MATHEMATICAL CENTRE TRACTS 95

# BAYESIAN CONTROL OF MARKOV CHAINS

K.M. VAN HEE

MATHEMATISCH CENTRUM AMSTERDAM 1978

AMS (MOS) Classification scheme (1970): primary 90C40, 62C10, 49D99,
secondary ~~62C~~10

ISBN: 90 6196 163 7

CONTENTS


| | | |
|---|---|---|
| | Acknowledgement | VII |
| 1. | INTRODUCTION | |
| 1.1 | Historical perspective | 1 |
| 1.2 | Informal description of the model | 4 |
| 1.3 | Summary of the following chapters | 8 |
| 1.4 | Notations, conventions and prerequisites | 10 |
| 2. | THE MODEL AND THE PROCESS OF POSTERIOR DISTRIBUTIONS | |
| 2.1 | The Bayesian control model | 19 |
| 2.2 | Posterior distributions | 26 |
| 2.3 | Limit behaviour of the posterior distributions | 33 |
| 3. | THE EQUIVALENT DYNAMIC PROGRAM AND OPTIMAL REWARD OPERATORS | |
| 3.1 | Transformation into a dynamic program | 43 |
| 3.2 | A class of optimal reward operators | 53 |
| 3.3 | Miscellaneous results for the Bayesian control model | 65 |
| 4. | BAYESIAN EQUIVALENT RULES AND THE AVERAGE-RETURN CRITERION | |
| 4.1 | Bayesian equivalent rules and other approaches | 71 |
| 4.2 | Optimal strategies for the average-return criterion | 74 |
| 5. | BAYESIAN EQUIVALENT RULES AND THE TOTAL-RETURN CRITERION | |
| 5.1 | Preliminaries and the independent case | 91 |
| 5.2 | Linear system with quadratic costs | 98 |
| 5.3 | A simple inventory control model | 108 |
| 6. | APPROXIMATIONS | |
| 6.1 | Bounds on the value function and successive approximations | 121 |
| 6.2 | Discretizations | 138 |
| 7. | COMPUTATIONAL ASPECTS AND EXAMPLES | |
| 7.1 | Algorithm for models where I is a singleton | 149 |
| 7.2 | Algorithm for models with known transition law except for one state | 157 |
| 7.3 | Numerical examples | 161 |

APPENDIX A. RESULTS FROM ANALYSIS 175

APPENDIX B. REMARKS ON THE MINIMAX CRITERION 179

REFERENCES 183

LIST OF SYMBOLS 189

*Acknowledgement*

This monograph contains a part of the results of my research, carried out at the department of mathematics of the Eindhoven University of Technology, during the years 1974 until 1978.
I am very grateful to my thesis advisor prof. dr. Jaap Wessels for his encouragement and many stimulating discussions, and to my co-promotor dr. Fred Steutel for his very careful reading of the manuscript which led to many improvements.
To prof. dr. Arie Hordijk I wish to express my gratitude for arousing my enthusiasm for Bayesian decision theory and dynamic programming. I am indebted to dr. Fred Simons for his support to the measure theoretical problems that appeared during my research.
Further I would like to thank prof. dr. Jaap Wessels and my colleagues in his group dr. Luuk Groenewegen, dr. Jo van Nunen, Jan van der Wal and dr. Jacob Wijngaard for creating an excellent working atmosphere.
I thank Ruth Kool for his computing assistance and Lieke Janson and Marèse Wolfs for their excellent typing of the manuscript.
Finally I am indebted to the Mathematical Centre for the opportunity to publish this monograph in their series: Mathematical Centre Tracts, and to all those at the Mathematical Centre who have contributed to its technical realization.

# 1. *INTRODUCTION*

In this monograph we study the control of Markov chains with incompletely known transition law. The Bayes criterion, which is used explains the name of the monograph. We start this chapter with a short historical overview of the problem field (section 1.1), In section 1.2 we give an informal description of the model we are dealing with.
Then we summarize the contents of the following chapters (section 1.3). We conclude this chapter with a summary of notations and prerequisites (section 1.4).

## 1.1 *Historical perspective*

After A. Wald founded (statistical) sequential analysis, it was R. Bellman who recognized that the technique of backward induction, which is frequently used in sequential analysis, is also applicable to a wide range of non-statistical sequential decision problems (cf. [Wald (1947)], [Bellman (1957)]). Bellman formalized the technique and called it *dynamic programming*. In [Howard (1960)] the first extensive treatment is found on the relations between dynamic programming and the control of Markov chains. Independently, in [Shapley (1953)] sequential control problems concerning Markov chains are studied, using a game theoretic formulation. Later on, in [Blackwell (1965)] and [Derman (1966)] the results of Howard are refined and extended for the criterion of *expected total rewards* and the criterion of *expected average rewards*, respectively. Blackwell and Derman started an explosive development of the theory of control of Markov chains.
Before enclosing the problem field we first specify what is meant by a *dynamic program* or a *Markov decision process*. A dynamic program is a system that is determined by a *state space*, an *action space*, a *reward function* and a *transition law*, such that for each pair (state, action) a probability distribution on the state space is specified. At discrete points in time, called *moments* or *stages*, the *controller* or *decision maker* chooses an action from the action space. Then, according to the transition law, the system moves to a new state and an immediate reward is obtained, depending on the state before the transition, on the action itself and on the new state. A recipe for choosing an action at each stage, is called a *strategy*.
To apply the results of dynamic programming in practice, one has to know the transition law. Unfortunately it seldom happens that these probability distributions are known. So the controller has to estimate the transition

law during the course of the process. Therefore, apart from the control problem, there is an estimation problem.

From now on we assume that the transition law depends on an unknown *parameter*, which belongs to some *parameter set*. Therefore the expected return at each stage depends on the unknown parameter and so we have to choose a criterion to measure the return at each stage. In literature the *Bayes criterion* is mainly used (cf. section 1.2 for a definition). The first attempts in the field of dynamic programming with an incompletely known transition law have been made by Bellman (see [Bellman (1961)]). He used the term *adaptive control* of Markov chains. Bellman noticed that, if the Bayes criterion is used, the problem can be transformed into an equivalent dynamic program with a completely known transition law and with a state space which is the Cartesian product of the original one and the set of all probability distributions on the parameter set. This transformation is also suggested in [Shiryaev (1964)], [Dynkin (1965)] and [Aoki (1967)] for models, which allow unobservability of the states, and in [Wessels (1971),(1972)]. In [Hinderer (1970)] the first systematic proof is given for the case that the state and action spaces are both countable, and afterwards in [Rieder (1972), (1975)] the transformation is given for complete separable metric state and action spaces. In fact it is shown that, for the Bayes criterion, the *posterior distributions* of the unknown parameter are *sufficient statistics*. In [Wessels (1968)], among other things, the problem of sufficient statistics is studied in connection with several other criteria, such as the *minimax criterion*. Almost all other authors considered only the Bayes criterion and studied the equivalent dynamic program, mentioned above. In [Martin (1967)], [Rieder (1972)], [Satia and Lave (1973)], and [Waldmann (1976)] the method of successive approximations for the equivalent dynamic program is studied. Only Satia and Lave tried to exploit the special structure of this dynamic program. In [Fox and Rolph (1973)], [Mandl (1974), (1976)], and in [Rose (1975)] optimal strategies are constructed for the criterion of expected average return. Here it is possible to construct strategies which are at least as good as all other strategies, for all parameter values, hence it is not necessary to work with the Bayes criterion or anything like it.

Special models, arising in control theory are studied in [Sworder (1966)] and [Aoki (1967)]. Inventory control models with an incompletely known demand distribution are studied in [Scarf (1959)], [Iglehart (1964)], [Wessels (1971), (1972)], [Rieder (1972)], [Zacks and Fennel (1973)] and in

[Waldmann (1976)]. A number of other problems can be found in the literature. The most famous one is the two-armed bandit problem. We will return to most of the contributions of the above-mentioned authors in the other chapters of this monograph. The number of publications in the field of dynamic programming with an incompletely known transition law is very small compared with the overwhelming amount of literature on dynamic programming with a known transition law.

We conclude this section with a sketch of the problems we examine in this monograph. We choose the Bayes criterion too. From a mathematical point of view this criterion has the advantage, as compared with the minimax criterion, that the model can be transformed into the so-called equivalent dynamic program. Further it has the nice property that the decision maker may express his opinion on the importance of the various parameter values, which characterize the unknown transition law, by a weight function. Even if the model with known transition law has finite state and action spaces, the equivalent dynamic program has a state space which is *essentially* infinite. However, the method of successive approximations to determine the optimal expected total return is workable, since in order to determine the n-th approximation we have to consider all possible paths through n stages of which there are a finite number, if the state and action spaces are finite. The effort needed to obtain good approximations proved to be very large in the studies of Martin and Satia and Lave (in [Martin (1967)] examples with only two states and two actions turn out to be very time-consuming and in [Satia and Lave (1973)] examples with four states and two actions are considered to be of "moderate-size"). One of the objectives of our study is to show that the method of successive approximations can be applied successfully to rather large models, that have a suitable parameter structure. Our analysis is based on the construction of special scrap-vectors for the successive approximation method and on the exploitation of the convergence of the posterior distributions. We note that some results of our analysis are also interesting for the problem of *robustness* of the model under variations in the parameter value. In section 1.3 we specify the approximation methods we advocate, in an informal way.

