Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

A transformation method for stochastic control problems with partial observations

G. Koole

Department of Operations Reasearch, Statistics, and System Theory

BS-R9601 1996

Report BS-R9601
ISSN 0924-0659

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

CWI is the National Research Institute for Mathematics and Computer Science. CWI is part of the Stichting Mathematisch Centrum (SMC), the Dutch foundation for promotion of mathematics and computer science and their applications.
SMC is sponsored by the Netherlands Organization for Scientific Research (NWO). CWI is a member of ERCIM, the European Research Consortium for Informatics and Mathematics.

Copyright © Stichting Mathematisch Centrum
P.O. Box 94079, 1090 GB Amsterdam (NL)
Kruislaan 413, 1098 SJ Amsterdam (NL)
Telephone +31 20 592 9333
Telefax +31 20 592 4199

A Transformation Method for Stochastic Control Problems with Partial Observations

Ger Koole

INRIA, B.P. 93, 06902 Sophia Antipolis, France

Abstract

Two stochastic control problems with partial observations are studied, one where the policy or control law depends only on the latest observation (the controller does not have recall of observations), and the other with the standard partial observations model. The equivalent full observation problems are formulated, and the equivalence is proven using a new method. The results are illustrated with a simple example.

AMS Subject Classification (1991): 90C40, 93E20

Keywords & Phrases: Stochastic control, Markov decision problems, partial observations, separation of estimation and control

Note: Supported by the Netherlands Organization for Scientific Research (NWO) through project NFI 62-354 and by the European HCM Grant ERBCHBGCT930254.

1. Introduction

In this paper we study stochastic control problems with partial observations. At each point in time the state observation is disturbed by some noise, and the control has to be based on these noisy observations. In the first problem that we study, the *no recall* model, the control can only use the current observation. The second model is the standard partial observation model where the control is allowed to depend on all observations up to the current epoch.

For both models we formulate an additional model where policies are allowed to use full state observation. A new method, applicable to both information structures, is used to prove the equivalence of the original and the additional model. In the last Section we illustrate the results by analyzing a simple system.

We begin in Section 2 by describing the model in general terms. To simplify the analysis we assume that state, action and observation sets are finite. Then we formulate the method. A second model, using full state observations, is formulated. Both models are related by a functional relation between their allowable policies. Then conditions are given under which minimal costs are equal, and it is indicated how the optimal policy can be found for the original model, from the optimal policy for the second one. This is formulated in Theorem 2.1. What remains is solving an, often complicated, full observation problem.

One of the reasons to describe the method in general terms in a separate Section is that it can also be used for models with other information structures, like delayed or decentralized information. Work in this direction is currently going on.

The first model we consider (in Section 3) is one with partial observation, but *without* recall, i.e., allowable policies depend only on partial observation on the current state, on the

initial distribution, and of the decision rules used earlier (but not on the actions taken earlier, as this would give information on the earlier observations). The states of the equivalent full observation problem are distributions on the states of the original model.

The case where the policies are allowed to depend on all observations up to the current is studied in Section 4. We derive the well known result that, again, the equivalent problem uses distributions on the original states as states. The difference with the result of Section 3 for the model without recall lies in the transition mechanism. Note that the result can already be found in the literature (see e.g. Kumar & Varaiya [5] or Bertsekas [2]), but that it is obtained using different proofs.

In Section 5 the results are applied to a two-state model. In spite of its simplicity, several concepts regarding the structure of the optimal policies and the value of information are nicely illustrated.

2. The method

First we define the stochastic control problem we consider, in a general setting, without specifying the information structure. Then we define the possible information structures and present the method.

The stochastic control problem is defined by $(\mathcal{X}, \mathcal{Y}, \mathcal{A}, r, q, c)$, with $\mathcal{X}$ the finite state space, $\mathcal{Y}$ the finite observation space, $\mathcal{A}$ the finite set of actions (assumed to be equal in each state), r the transition probabilities (i.e., $r(x, a, x')$ is the probability is going from x to x' using action a), q the observation probabilities (i.e., $q(x, y)$ is the probability of observing y in state x), and c the direct costs (i.e., $c(x, a)$ are the costs incurred if in state x action a is chosen).

