, u_{s,t})$ are selected, the control agent must calculate the best possible $\mathscr{H}_t$-adapted conditional distribution $\pi^{t+j}_{x,t}(z) = \mathscr{P}(X_{t+j} = z \mid \pi_t)$, with initial condition $\pi_{x,t}$, as a function of the control values $u_{a,t+n}, u_{s,t+n}, n = 0, 1, \ldots, j-1$. Starting at time t in $\Omega_{x,t}$ the state transition to $X_{t+1} = z$ with output $Y_{t+1} = y$ occurs with probability $\Pi(x \to z; u_{a,t}) . Q(y; x, z; u_{s,t})$. The probability distribution of $Y_{t+1} = y$, given $\mathscr{H}_t$ and restricted to $\Omega_{x,t}$ is then $\mathscr{P}(Y_{t+1} = y \mid \mathscr{H}_t, \Omega_{x,t}) = \Sigma_{z \in \mathbb{X}} \Pi(x \to z; u_{a,t}) . Q(y; x, z; u_{s,t})$, independent of $\mathscr{H}_t$. This argument can be extended to the joint distribution

$$\mathscr{P}(Y_{t+n} = y_n, n = 1, \ldots, H \mid \Omega_{x,t}) = \tag{15}$$
$$= \Sigma_{(z_1,\ldots,z_H) \in \mathbb{X}^H} Pr(z_1, \ldots, z_H)$$

with

$$Pr(z_1, \ldots, z_H) = \Pi(x \to z; u_{a,t}) . Q(y; x, z; u_{s,t}) \times \Pi_{n=1}^H \Pi(z_{n-1} \to z_n; u_{a,t+n-1}) . Q(y_n; z_{n-1}, z_n; u_{s,t+n-1})$$

The calculation of $\pi^{t+j}_{x,t}(z)$ uses only $\mathscr{H}_t$ and the fact that $\omega \in \Omega_{x,t}$. The trajectories of $\pi^{t+j}_{x,t}(z)$ are obtained by recursively applying Bayes' algorithm (see section 3) for each possible sequence of observations $Y_{t+n} = y_n, n = 1, \ldots, H$, generating $\pi^{t+j}_{x,t}(z)(y_n, n = 1, \ldots, H)$ and then averaging according to $\pi^{t+j}_{x,t}(z) = \Sigma_{y_n, n=1,\ldots,H} \mathscr{P}(Y_{t+n} = y_n, n = 1, \ldots, H \mid \Omega_{x,t}) . \pi^{t+j}_{x,t}(z)(y_n, n = 1, \ldots, H)$ assigning the probability calculated in (15) to each of these observation sequences $y_n, n = 1, \ldots, H$.

These distributions $\pi^{t+j}_{x,t}(z)$ allow the calculation of

$$\delta_t(x, u_{a,j}, u_{s,j}, j = 0, \ldots, H-1, \mathscr{H}_t)$$

according to (13). In order to obtain a good suboptimal control law the control agent must select values $(u_{a,j}, u_{s,j}) \in (U_a \times U_s)^H$ that minimize

$$\delta_t(u_{a,j}, u_{s,j}, j = 0, \ldots, H-1, \pi_t, \pi_{z,t}, z \in \mathbb{X}) = \tag{16}$$
$$\max_x \delta_t(x, u_{a,j}, u_{s,j}, j = 0, \ldots, H-1, \pi_t, \pi_{x,t})$$

The lowest possible value of (16) achievable by an optimal choice of $u_{a,j}, u_{s,j}, j = 0, \ldots, H-1$ is denoted δ_t^{minmax}. Since eqns (13, 14, 15) depend on both $u_{a,t+j}$ and $u_{s,t+j}$, the optimization defining the control laws must be done jointly. This makes the problem a lot more complicated than the special case for LQG models with observation cost where the separation theorem allows separate selection of the actuator and of the sensor activation controllers. By using an expensive sensor activation control value $u_{s,t+j}$ it may be possible to get very accurate estimates of the states $X_{t+j+\ell}$ which in turn may allow the control agent to select a very good

actuator control value $u_{t+j+\ell}$. In fact for some values of u_s some particularly bad states z' may be excluded completely, if $Q(y;z,z';u_s)=0$ for some values y.