Another objective of our study is to show that there are easy-to-handle optimal strategies for maximizing the average expected return, and also for some practical examples of our model for maximizing the expected total return. At the end of section 1.2 we consider this matter in more detail.

## 1.2 *Informal description of the model*

We start this section with a motivation of the choice of the model we study in this monograph: the *Bayesian control model*.

Consider a dynamic program with finite state and action spaces. It sometimes happens that a transition is affected by a random variable which is observable for the decision maker, but the value of which cannot be reconstructed from the state values of the process. For example consider a waiting-line model in discrete time, where $Y_{n+1}$ is the number of arrivals in the time period $[n,n+1)$ and where $X_n$ is the number of customers in the system at time n. Then it is obvious that the value of $Y_{n+1}$ is not determined by $X_n$ and $X_{n+1}$, if the number of services completed in each time interval is random. If the distribution of the random variable $Y_n$ is incompletely known, then it is useful to keep this random variable as a *supplementary state variable*. Confining ourselves to the state values of the original process only, means that we throw away information concerning the transition law. In our model we assume that for each state and action the transition may be affected by a random variable, the value of which is observed by the decision maker immediately *after* the transition. The value of this random variable is obtained by a random drawing from a distribution, depending only on the actual state and action. There are at most countably many different distributions from which is sampled. Further we assume that only these distributions are incompletely known. We call these random variables *supplementary state variables*. In case the transition, for some state and action, is not affected by a supplementary state variable we may consider the next state variable itself as a supplementary state variable. We return to this point in chapter 2.

We now continue with the model description. For simplicity, we assume here that all considered sets are finite. Let the state space be denoted by X and the action space by A. Further let the random variables $X_n$ and $A_n$ denote the state and action at stage n, respectively. The transition to state $X_{n+1}$, given $X_n$ and $A_n$ is also affected by the outcome of the supplementary state variable $Y_{n+1}$ which is observed at stage n + 1 and which takes on values in the set Y. This works in the following way. The conditional probability of $X_{n+1}$, given $X_n = x$, $A_n = a$ and $Y_{n+1} = y$, is

$$\mathbb{P}[X_{n+1} = x' \mid X_n = x\ ,\ A_n = a\ ,\ Y_{n+1} = y] = P(x'|x,a,y)$$

where the function P is assumed to be known.
However, the random variables $Y_{n+1}$, $X_n$ and $A_n$ are dependent, while the conditional distribution of $Y_{n+1}$, given $X_n$ and $A_n$ depends on some unknown parameter $\theta$, which belongs to a given parameter set $\Theta$, i.e. we have

$$\mathbb{P}_\theta[Y_{n+1} = y \mid X_n = x,\ A_n = a] = \sum_{i \in I} 1_{K_i}(x,a)\ p_i(y|\theta)$$

where $\{K_i,\ i \in I\}$ is a partition of $X \times A$, and I is some index set. Hence the distribution in the set $\{p_i(\cdot|\theta),\ i \in I\}$ from which the random variable $Y_{n+1}$ is sampled depends on the state and action at stage n. Further, if $X_n = x$ , $A_n = a$ and $Y_{n+1} = y$ there is an immediate, possibly negative, reward: $r(x,a,y)$.
Although the model may seem to be rather artificial, there are many well-known models which fit into this framework. For example, inventory control models, where $X_n$ is the inventory level at time n and $Y_{n+1}$ is the demand during the interval $[n,n+1)$. Here we always sample $Y_n$ from the same distribution, hence I is a singleton. Also the ordinary dynamic program with finite state and action spaces and all transition probabilities unknown, is included in our model. We return to this matter in chapter 2.
We note that, if the parameter $\theta$ is known, we are dealing with a dynamic program with state space X, action space A, transition law:

$$\mathbb{P}[X_{n+1} = x' \mid X_n = x,\ A_n = a] = \tilde{P}(x'|x,a) := \sum_{i \in I} 1_{K_i}(x,a) \sum_{y \in Y} P(x'|x,a,y)p_i(y|\theta),$$

and reward function:

$$\tilde{r}(x,a) := \sum_{i \in I} 1_{K_i}(x,a) \sum_{y \in Y} p_i(y|\theta)r(x,a,y)\ .$$

In this monograph X,Y,A and $\Theta$ are complete separable metric spaces, but the index set I is at most countable. Hence we do not allow more than countably many unknown distributions $p_i(\cdot|\theta)$, $i \in I$ and $\theta \in \Theta$.
A strategy $\pi$ is a procedure which chooses at each stage n an action, based on the *history* of the process, i.e. $X_0, A_0, Y_1, X_1, A_1, \ldots, Y_n, X_n$.
Each strategy $\pi$, each parameter value $\theta$, and each starting state x together determine a probability on the *sample space* of the process . The expectation with respect to this probability of the immediate reward at stage n is denoted by:

$$\mathbb{E}^{\pi}_{x,\theta}[r(X_n,A_n,Y_{n+1})] \ .$$

The *expected total discounted return* $v(x,\theta,\pi)$ is:

$$v(x,\theta,\pi) := \mathbb{E}^{\pi}_{x,\theta}[\sum_{n=0}^{\infty} \beta^n r(X_n,A_n,Y_{n+1})]$$

where $\beta \in [0,1)$ is called the *discount factor*.
Only in trivial situations there is a strategy $\pi^*$ such that $v(x,\theta,\pi^*) \geq v(x,\theta,\pi)$ for all $x \in X$, $\theta \in \Theta$ and all strategies $\pi$. So it is unwise to use this as a criterion for a strategy to be optimal. Criteria for which there are always (nearly) optimal strategies, are the already mentioned minimax and Bayes criteria. A strategy $\pi^*$ is called $\varepsilon$-*optimal*, $\varepsilon \geq 0$, for the minimax criterion, if

$$\min_{\theta\in\Theta} v(x,\theta,\pi^*) \geq \min_{\theta\in\Theta} v(x,\theta,\pi) - \varepsilon \quad \text{for all } x \in X,\ \theta \in \Theta$$

and all strategies $\pi$. We do not use this criterion. In appendix B we consider an example, which shows that the use of this criterion has some odd implications.
We use the Bayes criterion. So, we fix some probability distribution $q$ on the parameter set $\Theta$ and we call a strategy $\pi^*$ $\varepsilon$-*optimal*, $\varepsilon \geq 0$, if

$$\sum_{\theta\in\Theta} q(\theta)v(x,\theta,\pi^*) \geq \sum_{\theta\in\Theta} q(\theta)v(x,\theta,\pi) - \varepsilon$$

for all $x \in X$ and all strategies $\pi$. If a strategy is 0-optimal we call it *optimal*. We note again that the so-called *prior distribution* $q$ can be considered as a weight function, expressing the importance of the various parameter values in the opinion of the decision maker.
In chapter 4 we consider the *average expected return* instead of the expected total discounted return. We call a strategy $\pi^*$ $\varepsilon$-*optimal*, $\varepsilon \geq 0$, with respect to this criterion, if

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{\theta\in\Theta} q(\theta) \sum_{n=0}^{N-1} \mathbb{E}^{\pi^*}_{x,\theta}[r(X_n,A_n,Y_{n+1})] \geq$$

$$\geq \liminf_{N\to\infty} \frac{1}{N} \sum_{\theta\in\Theta} q(\theta) \sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,\theta}[r(X_n,A_n,Y_{n+1})] - \varepsilon$$

for all $x \in X$ and all strategies $\pi$ (again, a 0-optimal strategy is called *optimal*).
The Bayes criterion allows us to consider another interpretation of the Bayesian control model. In this interpretation we consider the unknown para-

meter as a random variable with distribution q. The *posterior distributions* of this random variable, or in other words the conditional distributions of this random variable, given the history of the process,play an important role in this monograph. It is well-known that the name of Bayes is connected with the criterion since he suggested to consider the unknown parameter of a distribution as a random variable itself in statistical inference. It turns out that the Bayesian control model is equivalent to a dynamic program with a known transition law and with a compound state space $X \times W$, where W is the set of all probability distributions on $\theta$. For each starting state and each strategy, we are dealing with a stochastic process $(X_n, Q_n, A_n)$ where $Q_n$ is the actual posterior distribution of the random variable that represents the parameter.

It is desirable to have good strategies that are easy to handle, i.e. to have a formula or a simple recipe which yields an action as a function of the actual state $x \in X$ and the actual posterior distribution $q \in W$. A way of deriving easy to handle strategies is based on the following idea. If the parameter is known to be $\theta$ and if there is an optimal strategy then an optimal action in state $x \in X$ often is a maximizer of $F(x,\theta,\cdot)$ where $F : X \times \Theta \times A \to \mathbb{R}$. Note that the action depends on the parameter $\theta$ and that the function F is assumed to be known. Now let the parameter be unknown. Then we may use an action a which maximizes the function $a \to \int q(d\theta)F(x,\theta,a)$ if the actual state is x and the actual posterior distribution is q (provided that integration is possible and the maximum exists). Such a rule is called a *Bayesian equivalent rule*. It will be proved that such a rule yields an optimal strategy, if we are maximizing the average expected return, under conditions which guarantee that in the long run the decision maker obtains enough information about the unknown parameter, i.e. the sets $K_i$ have to be recurrent. For maximizing the expected discounted total return we do not know a Bayesian equivalent rule that is optimal in general, however for some special models, such as the linear system with quadratic cost and a simple inventory control model, there is an optimal Bayesian equivalent rule. For the linear system with quadratic cost this rule can be considered as a generalization of the well-known certainty equivalent rule.