The finiteness of the sets involved is important, as it will simplify the analysis considerably. Other choices, as the time homogeneity, or the fact that the observation depend only on the state and not on the action, are less important. We did not make the model more general as to avoid further notational complexity.

Assume that the system is to be controlled from 1 to T (possibly ∞). A policy or control law ϕ is a set of decision rules $\phi_1, \ldots, \phi_T$, with $\phi_t : \mathcal{H}_t \to \mathcal{A}$. Here $\mathcal{H}_t = (\mathcal{X} \times \mathcal{Y} \times \mathcal{A})^{t-1} \times \mathcal{X} \times \mathcal{Y}$.

We could make the model more general by allowing random actions, that is by letting $\phi_t(h_t)$ be a distribution on the action set. Again, as to keep the notation simple, we will not do that. This choice can be justified: we prove for the models we consider that they are equivalent to full observation models, and for these models it is well known that there exist optimal non-randomized policies.

Define, for a fixed policy ϕ, the r.v.'s X_t, Y_t and A_t as follows (sometimes we write $X_t(\phi)$, etc., to denote the dependence on the policy). Assume that the initial distribution of X_1 is given. For each partial realization $h_t = (x_1, y_1, a_1, \ldots, x_{t-1}, y_{t-1}, a_{t-1}, x_t, y_t) \in \mathcal{H}_t$ of $(X_1, Y_1, A_1, \ldots, X_t, Y_t) = H_t$, A_t is given by $\phi_t(h_t)$, then, given $H_t = h_t$, X_{t+1} has the value x with probability $r(x_t, a_t, x)$, and Y_{t+1} (given $X_{t+1} = x$) then has value y with probability $q(x, y)$. For obvious reasons h_t is called the history of the system at t.

Now define $C(\phi) = E \sum_{t=1}^{T} \beta^t c(X_t, A_t)$, the expected discounted costs (with $0 < \beta \leq 1$ the discount factor) under ϕ. The problem is to find a policy ϕ^* such that $C(\phi^*) = \min\{ C(\phi) \mid$

$\phi \in \Phi'\}$, for some $\Phi' \subset \Phi$, with Φ the set of all policies. This optimal policy ϕ^* (if one exists) is notated as $\phi^*(\Phi')$, or just ϕ^*.

The ideas introduced so far can be found in most books on Markov decision problems or stochastic control theory, like Kumar & Varaiya [5] or Bertsekas [2].

In this paper we study models in which ϕ_t is a function of one or all of the $\mathcal{Y}$-components. By this we mean that $\phi_t(h_t) = \phi_t(h'_t)$ if $y_t = y'_t$, or $y_s = y'_s$ for $s \leq t$, respectively, with $h_t = (x_1, y_1, a_1, \ldots, x_t, y_t)$ and $h'_t = (x'_1, y'_1, a'_1, \ldots, x'_t, y'_t)$.

Note that other information structures can be modeled as well, like delayed information in which ϕ_t depends on $x_1, \ldots, x_{t-K}$ for some delay $K > 0$, or decentralized information structures where different controllers observe different components of the multi-dimensional states. These other structures are the subject of ongoing research.

Let us now consider the solution method.

For the problems in the next Sections we formulate a second problem $(\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, \tilde{r}, \tilde{q}, \tilde{c})$, where the allowable policies depend on the actual state of the system (therefore we omitted the observation from the definition of the system). All variables related to this new problem are indicated with a $\tilde{}$. The relation between the original system and the new one is defined by a function $\Gamma : \Phi' \to \tilde{\Phi}$. Our objective is to formulate the second problem and Γ in such a way that $\Gamma(\phi^*(\Phi')) = \tilde{\phi}^*(\tilde{\Phi})$, and that the costs of these policies are equal. Having found $\tilde{\phi}^*(\tilde{\Phi})$, we know then that a $\phi \in \Phi'$ with $\Gamma(\phi) = \tilde{\phi}^*(\tilde{\Phi})$ is optimal. Let $\stackrel{\mathrm{d}}{=}$ stand for equally distributed. Define an equivalence relation $\sim$ on $\tilde{\Phi}$ as follows: $\tilde{\phi} \sim \tilde{\phi}'$ if $\tilde{H}_t(\tilde{\phi}) \stackrel{\mathrm{d}}{=} \tilde{H}_t(\tilde{\phi}')$ for all $t \leq T$. It is easily verified that this is indeed an equivalence relation.