Properties of the proposed suboptimal controllers:

Selecting the $\mathscr{H}_t$-adapted control laws $u^{\delta}_{a,t}, u^{\delta}_{s,t}$ according to the first u_a and u_s components of

$$(u^{\delta}_{a,t}, u^{\delta}_{s,t}) = argmin_{(u_a,u_s)\in(U_a\times U_s)^H} \quad (17)$$
$$\delta_t(u_{a,j}, u_{s,j}, j=0,\ldots,H-1, \pi_t, \pi_{x,t}, x\in\mathbb{X})$$

ensures that the expected discounted cost of operating the system, starting in any state $X_t = x$, is at most

$$V(x) + \sum_{t\in\mathbb{Z}} \gamma^t.\delta^{minmax}_t$$

Remark 1: Note that δ^{minmax}_t includes a cost for sensor activation, depending on $h(u_{s,t})$. This must be taken into account when interpreting the cost increase compared to $V_\gamma(X_t)$, due to applying $(u^{\delta}_{a,t}, u^{\delta}_{s,t})$. The sensing cost of course also is included in the POMDP minimal cost, which is higher than the full state information cost even when $h(u_{s,t})=0$.

Remark 2: The horizon H must be selected long enough so that the future cost reduction thanks to the sensor activation is properly quantified in (16). This choice is however limited by the computational complexity of the online algorithm.

5. examples

The machine repair problem of section 2, in the case where both the repair and the inspection are not perfect, is a simple example of the application of the suboptimal control law proposed in section 4. The δ-values on $\Omega_{z,t}, z=0=$ good and $=1=$ bad for a window of size H are

$$\delta(z, u_{a,j}, u_{s,j}, j=0,\ldots,H-1) \quad (18)$$
$$= E_{u_{a,j},u_{s,j},j=0,\ldots,H-1}\{\sum_{j=0}^{H-1}\gamma^j.[Y_{t+j}) + r.u_{a,t+j} + \quad h.u_{s,t+j}]$$
$$+\gamma^H.V_\gamma(X_{t+H}) - V_\gamma(z) \mid \pi_t, \pi_{x,t}\}$$

The value functions $V(0) = \frac{\lambda}{(1-\gamma)} + (r+1).\frac{\rho.\gamma.(1-\rho)}{1-\gamma+\gamma.\rho^2}$ and $V(1) = \gamma.V(0) + r + 1$, and the transition probabilities for the corresponding Markov chain, have been calculated in section 2. The possible control actions are at each time t to repair or not to repair, and to inspect or not to inspect. Evaluate (18) for each possible choice of inspection and repair times in the interval $t, t+1, \ldots, t+H-1]$, and select $(u_{a,j}, u_{s,j}, j=0,\ldots,H-1)$ that minimizes $max_{z=0,1}\delta(z, u_{a,j}, u_{s,j}, j=0,\ldots,H-1)$.This minimum is denoted δ^{minmax}. This specifies the suboptimal inspection and repair times according to the suboptimal control law of section 4. Theoretically this requires 2.2^H evaluations of $\delta(z, u_{a,j}, u_{s,j}, j=0,\ldots,H-1)$, but in practice many cases can be excluded (like inspection immediately following a repair). Moreover one can calculate the cost increment $\delta^{minmax}/(1-\gamma)$ of this repair and inspection strategy, with respect to the full state information case.

Other more complcated examples can be treated by the same approach. Approximate solutions can be found using the pinned conditional distribution approach for the control of the optimal arrival rate or of the optimal service intensity in networks of queues. We have also considered applications to jump Markov LQG problemms in the case where the underlying Markov state is not dirrectly observable.

6. Conclusions

This paper has introduced a novel approach to designing actuator and sensor activation control laws, for the case where the system state can only be observed partially and at a cost. The approach provides a bound on the increment in the cost, due to operating the plant under the proposed partial information strategy as compared to the full state case.

References

D. Bertsekas: Dynamic programming. Deterministic and Stochastic Models. Prentice Hall, Inc Englewoods Cliff, 1987

R. Boel and J. H. van Schuppen: Control of the observation matrix for control purposes, MTNS 2010, Budapest, pp. 1261-1268

G. Koole: A transformation method for stochastic control problems with partial observations, Systems and Control Letters 35:301-308, 1998

P.R. Kumar and P. Varaiya: Stochastic systems: Estimation, identification and adaptive control, Prentice-Hall, 1986

D. Q. Mayne, J. B. Rawlings, C. V. Rao, P. O. M. Scokaert: Constrained model predictive control: Stability and optimality, Automatica, 36 (2000) pp. $789-814$

M. Zhong et al.: Function Approximation and Model Predictive Control, 2013 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), Singapore