## 1.3 *Summary of the following chapters*

In chapter 2 we start with a formal description of the Bayesian control model and we consider some examples. Then we study the process of posterior distributions. The main result is the convergence of the posterior distributions to a degenerate distribution, under each strategy which assures the number of visits to each set $K_i$, $i \in I$ to be infinite, with probability one. This result is used in several places in chapters 4 and 6.

In chapter 3 we deal with two rather technical points. First we show that the Bayesian control model is equivalent to a dynamic program (see section 1.2) and after that we study a class of optimal reward operators for dynamic programs in general. Here we consider optimal reward operators based on stopping times, for dynamic programs as introduced by Wessels (cf. [Van Nunen and Wessels (1977)]). We generalize the operators for dynamic programs with complete separable state and action space and we derive some new properties of these operators. These operators determine the maximal expected total return until some stopping time, and with a terminal reward at the stopping time, depending on the state at the time. Successive applications of these operators yield a sequence of functions on the state space, which converges to the function of optimal values. We use these operators in chapter 6 where we consider the method of successive approximations for the equivalent dynamic program.

In chapter 4 we first introduce the Bayesian equivalent rules. Then we construct optimal strategies in order to maximize the average expected reward.

Chapter 5 is devoted to the study of optimal strategies for the expected total-return criterion. For three examples of our model we show that a Bayesian equivalent rule provides an optimal strategy. The first example we call the *independent case* since the rewards are independent of the state, i.e. r is constant in the first coordinate. In all examples it is assumed that the index set I is a singleton, so the random variables $Y_n$, $n \in \mathbb{N}$ are sampled from the same (unknown) distribution at each stage. The second example is the linear system with quadratic cost and the last one is a simple inventory control model. For this inventory model the Bayesian equivalent rule is not always optimal. However, we give an upper bound for the loss we incur by using this rule when it is not optimal.

In chapter 6 we consider approximations for the "function of optimal values" when maximizing the expected discounted total return. This function is called

the *value function* and is defined on X × W by:

$$v(x,q) := \sup_{\pi} \sum_{\theta} q(\theta)v(x,\theta,\pi)$$

where the supremum is taken over all strategies. We first indicate an upper bound on v and several lower bounds. These bounds have simple interpretations and are computable if the parameter set is finite or equivalently, if the prior distribution is concentrated on a finite set. We study the use of these bounds for successive approximations of the value function. We also give a lower bound on the expected discounted total return if a special Bayesian equivalent rule is used and we construct an other easy-to-handle strategy which is not a Bayesian equivalent rule but which behaves nicely. Further we specialize the parameter structure as follows: there is a subset B of the state space X with the property that, if $X_n \in B$ then $Y_{n+1}$ is sampled from the same unknown distribution for all actions chosen, for $X_n \in X \setminus B$ the distribution of $Y_{n+1}$ is known (hence $K_1 = B \times A$ and $\theta \to p_i(\cdot|\theta)$ is constant for $i \neq 1$). A special example of this structure arises in the model where B = X, e.g. the models studied in chapter 5. Here we use an optimal reward operator as studied in chapter 3, with the entrance time in the set B as stopping time. In fact, this operator allows us to consider the process which is embedded on the set B. For this parameter structure we use the convergence of the posterior distributions to a degenerate distribution, and also the upper and lower bounds, to compute in advance an error estimate on the n-th successive approximation, starting with a fixed prior distribution. If the error estimate for the n-th approximation is small enough, then we may compute the value function for this prior distribution by backward induction. The effort needed for the computation of the n-th error estimate is small compared with the backward induction procedure. Since usually the computed quantities to determine the n-th approximation cannot be used to compute the n + 1-st approximation, it is nice to know in advance whether the n-th approximation is sufficiently accurate.

We also consider in this chapter another type of approximations, namely *discretizations* of the parameter set. Here we split up the parameter set into a finite partition, and in each set of the partition we choose a representative point. We give bounds for the error caused by replacing the given prior distribution q by the discrete prior distribution which attributes probabilities to the representative points equal to the given probabilities of the corresponding sets. In [Fox (1973)] and [Whitt (1976)] also discre-

tizations of dynamic programs are studied. To apply their method, we would have to split up the set of all distributions on the parameter set into a finite partition and, in the equivalent dynamic program, the process would then jump between representative points in these partition sets. However, we then loose the nice property that the second state-coordinate of the process (i.e. $Q_n$) is the posterior distribution of the unknown parameter, at every stage.

Our discretizations are of interest, since in general we can compute the upper and lower bounds, mentioned above, only if the prior distribution is concentrated on a finite set of parameters. As a byproduct of our analysis of discretizations we obtain a bound for the difference between the value function of the Bayesian control model and the model that is obtained by replacing in advance the distributions $p_i(\cdot|\theta)$ by their Bayes estimates based on the prior distribution and considering these estimates as the true distributions. This last model is used very frequently in practice, instead of the Bayesian model.

Finally, in chapter 7 we construct algorithms, based on the approximations of chapter 6, which compute the value function $v(x,q)$ for a fixed prior distribution, and which also determine $\varepsilon$-optimal strategies. We illustrate the quality of the algorithms by numerical data for some examples.

In appendix A we collect some results of measure theory which are used in chapter 3. In appendix B we illustrate the odd implications of the minimax criterion by an example.

We note that it is possible to start reading at chapter 4 after reading the model description in chapter 2 and the assertions of the theorems and corollaries of chapters 2 and 3.

## 1.4 *Notations, conventions and prerequisites*

We start with some conventions. A *numbered sentence* indicates a definition a result or a formula. Such a sentence may occupy several lines, each one of which is indicated by an indentation. Symbols used for objects, which are defined in a numbered sentence have a *global* meaning, i.e. if we use a symbol without defining it in the theorem proof, example or comment where it is used, then it has the meaning given in the numbered sentence where it is defined. References to lemmas, theorems, corollaries, examples, sections and chapters are preceded by the words "lemma","theorem", etc. Each chapter has its own numbering, for example 2.4 is the fourth numbered sentence in

chapter 2. References to appendix A are preceded by the capital: A. The end of a proof is indicated by: □. If there is no ambiguity concerning the domain of some index or variable, we omit the domain in the notations.

We continue with a list of notations.

1.1 $\mathbb{N} := \{0,1,2,\ldots\}$, $\overline{\mathbb{N}} := \mathbb{N} \cup \{\infty\}$, $\mathbb{N}^* := \{1,2,3,\ldots\}$, $\overline{\mathbb{N}}^* := \mathbb{N}^* \cup \{\infty\}$.

1.2 $\mathbb{R}$ is the set of real numbers, $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty,\infty\}$.

1.3 $\delta(\cdot,\cdot)$ is the Kronecker symbol, i.e. $\delta(i,i) = 1$ and $\delta(i,j) = 0$ if $i \neq j$.

1.4 $\#$ A is the cardinality of the set A.

1.5 $x^+ := \max(x,0)$, $x^- := -\min(x,0)$.

1.6 Let $(X_i,\mathcal{X}_i)$ be measurable spaces for $i \in I$, where I is a countable set then $X := \prod_{i\in I} X_i$ is the Cartesian product and $\mathcal{X} := \bigotimes_{i\in I} \mathcal{X}_i$ the product-σ-field on X. If $\mu_i$ is a probability on $\mathcal{X}_i$ then $\mu := \bigotimes_{i\in I} \mu_i$ is the product measure on $\mathcal{X}$, if I is finite and $\mu_i$ a σ-finite measure on $\mathcal{X}_i$ then μ is also the product measure on $\mathcal{X}$.

Let A, X and Y be sets, such that $A \subset X \times Y$ then

1.7 $\text{proj}_X(A) := \{x \in X \mid \text{there is some } y \in Y \text{ with } (x,y) \in A\}$.

1.8 i.i.d. means "independent and identically distributed", iff means "if and only if" and a.s. means "almost surely".

Let $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ be measurable spaces and let $f : X \to Y$ be measurable then

1.9 $\sigma(f)$ is the sub-σ-field of $\mathcal{X}$ induced by f, i.e.

$$\sigma(f) := \{A \in \mathcal{X} \mid A = f^{-1}(B),\ B \in \mathcal{Y}\}, \text{ where } f^{-1}(B) := \{x \in X \mid f(x) \in B\}.$$

1.10 $\mathcal{P}(\mathcal{X})$ is the set of all probabilities on a measurable space $(X,\mathcal{X})$.

Let f be a function on a set X then

1.11 $x \to f(x)$, $x \in X$ is a notation for this function.

1.12 $\emptyset$ is the empty set.

Let $x_1,x_2,x_3,\ldots$ be a sequence of real numbers, then

1.13 $\inf\{x_i \mid i \in \emptyset\} := \infty$, $\sum_{i\in\emptyset} x_i := 0$ and $\prod_{i\in\emptyset} x_i := 1$ .