Theorem 2.1 *Consider the stochastic control problem for the two models defined above. Assume that there exists a function $\Gamma : \Phi' \to \tilde{\Phi}$. If:*
i) $Ec(X_t(\phi), A_t(\phi)) = E\tilde{c}(\tilde{X}_t(\Gamma(\phi)), \tilde{A}_t(\Gamma(\phi)))$ for all $\phi \in \Phi'$ and t;
ii) in every equivalence class there is a $\tilde{\phi} \in \tilde{\Phi}$ (which we call the representative of its class) such that there exists a $\phi \in \Phi'$ with $\tilde{\phi} = \Gamma(\phi)$;
iii) there exists an optimal policy for the second problem;
then every $\phi \in \Phi'$ with $\Gamma(\phi)$ optimal in the second problem is also optimal in the original, and at least one such ϕ exists.

Proof From $\tilde{H}_t(\tilde{\phi}) \stackrel{\mathrm{d}}{=} \tilde{H}_t(\tilde{\phi}')$ it follows that $\tilde{C}(\tilde{\phi}) = \tilde{C}(\tilde{\phi}')$. From this we conclude that if a policy is optimal for the second problem, then so are all other policies in its equivalence class. Thus, by iii), there exists an optimal policy $\tilde{\phi}^*$ which is also the representative of its class, and therefore by ii) there exists also a $\phi^* \in \Phi'$ with $\Gamma(\phi^*) = \tilde{\phi}^*$. Thus, by i), $\min\{ C(\phi) \mid \phi \in \Phi' \} \leq \min\{ \tilde{C}(\tilde{\phi}) \mid \tilde{\phi} \in \tilde{\Phi} \}$. Also by i) $\tilde{C}(\Gamma(\phi)) = C(\phi)$ for all $\phi \in \Phi'$, and thus $\min\{ C(\phi) \mid \phi \in \Phi' \} \geq \min\{ \tilde{C}(\tilde{\phi}) \mid \tilde{\phi} \in \tilde{\Phi} \}$. As $C(\phi^*) = \tilde{C}(\tilde{\phi}^*)$, it follows that ϕ^* is optimal. Any other $\phi \in \Phi'$ such that $\Gamma(\phi) = \tilde{\phi}^*$ is also optimal. □

For the models we are going to study in the next Sections we formulate the second problem and then we show i) and ii). As for both models this second problem has compact state spaces, finite action spaces and uniformly bounded costs, optimal policies exists for the finite horizon (i.e., $T < \infty$) problem and for the infinite horizon discounted (i.e., $T = \infty$ and

$0 < \beta < 1$) problem (Schäl [7]). Having thus transformed the problem into a problem with a full information pattern, for which an optimal policy exists in the cases indicated above, we can use the standard techniques for this (most notably dynamic programming) to find the optimal policy.

Let us finish this Section with some remarks on computational issues. Theoretically speaking, the results of this paper gives the means to solve partial observation problems with the correct information structure. Numerically however, the compactness of the state spaces of the problems to be solved prohibits the direct use of standard methods like dynamic programming. This explains the lack of models studied in the literature. A discrete approximation is used for the system of Section 5, explaining partly why we only consider a two-state model (the other reason being that there is no need to study a bigger model; the current illustrates the results of this paper already perfectly).

For average or discounted optimality Hordijk & Loeve [4] provide a numerically attractive computational method to solve partial observation models, although it is not guaranteed to give the optimal policy. The first results however, as reported in Hordijk et al. [3], are promising. It can also be applied to the no recall model of Section 3.

Several other methods, using methods such as linear programming and branch-and-bound, can be found in the literature. See Loeve [6] for an overview.

3. No recall information pattern

A simple but interesting model to apply this to is the model without recall. Here ϕ_t can be written as $\phi_t(y_t)$, thus Φ' consists of all functions which depend on the last component $y_t \in \mathcal{Y}$ only: the controller forgets previous observations.