Let $(X,\mathcal{X})$ be a measurable space and let $q$ be a measure on $\mathcal{X}$ and $f$ a non-negative Borel measurable function on $(X,\mathcal{X})$, then

1.14 $f(x)q(dx)$ is a notation for the measure $\mu$ defined by

$$\mu(A) := \int_A f(x)q(dx), \quad A \in \mathcal{X}.$$

1.15 Let $f$ and $g$ be functions on some set $X$ with range $\overline{\mathbb{R}}$ and let $y \in \overline{\mathbb{R}}$, $f \leq g$ if and only if $f(x) \leq g(x)$ for all $x \in X$, $f \leq y$ if and only if $f(x) \leq y$ for all $x \in X$. The analogous convention is used if $\leq$ is replaced by $<$, $\geq$, $>$ or $=$ .

We continue with some pertinent facts on transition probabilities and conditional expectations. Let $(\Omega,\mathcal{F},\mathbb{P})$ be a probability space, $(A,\mathcal{A})$ a measurable space, and let $Y : \Omega \to A$ be measurable. Then we call $Y$ a *random variable* and we write

1.16 (i) $\mathbb{P}[Y \in B] := \mathbb{P}[\{\omega \in \Omega \mid Y(\omega) \in B\}]$ , $B \in \mathcal{A}$ .

(ii) $\mathbb{E}[Y] := \int Y(\omega)\,\mathbb{P}(d\omega)$ .

A real-valued function on $\Omega$ is called *$\mathcal{F}$-measurable* or simply measurable, if it is measurable with respect to the Borel $\sigma$-field on $\mathbb{R}$. The following lemma is well-known (cf. [Bauer (1968) lemma 55.1]).

**Lemma 1.1**

Let $(\Omega,\mathcal{F})$ and $(A,\mathcal{A})$ be measurable spaces, and let $f : \Omega \to A$ be measurable. Then a real-valued function $g$ on $\Omega$ is $\sigma(f)$-measurable iff there is a real-valued measurable function $h$ on $A$ such that $g = h(f)$. If $f$ is a surjection then the function $h$ is unique.

1.17 A measurable space $(A,\mathcal{A})$ is called *Borel space* if $A$ is a non-empty Borel subset of a complete separable metric space and $\mathcal{A}$ is the Borel $\sigma$-field on $A$ (note that in [Hinderer (1970) page 187] such a space is called a standard Borel space and in [Blackwell (1965)] a Borel set).

1.18 The topological product of at most countably many Borel spaces which, because of the separability of the spaces, coincides with the measure theoretic product, is again a Borel space (cf.[Parthasarathy (1967) p. 135]).

Let $(\Omega,F)$ and $(A,\mathcal{A})$ be measurable spaces, then a function $P$ from $A \times \Omega$ to $[0,1]$ is called a *transition probability from* $(\Omega,F)$ *to* $(A,\mathcal{A})$, or simply from $\Omega$ to $A$, if

1.19 (i) $P(B|\cdot)$ is $F$-measurable for each $B \in \mathcal{A}$.

(ii) $P(\cdot|\omega)$ is a probability on $\mathcal{A}$, for each $\omega \in \Omega$.

Let $(\Omega,F,\mathbb{P})$ be a probability space, let $\mathcal{B}$ be a sub-$\sigma$-field of $F$ and let $X$ be a real-valued measurable function on $\Omega$, with $\mathbb{E}[X^+] < \infty$ .

1.20 (i) The *conditional expectation* of $X$ given $\mathcal{B}$ is denoted by $\mathbb{E}[X|\mathcal{B}]$ and defined as a real-valued $\mathcal{B}$-measurable function on $\Omega$ such that $\mathbb{E}[X1_B] = \mathbb{E}[\mathbb{E}[X|\mathcal{B}]1_B]$ for all $B \in \mathcal{B}$.
(Here $1_B$ is the *indicator function* of the set $B$.)

(ii) If $Y$ is another a real-valued measurable function on $\Omega$ we define $\mathbb{E}[X|Y] := \mathbb{E}[X|\sigma(Y)]$ .

(iii) For every $A \in F$ we define the *conditional probability of* $A$ *given* $\mathcal{B}$, respectively the *conditional probability of* $A$ *given* $Y$ by $\mathbb{P}[A|\mathcal{B}] := \mathbb{E}[1_A|\mathcal{B}]$, respectively $\mathbb{P}[A|Y] := \mathbb{E}[1_A|Y]$.

Note that the conditional expectation is not uniquely defined, however two versions of it are equal $\mathbb{P}$-a.s.

Theorem 1.2

Let $(\Omega,F)$ be a Borel space and let $\mathbb{P}$ be a probability on $F$. Then for every sub-$\sigma$-field $\mathcal{B}$ of $F$ the conditional probability is *regular*, i.e. there exists a transition probability $P$ from $(\Omega,\mathcal{B})$ to $(\Omega,F)$ such that for every real-valued $F$-measurable function $X$ that is bounded from above, we have $\omega \to \int X(\tilde{\omega})P(d\tilde{\omega}|\omega)$ is a version of $\mathbb{E}[X|\mathcal{B}]$ . If $P'$ is another transition probability from $(\Omega,\mathcal{B})$ to $(\Omega,F)$ with this property, then

$$\mathbb{P}[\{\omega|P(\cdot|\omega) \neq P'(\cdot|\omega)\}] = 0 \ .$$

For a proof cf. [Bauer (1968) th. 56.5].

We sometimes need the following corollary of th. 1.2.

Corollary 1.3

Let $(\Omega,F)$ be a Borel space, let $\mathbb{P}$ be a probability on $F$, let $(A,\mathcal{A})$ be a measurable space and let $Y$ be a measurable map from $\Omega$ to $A$. The probability $Q$ on $\mathcal{A}$ is defined by $Q(B) := \mathbb{P}[Y \in B]$, $B \in \mathcal{A}$. Then there is a

transition probability P from $(A,\mathcal{A})$ to $(\Omega,\mathcal{F})$ such that

$$(*) \qquad \mathbb{P}[D \cap Y^{-1}(B)] = \int_B P(D|y)Q(dy) \quad \text{for all } B \in \mathcal{A} \text{ and } D \in \mathcal{F}.$$

If P' is another transition probability from $(A,\mathcal{A})$ to $(\Omega,\mathcal{F})$ with this property, then

$$Q[\{y \mid P(\cdot|y) \neq P'(\cdot|y)\}] = 0 .$$

P is called a *regular conditional probability given* $Y = y$ and we usually write $\mathbb{P}[\cdot|Y = y]$ for $P(\cdot|y)$.

**Proof.**

By th. 1.2 there is a transition probability $\tilde{P}$ from $(\Omega,\sigma(Y))$ to $(\Omega,\mathcal{F})$ such that for all $D \in \mathcal{F}$ and $B \in \mathcal{A}$:

$$\mathbb{P}[D \cap \{\omega|Y(\omega) \in B\}] = \int_{Y^{-1}(B)} \tilde{P}(D|\omega)\,\mathbb{P}(d\omega) .$$

By lemma 1.1 there is for each $D \in \mathcal{F}$ a real-valued measurable function on A, denoted by $P(D|\cdot)$ such that

$$P(D|Y(\omega)) = \tilde{P}(D|\omega) , \qquad \text{for } \omega \in \Omega.$$

It is easy to verify that P, considered as a function on $A \times \mathcal{F}$ is a transition probability from $(A,\mathcal{A})$ to $(\Omega,\mathcal{F})$ with property (*).
Let P' be another transition probability on $A \times \mathcal{F}$ with property (*), and define $N := \{y \in A|P(\cdot|y) \neq P'(\cdot|y)\}$. Then

$$Y^{-1}(N) = \{\omega \in \Omega|P(\cdot|Y(\omega)) \neq P'(\cdot|Y(\omega))\}.$$

By th. 1.2 $\mathbb{P}[Y^{-1}(N)] = 0$. Hence $Q[N] = 0$. □

Let the assumptions of corollary 1.3 hold and let X be a real-valued measurable function on $\Omega$, bounded from above. Then we define

$$1.21 \qquad \mathbb{E}[X|Y = y] := f(y) := \int X(\omega)\,\mathbb{P}[d\omega|Y = y] .$$

It is easy to verify that $f(Y)$ is a version of the conditional expectation of X given Y.

We frequently use the following theorem of Ionescu Tulcea (cf. [Neveu (1965) page 165]).

Theorem 1.4

Let $(X_n,\mathcal{X}_n)$ , $n \in \mathbb{N}$ be measurable spaces and let $Q_{n+1}$ be a transition probability from $(\Pi_{t=0}^{n} X_t, \otimes_{t=0}^{n} \mathcal{X}_t)$ to $(X_{n+1}, \mathcal{X}_{n+1})$, $n \in \mathbb{N}$. Further let $(X,\mathcal{X}) := (\Pi_{t=0}^{\infty} X_t, \otimes_{t=0}^{\infty} \mathcal{X}_t)$ and let $\xi_0,\xi_1,\ldots$ be the coordinate functions on X i.e. $\xi_n(x) := x_n$, $x = (x_0,x_1,\ldots) \in X$. Then

(i) for all $n \in \mathbb{N}$ there is an unique transition probability P from $(\Pi_{t=0}^{n} X_t, \otimes_{t=0}^{n} \mathcal{X}_t)$ to $(X,\mathcal{X})$ denoted by $P(B|x_0,\ldots,x_n)$, $B \in \mathcal{X}$, $x_i \in X_i$ $i = 0,\ldots,n$, such that for cylinder sets of the form $B := A_0 \times \ldots \times A_m \times X_{m+1} \times X_{m+2} \times \ldots$ and $m \geq n$:

$$P(B|x_0,\ldots,x_n) = 1_{A_1\times\ldots\times A_n}(x_0,\ldots,x_n)\int_{A_{n+1}} Q_{n+1}(dx_{n+1}|x_0,\ldots,x_n) \ldots$$

$$\ldots \int_{A_m} Q_m(dx_m|x_0,\ldots,x_{m-1}) .$$

(ii) for every probability $\rho$ on $X_0$ there is a unique probability $\mathbb{P}_\rho$ on $\mathcal{X}$ given by $\mathbb{P}_\rho[B] = \int_{X_0} \rho(dx_0)P(B|x_0)$ , $B \in \mathcal{X}$ and for any measurable function Y on X that is bounded from above, $\int P(dx|\xi_0,\ldots,\xi_n)Y(x)$ is a version of the conditional expectation of Y given the $\sigma$-field $\sigma(\xi_0,\ldots,\xi_n)$. Hence one may define: (cf. lemma 1.1)

$$\mathbb{E}_\rho[Y \mid \xi_0 = x_0,\ldots,\xi_n = x_n] := \int P(dx|x_0,\ldots,x_n)Y(x)$$

or

$$\mathbb{E}_\rho[Y \mid \xi_0,\ldots,\xi_n] := \int P(dx|\xi_0,\ldots,\xi_n)Y(x) .$$

Finally we summarize some pertinent facts concerning the set $P(X)$ of all probabilities on a Borel space $(X,\mathcal{X})$.

1.22 On $P(X)$ we have the *topology of weak convergence*; this is the

coarsest topology such that for functions $f \in C(X)$ the map $\mu \to \int f(x)\mu(dx)$ is continuous, $\mu \in P(X)$, where $C(X)$ is the set of bounded real-valued continuous functions on X (cf. [Parthasarathy (1967)]).

**Lemma 1.5**

Let $\mathcal{E}$ be the topology of weak convergence on $P(X)$ and $\mathcal{F}$ the $\sigma$-field generated by $\mathcal{E}$. Then $\mathcal{F}$ is also:

(i) the smallest $\sigma$-field such that the functions $\mu \to \mu(B)$ are measurable, $\mu \in P(X)$, $B \in \mathcal{X}$.

(ii) the smallest $\sigma$-field such that the functions $\mu \to \int f(x)\mu(dx)$ are measurable, $\mu \in P(X)$, $f \in C(X)$.

The proof of statement (i) can be found in [Rieder (1975) lemma 6.1]. Note that this implies that $\mathcal{F}$ is also the smallest $\sigma$-field such that $\mu \to \int f d\mu$, $\mu \in P(X)$ are measurable for all real-valued bounded measurable functions f on X.

Proof of statement (ii). Let $\mathcal{B}$ be the smallest $\sigma$-field in $P(X)$ such that $\mu \to \int f(x)\mu(dx)$ is measurable, for $f \in C(X)$. For each Borel subset $D \subset \mathbb{R}$ and every $f \in C(X)$ we have $\{\mu \mid \int f(x)\mu(dx) \in D\} \in \mathcal{B}$, for $f \in C(X)$. This is true in particular for all open sets of $\mathbb{R}$. Hence the topology $\mathcal{E}$ is contained in $\mathcal{B}$, i.e. $\mathcal{F} \subset \mathcal{B}$. On the other hand, since for all open subsets $D \subset \mathbb{R}$ $\{\mu \mid \int f(x)\mu(dx) \in D\} \in \mathcal{F}$ and since the Borel $\sigma$-field on $\mathbb{R}$ is generated by the open sets, we have $\{\mu \mid \int f(x)\mu(dx) \in D\} \in \mathcal{F}$ for all Borel subsets $D \subset \mathbb{R}$. Hence $\mathcal{B} \subset \mathcal{F}$. □

In lemma 1.6 we collect some miscellaneous results.

**Lemma 1.6**

(i) Let $(X,\mathcal{X})$ be a Borel space and $\mathcal{F}$ the $\sigma$-field on $P(X)$, generated by the topology of weak convergence, then $(P(X),\mathcal{F})$ is a Borel space.

(ii) The identification of elements of X with the point measures in $P(X)$ is a homeomorphism.

(iii) Let $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ be Borel spaces and f a nonnegative measurable function on $X \times Y$, then the function

$$(x,q) \to \int f(x,y)q(dy), \qquad x \in X,\ q \in P(Y)$$

is measurable.

The proof of (i) is found in [Hinderer (1970) th. 12.13], the proof of part (ii) in [Parthasarathy (1967) lemma 6.1 page 42] and part (iii) is an immediate consequence of lemma 1.5 (i) (cf. [Rieder (1975) lemma 6.2]).

# 2. THE MODEL AND THE PROCESS OF POSTERIOR DISTRIBUTIONS

In section 2.1 we define the *Bayesian control model*, the model we study in this monograph, and we present some examples. In section 2.2 the posterior distributions of the random variable, which represents the unknown parameter, are defined and some properties are derived. Finally, in section 2.3 the limit behaviour of the posterior distributions is studied and also the differences of successive posterior distributions.

## 2.1 *The Bayesian control model*

Our model is similar to models described in [Shiryaev (1964), (1967)], [Dynkin (1965)], [Martin (1967)] and [Hinderer (1970)]. In fact, it is a special case of the model considered in [Rieder (1975)], which will be shown later on in this section. In this monograph several models are considered, which are special cases of the Bayesian control model we describe now.

model 1: *Bayesian control model*

The model consists of the following objects.

2.1 (a) $(X,\mathcal{X})$ a Borel space. $X$ is called the *state space*.

(b) $(Y,\mathcal{Y})$ a Borel space. $Y$ is called the *supplementary state space*.

(c) $(A,\mathcal{A})$ a Borel space. $A$ is called the *action space*.

(d) $D$, a function from $X$ to the non-empty subsets of $A$ such that $K := \{(x,a) \mid x \in X,\ a \in D(x)\}$ is an element of $\mathcal{X} \otimes \mathcal{A}$. $D(x)$ is called the set of *admissible actions* in state $x$. It is assumed that $K$ contains the graph of some measurable function from $X$ to $A$.

(e) $I$ is a countable set, called the *index set*.

(f) For all $i \in I$ there is a Borel space $(\theta_i,\mathcal{T}_i)$ and $\theta_i$ is called the *parameter space of index* $i$. The Borel space $(\theta,\mathcal{T})$ is defined by $\theta := \Pi_{i\in I}\ \theta_i$, $\mathcal{T} := \otimes_{i\in I}\ \mathcal{T}_i$. The set $\theta$ is called the *parameter space*.

(g) $\{K_i,\ i \in I\}$ is a measurable partition of $X \times A$.

(h) $P$ is a transition probability from $X \times A \times Y$ to $X$ (cf. 1.19).

(i) $\nu$ is a $\sigma$-finite measure on $\mathcal{Y}$. If $Y$ is countable then $\nu$ is assumed to be the counting measure.

(j) $p_i$ is a nonnegative measurable function on $Y \times \theta_i$, for all $i \in I$ such that $\int_Y p_i(y|\theta_i)\nu(dy) = 1$ for all $\theta_i \in \theta_i$ and $i \in I$.

For all $i \in I$ and $\theta_i, \tilde{\theta}_i \in \Theta_i$, $\theta_i \neq \tilde{\theta}_i$ we assume

$$\nu(\{y \in Y \mid p_i(y|\theta_i) \neq p_i(y|\tilde{\theta}_i)\}) > 0 .$$

This property is called: the *separation property*.

(k) $r$ is a real-valued measurable function on $X \times A \times Y$, bounded from above, and called the *reward function*.

We continue with some definitions which clarify the meaning of the objects defined in 2.1.

Each $\theta \in \Theta$ can be described by $\theta = (\theta_i)_{i \in I}$ where $\theta_i \in \Theta_i$ is called the i-th coordinate of $\theta$.

For each $\theta \in \Theta$ we define a *transition probability* $\bar{P}_\theta$ from $X \times A$ to $Y \times X$, by

$$2.2 \qquad \bar{P}_\theta(E \times F \mid x,a) := \sum_{i \in I} 1_{K_i}(x,a) \int_E \nu(dy)\, p_i(y|\theta_i) \int_F P(dx'|x,a,y)$$

where $E \in \mathcal{Y}$, $F \in \mathcal{X}$, $x \in X$, $a \in A$ and $\theta_i$ the i-th coordinate of $\theta \in \Theta$.

(Note that $\bar{P}_\theta$ satisfies all requirements for a transition probability (cf. 1.19)).

2.3 The set of *histories* $H_n$ at stage n is defined by

(i) $H_0 := X$, $H_n := X \times (A \times Y \times X)^n$, $\quad n \in \bar{\mathbb{N}}^*$.

(ii) $\mathcal{H}_n$ is the product-$\sigma$-field on $H_n$ induced by $\mathcal{X}$, $\mathcal{A}$ and $\mathcal{Y}$ for $n \in \bar{\mathbb{N}}$.