Let $\mathcal{P}$ denote the set consisting of all probability distributions on $\mathcal{X}$. Thus $\mathcal{P} \subset [0,1]^{|\mathcal{X}|}$. For a $p \in \mathcal{P}$ we denote with $p(x)$ the probability mass on $x \in \mathcal{X}$. We take $\tilde{\mathcal{X}} = \mathcal{P}$. The action set is defined by $\tilde{\mathcal{A}} = \mathcal{A}^{|\mathcal{Y}|}$, thus the actions in the second model can be seen as consisting of a vector, for each possible observation one. For $\tilde{a} \in \tilde{\mathcal{A}}$ the yth component is denoted with $\tilde{a}(y)$.

The transitions are defined as follows: $\tilde{r}(p, \tilde{a}, p') = 1$ for the distribution p' such that $p'(x') = \sum_{x,y} p(x)q(x,y)r(x,\tilde{a}(y),x')$. We take $\tilde{c}(p,\tilde{a}) = \sum_{x,y} p(x)q(x,y)c(x,\tilde{a}(y))$ as direct costs. As initial state p_1 we take the distribution of X_1, thus $p_1(x) = P(X_1 = x)$.

Now take an allowable policy ϕ. We define $\tilde{\phi} = \Gamma(\phi)$ as follows: $\tilde{\phi}_t(\tilde{h}_t) = \tilde{a}$ with $\tilde{a}(y) = \phi_t(y)$, for every history $\tilde{h}_t$. Thus to every policy ϕ belongs a policy $\tilde{\phi}$, independent of the history, which has as yth *component* the action belonging to *observation* y under ϕ.

This completely defines the second problem and the relation between the two problems.

Theorem 3.1 *For the problems* $(\mathcal{X}, \mathcal{Y}, \mathcal{A}, r, q, c)$, $(\tilde{\mathcal{X}}, \tilde{\mathcal{A}}, \tilde{r}, \tilde{q}, \tilde{c})$, *and the function* Γ *as defined above the conditions i) and ii) of Theorem 2.1 are satisfied, and therefore an equivalent problem with full observation exists.*

Proof First note that, for an arbitrary policy $\tilde{\phi}$, in view of the transition structure, the history $\hat{h}_t = (p_1, \tilde{a}_1, \ldots, p_{t-1}, \tilde{a}_{t-1}, p_t)$ with $\tilde{a}_s$ such that $\tilde{a}_s = \tilde{\phi}_s(\hat{h}_s)$ and p_{s+1} such that

$\tilde{r}(p_s, \tilde{a}_s, p_{s+1}) = 1$ occurs a.s.

We show that, for $\tilde{\phi} = \Gamma(\phi)$, $p_t(x) = P(X_t = x)$ for all x and t. Because of our choice of initial distribution it holds for $t = 1$. Given $p_t(x) = P(X_t = x)$ for some t,

$$P(X_{t+1} = x') = \sum_{x,y} P(X_t = x) q(x,y) r(x, \phi_t(y), x') =$$

$$\sum_{x,y} p_t(x) q(x,y) r(x, \tilde{a}_t(y), x') = p_{t+1}(x'),$$

according to the definition of Γ and $\tilde{r}$. Therefore

$$Ec(X_t, A_t) = \sum_{x,y} p(x) q(x,y) c(x, \phi_t(y)) = \sum_{x,y} p(x) q(x,y) c(x, \tilde{a}_t(y)) = E\tilde{c}(\tilde{X}_t, \tilde{A}_t),$$

giving i).

To prove ii) note that we observed that for an arbitrary $\tilde{\phi}$ there is a single path with non-zero probability of occurring, with histories $\hat{h}_t$. The values of $\tilde{\phi}_t$ for other values of $\tilde{h}_t$ have no influence on $\tilde{H}_T$. This defines the equivalence relation. As a representative for each class take the policy with $\tilde{\phi}_t$ constant, i.e., with $\tilde{\phi}_t(\tilde{h}_t) = \tilde{\phi}_t(\hat{h}_t)$ for all $\tilde{h}_t$. It is clear that ϕ for which $\Gamma(\phi) = \tilde{\phi}$ is the policy with $\phi_t(y) = \tilde{a}(y)$, where $\tilde{a} = \tilde{\phi}_t$ (note that $\tilde{\phi}_t$ is constant). This proves ii). □

Remark 3.2 Theorem 3.1 shows that a *separation principle* [9, 8] holds: based on $\phi_1, \ldots, \phi_{t-1}$ (but not on the actual observations) the state distribution at t is computed (the estimation part), and based on this distribution the optimal control is computed (the control part). The crucial difference with the model with recall in the next Section is that there the estimation is based also on the actual observations before t.

Remark 3.3 Note the crucial role that the initial state plays; the equivalence relation depends strongly on it. Most solution methods solve full observation problems for each initial state. Here we are only interested in a single initial state (being the initial distribution in the original problem). If we are also interested in other initial states, the various policies within an equivalence class get a meaning too. It also illustrates the dependence of the optimal policy on the initial distribution.

4. Partial observation information pattern

In the second model ϕ_t is allowed to depend on all previous observations $y_1, \ldots, y_t$. Note that we could let ϕ_t also depend on the previous actions, but as we can reconstruct all actions taken from the old observations, there is no need in doing so.

The result states that the distribution on the states, conditional on the previous observations and actions, gives the states for the second problem. At each stage this distribution is updated in a Bayesian manner (this in contrast with the no recall model).

This result is well known, see for example the textbooks Kumar & Varaiya [5] (p. 84–87) or Bertsekas [2] (p. 98–102); these results are based on the work of Striebel [8] and Witsenhausen [9].

Let us define the second model, formulating the well known result. As state space we take $\tilde{\mathcal{X}} = \mathcal{P}$ (defined in Section 3 as the set of distributions on $\mathcal{X}$), the action set remains the same, thus $\tilde{\mathcal{A}} = \mathcal{A}$. The transitions are as follows: $\tilde{r}(p,a,p') = \sum_{x,x'} p(x)r(x,a,x')q(x',y)$, for p' such that $p'(x') = \sum_x p(x)r(x,a,x')q(x',y) \Big/ \sum_{x,x'} p(x)r(x,a,x')q(x',y)$. This p' is notated as $T(p,y)$. In the case that for fixed (p,a) there are different y and y' such that $T(p,y) = T(p,y')$, then the probabilities in the definition of $\tilde{r}$ should be summed, or, alternatively, we could define $\tilde{r}$ as $\tilde{r}(p,a,p') = \sum_{x,x',y} p(x)r(x,a,x')q(x',y)\mathrm{I}\{T(p,y) = p'\}$ (with $\mathrm{I}\{\cdot\}$ the indicator function). As direct costs we have $\tilde{c}(p,a) = \sum_x p(x)c(x,a)$, and as initial distribution p_1 we take $X_1|Y_1$.

The interpretation of the variables is as follows: p should be considered as the state distribution, based on all observations, just after an observation. Then an action a is chosen, the system changes state and a new observation occurs. These two steps are reflected in the transition function $\tilde{r}$.

Theorem 4.1 *For the problems $(\mathcal{X},\mathcal{Y},\mathcal{A},r,q,c)$, $(\tilde{\mathcal{X}},\tilde{\mathcal{A}},\tilde{r},\tilde{q},\tilde{c})$, and the function Γ as defined above the conditions i) and ii) of Theorem 2.1 are satisfied, and therefore an equivalent problem with full observation exists.*

Proof Unfortunately we cannot follow right away the proof of Theorem 3.1, because it can occur that different histories result in the same distribution, for example because different observations may result in the same conditional distribution of the state given past observations. We cannot replace the for the controller identical observations by one, as it can be the case that there is a difference in other states. Thus different histories for the partial observation model can lead to the same history for the full observation model. This gives us problems defining Γ. Therefore we introduce a third problem, based on the full observation of distributions, but which states contain also the observation. Now we can define Γ, and it is easily shown that adding the observation in this way to the second problem does not give any additional information, i.e., the value functions of the second and the third model are the same.

The third problem is defined by $(\hat{\mathcal{X}},\hat{\mathcal{A}},\hat{r},\hat{q},\hat{c})$. Here $\hat{\mathcal{X}} = \tilde{\mathcal{X}} \times \mathcal{Y}$, $\hat{\mathcal{A}} = \mathcal{A}$, $\hat{r}(\hat{x},a,\hat{x}') = \tilde{r}(p,a,p')$ with $\hat{x} = (p,y)$ for some y and $\hat{x}' = (p',y')$ with y' the observation occurring in the definition of $\tilde{r}(p,a,p')$. Finally $\hat{c}(\hat{x},a) = \tilde{c}(p,a)$ for $\hat{x} = (p,y)$ and all y.