2.4 A *strategy* $\pi$ is a sequence: $\pi = (\pi_0, \pi_1, \ldots)$ where $\pi_n$ is a transition probability from $(H_n, \mathcal{H}_n)$ to $(A, \mathcal{A})$ such that
$\pi_n(\cdot|x_0,a_0,y_1,x_1,a_1,\ldots,y_n,x_n)$
is concentrated on the set $D(x_n)$. The set of all possible strategies is denoted by $\Pi$.

It is easy to verify, by the condition on K (cf. 2.1 (d)), that $\Pi$ is non-empty.

2.5 The *sample space* of the Bayesian control process is $\Omega := \Theta \times H_\infty$, and on $\Omega$ we have the product-$\sigma$-field $\mathcal{H} := \mathcal{T} \otimes \mathcal{H}_\infty$.

Note that $(\Theta, \mathcal{T})$ and $(\Omega, \mathcal{H})$ are Borel spaces (cf. 1.18). On $\Omega$ we define the *coordinate functions* Z, $X_n$, $Y_n$, $A_n$, $n \in \mathbb{N}$, also called *random variables:*

2.6 $Z(\omega) := \theta$, $X_n(\omega) := x_n$, $Y_n(\omega) := y_n$, $A_n(\omega) := a_n$ for $\omega = (\theta, x_0, a_0, y_1, x_1, a_1, \ldots) \in \Omega$.

According to the Ionescu Tulcea theorem (cf. th. 1.4) we have for each so-called *starting distribution* $\rho \in P(X)$, each so-called *prior distribution* $q \in P(T)$ and each strategy $\pi \in \Pi$, a *probability* $\mathbb{P}^{\pi}_{\rho,q}$ on $(\Omega, H)$, defined by

2.7 $$\mathbb{P}^{\pi}_{\rho,q}[Z \in B,\ X_0 \in C,\ A_0 \in D_0,\ (Y_1,X_1) \in E_1, \ldots, (Y_n,X_n) \in E_n] :=$$

$$\int_B q(d\theta) \int_C \rho(dx_0) \int_{D_0} \pi_0(da_0|x_0) \int_{E_1} \bar{P}_\theta(d(y_1,x_1)|x_0,a_0) \ \ldots$$

$$\ldots \int_{D_{n-1}} \pi_{n-1}(da_{n-1}|x_0,a_0,y_1,x_1,a_1,\ldots,y_{n-1},x_{n-1}) \int_{E_n} \bar{P}_\theta(d(y_n,x_n)|x_{n-1},a_{n-1})$$

where $B \in T$, $C \in X$, $D_n \in A$ and $E_n \in Y \otimes X$, $n \in \mathbb{N}$.

2.8 The *expectation* with respect to $\mathbb{P}^{\pi}_{\rho,q}$ is denoted by $\mathbb{E}^{\pi}_{\rho,q}$.

2.9 Define $W := P(T)$ and let $\mathcal{W}$ be the $\sigma$-field on $W$ generated by the weak topology (cf. 1.22).

We identify each $\theta \in \Theta$ with the element of $W$ which is *degenerate* in $\theta$, i.e. $\theta$ represents the probability that is concentrated on $\{\theta\}$. (By lemma 1.6(ii) this identification is a homeomorphism). And similarly we identify each $x \in X$ with the degenerate distribution in $P(X)$. Hence, for $\pi \in \Pi$, $x \in X$, $\theta \in \Theta$ the probability $\mathbb{P}^{\pi}_{x,\theta}$ is well-defined.
Using th. 1.4 and the identification we easily derive:

2.10 The conditional probability may be chosen as:

$$\mathbb{P}^{\pi}_{\rho,q}[\cdot|Z=\theta] = \mathbb{P}^{\pi}_{\rho,\theta}[\cdot]$$

or

$$\mathbb{P}^{\pi}_{\rho,q}[\cdot|Z] = \mathbb{P}^{\pi}_{\rho,Z}[\cdot] \ .$$

Note that the difference in these expressions is that the first one is a function on $\theta$, while the second one is a function on $\Omega$, depending on the first coordinate only.
Using 2.10 we find, for $B \in T$ and $C \in H_\infty$ :

2.11 $\mathbb{P}^{\pi}_{\rho,q}[Z \in B,\ (X_0,A_0,Y_1,X_1,A_1,\ldots) \in C] =$

$$\int_B q(d\theta)\,\mathbb{P}^{\pi}_{\rho,\theta}[(X_0,A_0,Y_1,X_1,A_1,\ldots) \in C] \ .$$

Further we define *criterion functions* for the discrimination of strategies.

2.12 (i) The *Bayesian discounted total return* v is a real-valued function on $X \times W \times \Pi$:

$$v(x,q,\pi) := \mathbb{E}^{\pi}_{x,q}\Big[\sum_{n=0}^{\infty} \beta^n r(X_n,A_n,Y_{n+1})\Big]$$

where $\beta \in [0,1)$ is the *discount factor*.

(ii) The *value function* v is a real-valued function on $X \times W$:

$$v(x,q) := \sup_{\pi\in\Pi} v(x,q,\pi).$$

Note that we use the symbol v for two different, but related functions, and note that we use the name "value function" only in connection with the discounted total return.

2.13 The *Bayesian average return* g is a real-valued function on $X \times W \times \Pi$:

$$g(x,q,\pi) := \liminf_{N\to\infty} \frac{1}{N}\,\mathbb{E}^{\pi}_{x,q}\Big[\sum_{n=0}^{N-1} r(X_n,A_n,Y_{n+1})\Big] \ .$$

Finally we define (nearly) optimal strategies. Let $\varepsilon \geq 0$.

2.14 (i) A strategy $\pi$ is called $\varepsilon$-*optimal for the total return criterion* in $x \in X$ and $q \in W$, if $v(x,q,\pi) \geq v(x,q) - \varepsilon$.

(ii) A strategy $\pi$ is called $\varepsilon$-*optimal for the average return criterion* in $x \in X$ and $q \in W$, if $g(x,q,\pi) \geq \sup_{\tilde{\pi}\in\Pi} g(x,q,\tilde{\pi}) - \varepsilon$ .

A 0-optimal strategy is simply called *optimal*.

Now the Bayesian control model has been described completely. Note that for each starting distribution $\rho \in P(X)$, each prior distribution $q \in W$ and each strategy $\pi \in \Pi$ the probability $\mathbb{P}^{\pi}_{\rho,q}$ and the stochastic process $(Z,X_0,A_0,Y_1,X_1,A_1,\ldots)$ are completely described. Only in chapter 4 we shall consider the average-return criterion, everywhere else we consider the total-return criterion.

The Bayesian control model is an example of the so-called *Bayesian decision model* studied in [Rieder (1975)]. This relationship is not used in our monograph. However, it simplifies comparisons of our results with the literature. To substantiate this we introduce the following notations.

2.15 (i) $S := Y \times X$ , $\mathcal{S} := \mathcal{Y} \otimes \mathcal{X}$ .

(ii) $P^*_\theta$ is a transition probability from $S \times A$ to $S$, defined by

$$P^*_\theta(E \times F \mid (x,y),a) := \bar{P}_\theta(E \times F \mid x,a) \quad \text{for all } y \in Y,\ E \in \mathcal{Y},\ F \in \mathcal{X} \text{ and } \theta \in \Theta.$$

(iii) $D^*$ is a function from $S$ to the non-empty subsets of $A$, such that $D^*((y,x)) := D(x)$ for all $y \in Y$.

(iv) $r^*$ is a real-valued function on $S \times A \times S$ defined by
$r^*((y,x),a,(y',x')) := r(x,a,y')$ , $x,x' \in X$, $a \in A$, $y,y' \in Y$.

The 8-tuple $((S,\mathcal{S}), (A,\mathcal{A}), D^*, (\Theta,\mathcal{T}), P^*_\theta, q, \rho^*, r^*)$, where $\rho^* \in P(S)$ and $q \in W$, satisfies all assumptions of the model of Rieder. Note that, in our model the starting distribution $\rho$ is specified only on $X$ and in Rieder's formulation of our model the starting distribution $\rho^*$ on $Y \times X$ is required. However, only the marginal distribution of $\rho^*$ on $X$ plays a role, since the transition probability $P^*_\theta$ has the property: $y \to P^*_\theta(B \mid (y,x),a)$ is constant, by 2.15(ii).

We conclude this section with some examples, illustrating the applicability of our model.

Example 2.1

If the parameter set $\Theta$ is a singleton, or equivalently if the prior distribution $q \in W$ is degenerate in $\theta \in \Theta$, the Bayesian control model is an ordinary dynamic program, with state space $(X,\mathcal{X})$, action space $(A,\mathcal{A})$ and transition probability $\tilde{P}_\theta$, given by

$$\tilde{P}_\theta(B \mid x,a) := \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y \mid \theta_i) \int_B P(dx' \mid x,a,y), \quad B \in \mathcal{X}$$

and reward function $\tilde{r}_\theta$:

$$\tilde{r}_\theta(x,a) := \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y \mid \theta_i),\ r(x,a,y) .$$

**Example 2.2**

Each dynamic program with countable state space $\tilde{X}$, countable action space $\tilde{A}$ and incompletely known transition probability $\tilde{P}$ from $\tilde{X} \times \tilde{A}$ to $\tilde{X}$ and real-valued reward function $\tilde{r}$ on $\tilde{X} \times \tilde{A}$ can be transformed into a Bayesian control model. To verify this define $X := \tilde{X}$, $\mathcal{X}$ is the power set of $\tilde{X}$, $A := \tilde{A}$, $\mathcal{A}$ is the power set of $\tilde{A}$, $Y := X$, $\mathcal{Y} := \mathcal{X}$ and $r(x,a,y) := \tilde{r}(x,a)$ for all $x \in X$, $a \in A$ and $y \in Y$. Further define $I := X \times A$, $K_i := \{i\}$, $i \in I$ and $\theta_i := P(X)$. Note that I is countable and that $(\theta_i, \mathcal{T}_i)$ is a Borel space if $\mathcal{T}_i$ is the $\sigma$-field on $\theta_i$ generated by the weak topology (cf. lemma 1.6). Finally define $P(\{x'\}|x,a,y) := \delta(x',y)$, $x,x' \in X$, $a \in A$, $y \in Y$ and $p_i(y|\theta_i) := \theta_i(\{y\})$, $y \in Y$, $\theta_i \in \theta_i$, $i \in I$.

It is straightforward to verify that all assumptions of 2.1 are satisfied. If, for some pair $x,a \in X \times A$, $\tilde{P}(\cdot|x,a)$ is known, then the marginal distribution on $\theta_{x,a}$ of $q \in W$ has to be degenerate in $\tilde{P}(\cdot|x,a)$. Similarly, if $\tilde{P}(\cdot|x,a)$ is unknown but belongs to some family of probabilities on $(X,\mathcal{X})$ then the marginal on $\theta_{x,a}$ of $q \in W$ has to be concentrated on this family. Consequently the models described in [Martin (1967)], [Wessels (1968)], [Rose (1975)] can be regarded as special cases of our model.

**Example 2.3**

The class of models considered here is specified by Euclidean spaces X, Y and A, and a measurable function F from $X \times A \times Y$ to X. The state $X_n$ at time n is a function of the action $A_{n-1}$ at time $n-1$, the state $X_{n-1}$ at time $n-1$, and a random variable $Y_n$ such that

$$X_n = F(X_{n-1}, A_{n-1}, Y_n) , \qquad n \in \mathbb{N}^*$$

where $X_n \in X$, $A_n \in A$ and $Y_n \in Y$. The random variables $\{Y_n, n \in \mathbb{N}^*\}$ are i.i.d. and cannot be *controlled* by the decision maker, however they can always be *observed* by him. For that reason the sequence $\{Y_n, n \in \mathbb{N}^*\}$ is called the *external process*. The external process can be considered as a nuisance process. It is assumed that the distribution of $Y_n$ is not completely known: $p(\cdot|\theta)$ is the probability density of $Y_n$ with respect to the $\sigma$-finite measure $\nu$ on Y for all $\theta \in \Theta$ where $(\Theta,\mathcal{T})$ is a Borel space. We also assume $\nu(\{y \in Y \mid p(y|\theta) \neq p(y|\tilde{\theta})\}) > 0$ for $\theta \neq \tilde{\theta}$. It is easy to transform these models into our framework. To this end let $P(\{F(x,a,y)\}|x,a,y) = 1$ for $x \in X$, $a \in A$ and $y \in Y$, and let $\mathcal{X}$ be the Borel $\sigma$-field on X, and let $\mathcal{A}$ and

$\mathcal{Y}$ be the Borel $\sigma$-fields on A and Y respectively. Further let I be a singleton, i.e. $I := \{1\}$ and $K_1 := X \times A$. At each stage $Y_n$ is sampled from the distribution with density $p(\cdot|\theta)$, $\theta \in \Theta$.

Let there be a reward function satisfying 2.1 (k). Then all conditions of 2.1 are satisfied.

Examples of this class are the *linear system* with unknown disturbance distribution as studied in [Aoki (1968)], and *inventory models* with unknown demand distribution with or without backlogging (in chapter 5 we study such a model extensively). Another example of these models is the *replacement model* with *additive damage* as considered in [Taylor (1975)] where the distribution of the so-called *shocks* is not completely known (in chapter 7 we consider this model too).

Example 2.4

A model that satisfies all conditions 2.1(a)- 2.1(j), but for which the reward function is not bounded from above, can sometimes be transformed into a model satisfying all conditions of 2.1. For this purpose we replace 2.1(k) by another condition which is due to Wessels (cf. [Wessels (1977]), who assumes the existence of a so-called *bounding function* b, i.e. a positive measurable function on X, and a positive number M such that for all $x \in X$, $a \in A$ and $y \in Y$:

$$\text{(i)} \quad \int P(dx'|x,a,y)b(x') \leq b(x)$$

$$\text{(ii)} \quad r(x,a,y) \leq Mb(x) \ .$$

We shall carry out this transformation for the case where X is countable. It is easy to extend the argument to the general case. Define:

$$P^*(x'|x,a,y) := P(\{x'\}|x,a,y)b(x')b(x)^{-1}$$

$$r^*(x,a,y) := r(x,a,y)b(x)^{-1} \ , \quad \text{for } x,x' \in X,\ a \in A,\ y \in Y.$$

As it may happen that $\sum_{x' \in X} P^*(\{x'\}|x,a,y) < 1$ we add a state $x^*$ to X and let $X^* := X \cup \{x^*\}$. Further we define for $x \in X$, $a \in A$ and $y \in Y$:

$$P^*(x^*|x^*,a,y) := 1, \quad P^*(x^*|x,a,y) := 1 - \sum_{x' \in X} P(\{x'\}|x,a,y) \ ,$$

$r^*(x^*,a,y) := 0$ and $b(x^*) := 1$.

Each strategy for the original model is also a strategy for the new model (except in state $x^*$). We denote the expectation for the transformed model by $\mathbb{E}^*$. Note that for $x_j \in X$, $a_j \in A$, $y_j \in Y$:

$$b(x_0)^{-1}\{\prod_{j=0}^{n-1} P(\{x_{j+1}\}|x_j,a_j,y_{j+1})\}b(x_n) =$$

$$= \prod_{j=0}^{n-1} \{P(\{x_{j+1}\}|x_j,a_j,y_{j+1})b(x_{j+1})b(x_j)^{-1}\} = \prod_{j=0}^{n-1} P^*(x_{j+1}|x_j,a_j,y_{j+1}) .$$

And therefore

$$b(x_0)^{-1}\{\prod_{j=0}^{n-1} P(\{x_{j+1}\}|x_j,a_j,y_{j+1})\}r(x_n,a_n,y_{n+1}) =$$

$$= \{\prod_{j=0}^{n-1} P^*(x_{j+1}|x_j,a_j,y_{j+1})\}r^*(x_n,a_n,y_{n+1}) .$$

Now it is straightforward to verify that for $x \in X$, $q \in W$ and $\pi \in \Pi$:

$$b(x)^{-1}\mathbb{E}^{\pi}_{x,q}[r(X_n,A_n,Y_{n+1})] = \mathbb{E}^{*\pi}_{x,q}[r^*(X_n,A_n,Y_{n+1})] .$$

This shows the equivalence of both models.

## 2.2 *Posterior distributions*

As already announced, the *posterior distribution* of the random variable Z, which represents the unknown parameter, plays an important role in this monograph. We define random variables on $(\Omega,H)$ with range the set W, the set of distributions on $(\theta,T)$ and afterwards we show that these random variables are versions of the conditional distribution of Z, given the observed histories of the process. This property justifies calling these random variables the *posterior distributions*.
We start with some definitions.

2.16 On $\Omega$ we define, for $i \in I$, the function $Z_i$: $Z_i(\omega) = \theta_i$ where $\omega = (\theta,x_0,a_0,y_1,x_1,a_1,\ldots) \in \Omega$ and where $\theta = (\theta_i)_{i\in I}$.

Hence $Z = (Z_i)_{i\in I}$ and we may interpret the random variable $Z_i$, $i \in I$ as the parameter of the distribution from which $Y_n$ is sampled, if

$(X_{n-1},A_{n-1}) \in K_i$.

On $\Omega$ we define, for $i \in I$, a sequence of *stopping times*, $\{\tau(i,n),\ n \in \mathbb{N}\}$:

2.17 $\tau(i,0)(\omega) := 0$, $\tau(i,n)(\omega) := \inf\{m \in \mathbb{N} \mid m > \tau(i,n-1)(\omega),\ (X_{m-1}(\omega),A_{m-1}(\omega)) \in K_i\}$

for $n \in \mathbb{N}^*$ and $\omega \in \Omega$.

Note that the n-th observation from the distribution determined by $p_i(\cdot \mid \theta_i)$, $\theta_i \in \Theta_i$ occurs at stage $\tau(i,n)$ and note also that for each $\omega \in \Omega$ and each $k \in \mathbb{N}^*$ there is exactly one pair $(i,n)$, with $i \in I$, $n \in \mathbb{N}^*$ such that

$$\tau(i,n)(\omega) = k .$$

In the rest of this chapter the sub-$\sigma$-fields in $H$, induced by the observable random variables, are used frequently, therefore we introduce the notation:

2.18 $$F_n := \sigma(X_0,A_0,Y_1,X_1,A_1,\ldots,Y_n,X_n,A_n) , \qquad n \in \overline{\mathbb{N}} .$$

For the stopping times $\tau(i,n)$ we define the usual $\sigma$-fields $F_{\tau(i,n)}$:

2.19 $$F_{\tau(i,n)} := \{B \in H \mid B \cap \{\tau(i,n) = k\} \in F_k \text{ for all } k \in \mathbb{N}\}.$$

Note that $\{\tau(i,n) = k\} \in F_{k-1}$ for $n,k \in \mathbb{N}^*$.