Writing $\hat{V}^t(\hat{x})$ and $\tilde{V}^t(p)$ for the corresponding dynamic programming value functions, it can be proven easily by induction that $\hat{V}^t(\hat{x}) = \tilde{V}^t(p)$ for $\hat{x} = (p,y)$, because the transitions and costs do not depend on y. Also the equivalence between the optimal policies for both models is easily seen. Thus it remains to prove that the first and the third problem are equivalent.

Thus we define $\hat{\phi} = \Gamma(\phi)$ as follows: for the history $\hat{h}_t = (\hat{x}_1,\ldots,\hat{x}_t)$ with $\hat{x}_s = (p_s,y_s)$ take $\hat{\phi}_t(\hat{h}_t) = \phi_t(h_t)$ with $h_t = (y_1,\ldots,y_t)$.

Fix ϕ and $\hat{\phi} = \Gamma(\phi)$. For a given history $h_t = (y_1,\ldots,y_t)$, there is a history $\hat{h}_t = (\hat{x}_1,\ldots,\hat{x}_t)$ with $\hat{x}_s = (p_s,y_s)$ which occurs with probability 1, namely the one where p_{s+1} is the distribution resulting from a transition from p_s, while observing y_s, and taking action $a = \phi_s(h_s)$.

Next we show that, for h_t and $\hat{h}_t$ as above, $P(H_t = h_t) = P(\hat{H}_t = \hat{h}_t)$ and that $p_t(x) = P(X_t = x \mid H_t = h_t)$. Assume that both hold for t. Then, with $a = \phi_t(h_t)$,

$$P(H_{t+1} = h_{t+1} \mid H_t = h_t) = P(Y_{t+1} = y_{t+1} \mid H_t = h_t) =$$
$$\sum_x P(X_t = x \mid H_t = h_t) P(Y_{t+1} = y_{t+1} \mid X_t = x, H_t = h_t) =$$
$$\sum_{x,x'} p_t(x) r(x, a, x') q(x', y_{t+1}) = \hat{r}(\hat{x}_t, a, \hat{x}_{t+1}),$$

for $\hat{x}_{t+1} = (p_{t+1}, y_{t+1})$, with p_{t+1} the distribution corresponding to the observation y_{t+1}. That X_{t+1} given $H_{t+1} = h_{t+1}$ has the conditional distribution p_{t+1} follows from

$$P(X_{t+1} = x' \mid H_{t+1} = h_{t+1}) =$$
$$\frac{P(X_{t+1} = x' \mid H_t = h_t) P(Y_{t+1} = y_{t+1} \mid X_{t+1} = x', H_t = h_t)}{\sum_{x'} P(X_{t+1} = x' \mid H_t = h_t) P(Y_{t+1} = y_{t+1} \mid X_{t+1} = x', H_t = h_t)} =$$
$$\frac{\sum_x P(X_t = x \mid H_t = h_t) P(X_{t+1} = x' \mid X_t = x, H_t = h_t) q(x', y_{t+1})}{\sum_{x,x'} P(X_t = x \mid H_t = h_t) P(X_{t+1} = x' \mid X_t = x, H_t = h_t) q(x', y_{t+1})} =$$
$$\frac{\sum_x p_t(x) r(x, a, x') q(x', y_{t+1})}{\sum_{x,x'} p_t(x) r(x, a, x') q(x', y_{t+1})}.$$

Now we can verify i) and ii). It follows that

$$Ec(X_t, A_t) = \sum_{h_t} P(H_t = h_t) E[c(X_t, A_t) \mid H_t = h_t] =$$
$$\sum_{h_t} P(H_t = h_t) \sum_x p_t(x) c(x, \phi_t(h_t)) =$$
$$\sum_{h_t} P(\hat{H}_t = \hat{h}_t) \sum_x p_t(x) c(x, \hat{\phi}_t(\hat{h}_t)) = E\hat{c}(\hat{X}_t, \hat{A}_t),$$

giving i). Also ii) follows easily. We already saw that for each h_t there was only $\hat{h}_t$ which could occur. Thus we can take policies $\hat{\phi} \in \hat{\Phi}$ which only depend on the $\mathcal{Y}$-component as representatives of their classes. The corresponding $\phi \in \Phi'$ takes exactly the same actions. This proves ii). □

Remark 4.2 As in the previous Section we have the separation of estimation and control, but in the current case the estimation is also based on the past observations.

5. An example

We compute the optimal policies for both information patterns for a simple two-state model, which is as follows. Consider a system which can be either up or down. Whether or not the system is up or down cannot be observed directly, but if the system is up this is observed with probability q. This models for example a machine that detects rare particles: when a particle is detected this means that the system is functioning, but when no particle is detected this can be due to the fact that the machine is down or because there is no particle to be detected.

When the system is up, it fails with probability r. At each moment in time we can choose to repair the system (action 1, with direct costs 1) or leave it as it is (action 0, costs 0). For each unit of time that the system is up we earn a reward R. Costs and rewards are discounted with a factor β.

This system can be modeled as a stochastic control problem with the following parameters:

$\mathcal{X} = \{0,1\}$ (with 0 the down state);

$\mathcal{Y} = \{0,1\}$ (observation 1 means a particle detected);

$\mathcal{A} = \{0,1\}$;

$r(0,0,0) = r(0,1,1) = r(1,1,1) = 1$, $r(1,0,0) = 1 - r(1,0,1) = r$;

$q(0,0) = 1$, $1 - q(1,0) = q(1,1) = q$;

$c(0,0) = 0$, $c(0,1) = 1$, $c(1,0) = -R$, $c(1,1) = 1 - R$.

Using the results presented in this paper we can formulate the equivalent dynamic programming problem. We computed the optimal policy by approximating the continuous state space ($\tilde{\mathcal{X}} = [0,1]$) by a discrete one ($\{0, 1/1000, 2/1000, \ldots, 1\}$). For this model we computed the discounted optimal policy. In the following table we present some of the results. State p is the probability that the system is down (i.e., in state 0).

q	0	0.1	0.25	0.5	0.75	1
no recall	-7.8065	-7.8549	-7.9307	-8.0984	-8.3615	-9.1324
partial obs.	-7.8065	-7.9448	-8.1137	-8.3659	-8.5903	-9.1388

Table. Minimal discounted costs for starting state 1 (the up state), $r = 0.05$, $R = 1$, $\beta = 0.9$ and varying q.

Some interesting phenomena can be derived from the table. First we see that the costs are decreasing in q. This is not surprising: q can be seen as the amount of information available to the controller, and costs decrease as the amount of information increases. When comparing both models it is seen that are the costs for the partial observation case are lower than for the no recall case. This is also easily explained, as the no recall policy is allowable in the partial observations model. For $q = 0$, there are no observations, and both methods coincide. For $q = 1$ there are complete observations, and thus the optimal policy depends only on the current observation. Here the optimal policies are also equal, the difference in the numbers in the tables can be explained by a combination of the discrete approximation and the difference in the methods for computing the optimal policies.

Let us now take a closer look at the optimal policy, for the case from the table with $q = 0.5$. For the no recall case the actions are two-dimensional, the first (second) entry being the action taken if no (a) particle is detected. The action $(0,0)$ (thus no repair under both observations) is taken in the states $\{0, \frac{1}{1000}, \ldots, \frac{212}{1000}\}$, and $(1,0)$ (thus a repair if no particle is detected) is

chosen in $\{\frac{213}{1000},\ldots,1\}$, the optimal policy is thus of threshold type. Starting from state 0, we finally end up in the following cycle, with the actions written on top of the arrow:

$$\frac{19}{1000} \overset{(0,0)}{\longrightarrow} \frac{115}{1000} \overset{(0,0)}{\longrightarrow} \frac{159}{1000} \overset{(0,0)}{\longrightarrow} \frac{201}{1000} \overset{(0,0)}{\longrightarrow} \frac{241}{1000} \overset{(1,0)}{\longrightarrow} \frac{19}{1000}.$$

The fact that this cycle does not return to 0 can be explained by the fact that if the system is in state 1, it is not repaired and can fail therefore. Thus the optimal policy repairs the system every 5th time unit, unless it is observed to be operating at that moment.

The optimal policy for the model with recall is also of threshold type. If 1 is observed, repair is never undertaken. If 0 is observed, action 0 is chosen up to $\frac{461}{1000}$, and from $\frac{462}{1000}$ on action 1 is chosen. Describing the policy however is somewhat more complicated. If 1 is observed, the system returns to 0. If 0 is observed then, starting from 0, state $\frac{639}{1000}$ is reached in 4 steps, and action 1 is taken. Thus the optimal policy repairs if it has observed 0 four times in a row.

Structural results In Albright [1] structural properties for partially observed stochastic control problems with two states are studied. Unfortunately, his results cannot be used to explain the optimality of the threshold policies found above: his condition in Theorem 3 that $p_{22}(a) - p_{12}(a)$ is isotone is not valid for the current model. (Indeed, on p. 1050 of [1] it is stated that this condition is equivalent to "more maintenance is felt more in the good state than in the bad state". That is clearly not the case in our model: maintenance is felt in state 0 (the bad state) and not in state 1.)

However, due to the simplicity of our model a proof of the optimality of threshold policies for the model with recall is easily given. The dynamic programming equation has the following form:

$$\tilde{V}^t(p) = \min\Big\{ -(1-p)R + \beta[1-(1-p)(1-r)q]\tilde{V}^{t+1}(p') + \beta(1-p)(1-r)q\tilde{V}^{t+1}(0),$$

$$1-(1-p)R + \beta\tilde{V}^{t+1}(0)\Big\},$$

for $p' = [T(p,0)](0)$, with $T(p,0)$ the a posteriori distribution after observing 0, as defined in Section 4. To show that a threshold policy is optimal it suffices to show inductively that $\tilde{V}^t(p) \leq \tilde{V}^t(p')$ if $p \leq p'$. Indeed, from this and $[T(p,0)](0) \leq [T(p',0)](0)$ it follows that

$$\Big(-(1-p)R + \beta[1-(1-p)(1-r)q]\tilde{V}^t([T(p,0)](0)) + \beta(1-p)(1-r)q\tilde{V}^t(0)\Big)$$

$$-\Big(1-(1-p)R + \beta\tilde{V}^t(0)\Big)$$

is increasing in p, which gives the optimality of a threshold policy. Proving $\tilde{V}^t(p) \leq \tilde{V}^t(p')$ is straightforward.

For the model without recall the situation is more complicated due to the fact that there are four possible actions, and because after action $(1,0)$ the system does not return to state 1. For this case we were unable to derive structural results.

6. Conclusion

A technique which transforms stochastic control problems with partial observations into equivalent full observation problems is formulated and applied to two specific models, the model with and the model without recall of observations. The results are applied to a simple system with two states and two possible observations. The optimal policy for the system without recall has a remarkably simple cyclic structure.

The technique is also applicable to models with delayed and/or decentralized information. This is subject of ongoing research.

Acknowledgements This work was initiated at CWI (Amsterdam), while working with Jan van Schuppen. I thank him for introducing me to this field of research and for his extensive comments on earlier versions of this paper.

References

[1] S.C. Albright. Structural results for partially observable Markov decision processes. *Operations Research*, 27:1041–1053, 1979.

[2] D.P. Bertsekas. *Dynamic Programming.* Prentice-Hall, 1987.

[3] A. Hordijk, G.M. Koole, and J.A. Loeve. Analysis of a customer assignment model with no state information. *Probability in the Engineering and Informational Sciences*, 8:419–429, 1994.

[4] A. Hordijk and J.A. Loeve. Undiscounted Markov decision chains with partial information; an algorithm for computing a locally optimal periodic policy. *ZOR - Mathematical Methods of Operations Research*, 40:163–181, 1994.

[5] P.R. Kumar and P. Varaiya. *Stochastic Systems.* Prentice-Hall, 1986.

[6] J.A. Loeve. *Markov Decision Chains with Partial Information.* PhD thesis, Leiden University, 1995.

[7] M. Schäl. Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. *ZOR - Mathematical Methods of Operations Research*, 32:179–196, 1975.

[8] C. Striebel. Sufficient statistics in the optimum control of stochastic systems. *Journal of Mathematical Analysis and Applications*, 12:576–592, 1965.

[9] H.S. Witsenhausen. Separation of estimation and control for discrete time systems. *Proceedings of the IEEE*, 59:1557–1566, 1971.