Since $(\theta,\mathcal{T})$ is a product space we define, for each $q \in W$ the *marginal distributions* $q_i$ on $(\theta_i,\mathcal{T}_i)$, for $i \in I$:

2.20 Let $B \in \mathcal{T}_i$ then

$$q_i(B) := \int_{\{\theta \mid \theta_i \in B\}} q(d\theta) .$$

It seems to be quite natural to work with prior distributions q that are product-measures, i.e. $q = \bigotimes_{i\in I} q_i$. However most results of this monograph are valid without this assumption. Note that the assumption that $q = \bigotimes_{i\in I} q_i$ is equivalent with the assumption that $Z_i$, $i \in I$ are independent. In th. 2.1 we return to this matter.

In order to define the posterior distributions we define, for $n \in \mathbb{N}^*$, the functions $\alpha_n$ on $\Omega$ with range the set of measures on the parameter space $(\theta,\mathcal{T})$ and for $i \in I$ the random variables $\alpha_{i,n}$ on $\Omega$ with range the set of measures on $(\theta_i,\mathcal{T}_i)$:

2.21 (i) $$\alpha_n(B) := \int_B \prod_{j=1}^{n} \sum_{i\in I} 1_{K_i}(X_{j-1},A_{j-1}) p_i(Y_j \mid \theta_i) q(d\theta) .$$

$$\text{(ii)}\quad \alpha_{i,n}(B_i) := \int_{B_i} \prod_{j=1}^{n} \{1_{K_i}(X_{j-1},A_{j-1})p_i(Y_j|\theta_i) + 1 - 1_{K_i}(X_{j-1},A_{j-1})\}q_i(d\theta_i)$$

where $B \in \mathcal{T}$, $B_i \in \mathcal{T}_i$ and $\theta_i$ the i-th coordinate of $\theta \in \Theta$ (for notational convenience we have omitted the dependence on $\omega \in \Omega$ in 2.21).

The integrand of 2.21 (i) may be considered as the *likelihood function* of the parameter $\theta$ at time n and similarly the integrand of 2.21 (ii) as the likelihood function of the parameter-coordinate $\theta_i$, at time n.
The following equality clarifies this. It is easy to verify that on $\Omega$ we have

$$2.22 \qquad \prod_{j=1}^{n} \{1_{K_i}(X_{j-1},A_{j-1})p_i(Y_j|\theta_i)+1-1_{K_i}(X_{j-1},A_{j-1})\} =$$

$$= \prod_{\{k>0 \mid \tau(i,k)\le n\}} p_i(Y_{\tau(i,k)}|\theta_i)\ , \qquad i \in I\ .$$

We shall use the convention:

2.23 For any real-valued function f on Y and a stopping time $\tau$

$$f(Y_{\tau(\omega)}(\omega)) := 0 \quad \text{if} \quad \tau(\omega) = \infty, \quad \text{for } \omega \in \Omega.$$

Finally, we are ready to define the *posterior distribution* $Q_n$ for the *prior distribution* $q \in W$, as a random variable on $\Omega$ with range the set W:

2.24 Let $B \in \mathcal{T}$, then $Q_o(B) := q(B)$ and for $\omega \in \Omega$ and $n \in \mathbb{N}^*$:

$$Q_n(B)(\omega) := \alpha_n(B)(\omega)\{\alpha_n(\theta)(\omega)\}^{-1}\ , \qquad \text{if } \alpha_n(\theta)(\omega) > 0$$

$$:= q(B) \text{ otherwise.}$$

(In th. 2.1 it turns out that $\alpha_n(\theta) > 0$, $\mathbb{P}^{\pi}_{\rho,q}$-a.s.).
And similarly we define the *posterior distributions* $Q_{i,n}$ for $i \in I$, $n \in \mathbb{N}$ :

2.25 Let $B \in \mathcal{T}_i$, then $Q_{i,0}(B) := q_i(B)$ and for $\omega \in \Omega$ and $n \in \mathbb{N}^*$

$$Q_{i,n}(B)(\omega) := \alpha_{i,n}(B)(\omega)\{\alpha_{i,n}(\theta_i)(\omega)\}^{-1}\ , \qquad \text{if } \alpha_{i,n}(\theta_i)(\omega) > 0$$

$$:= q_i(B) \text{ otherwise.}$$

Note that $Q_n(\cdot)(\omega)$ and $Q_{i,n}(\cdot)(\omega)$ are probabilities for all $\omega \in \Omega$.
The measurability of $Q_n$ and $Q_{i,n}$ is a direct consequence of lemma 1.5 (i).
The name "posterior distribution" is justified in th. 2.1.

In th. 2.1 we collect some obvious properties of the random variables $Q_n$ and $Q_{i,n}$. Throughout this chapter we fix a starting distribution $\rho \in P(X)$, a prior distribution $q \in W$ and a strategy $\pi \in \Pi$, and for notational convenience we write $\mathbb{P}$ and $\mathbb{E}$ instead of $\mathbb{P}^{\pi}_{\rho,q}$ and $\mathbb{E}^{\pi}_{\rho,q}$.

Theorem 2.1

Let $B \in \mathcal{T}$ and $B_i \in \mathcal{T}_i$, for $i \in I$. Then:

(i) $\mathbb{P}[Z \in B | \mathcal{F}_n] = Q_n(B)$ , $\mathbb{P}$-a.s.

(ii) if $q = \bigotimes_{i \in I} q_i$ then $Q_n = \bigotimes_{i \in I} Q_{i,n}$ .

(iii) if $q = \bigotimes_{i \in I} q_i$ then $\mathbb{P}[Z_i \in B_i | \mathcal{F}_n] = Q_{i,n}(B_i)$, $\mathbb{P}$-a.s.

(iv) if $q = \bigotimes_{i \in I} q_i$ then $\mathbb{P}[Z_i \in B_i | \mathcal{F}_{\tau(i,n)}] = Q_{i,\tau(i,n)}(B_i)$ on $\{\tau(i,n) < \infty\}$ $\mathbb{P}$-a.s.

(v) $$Q_{n+1}(B) = \sum_{i \in I} 1_{K_i}(X_n, A_n) \frac{\int_B p_i(Y_{n+1} | \theta_i) Q_n(d\theta)}{\int_\theta p_i(Y_{n+1} | \theta_i) Q_n(d\theta)}$$

(on the subset of $\Omega$ where the denominator is positive).

(vi) $\mathbb{E}[Q_n(B) | \mathcal{F}_m] = Q_m(B)$ if $n > m$, $\mathbb{P}$-a.s.

Proof.

Let $C := \theta \times E_0 \times F_0 \times D_1 \times E_1 \times F_1 \times \ldots \times D_n \times E_n \times F_n \times (Y \times X \times A)^{\mathbb{N}}$ where $D_i \in \mathcal{Y}$, $E_i \in \mathcal{X}$ and $F_i \in \mathcal{A}$ for $i \in \mathbb{N}$. Then $C \in \mathcal{F}_n$ and

$$\int_C \mathbb{P}[Z \in B | \mathcal{F}_n] d\mathbb{P} = \mathbb{P}[Z \in B, X_0 \in E_0, A_0 \in F_0, \ldots, Y_n \in D_n, X_n \in E_n, A_n \in F_n] =$$

$$= \int_B q(d\theta) \int_{E_0} \rho(dx_0) \int_{F_0} \pi_0(da_0 | x_0) \int_{D_1} \nu(dy_1) \int_{E_1} P(dx_1 | x_0, a_0, y_1) \ldots$$

We introduce some notations which are useful in the following chapters.
On $Y \times W$ we define real-valued functions $p_i$, $i \in I$:

2.27 $$p_i(y,q) := \int q(d\theta) p_i(y|\theta_i) \qquad y \in Y,\ q \in W.$$

Notice that these functions may be considered as extensions of the functions defined in 2.1(j), by the embedding of $\theta$ in W, in fact $p_i(y,\theta) = p_i(y|\theta_i)$, $\theta \in \Theta$.
It is a consequence of lemma 1.6(iii) that the function $(y,q) \to p_i(y,q)$ is measurable, $i \in I$.
Note that $p_i(\cdot,q)$ is a probability density with respect to the measure $\nu$, for $i \in I$ and $q \in W$.
Further we define functions $T_i$ on $Y \times W$ with range the set W, for $i \in I$:

2.28 $T_{i,y}(q)$ is a probability on $\Theta$ such that, for $B \in \Theta$

$$T_{i,y}(q)(B) := \int_B p_i(y|\theta_i) q(d\theta) \{p_i(y,q)\}^{-1}, \quad \text{if } p_i(y,q) > 0$$

$$:= q(B), \quad \text{otherwise.}$$

Again, using lemma 1.6(iii) we find the measurability of

$$(y,q) \to T_{i,y}(q)(B), \qquad B \in \mathcal{T},\ i \in I.$$

Hence $T_i$ is a transition probability from $Y \times W$ to $\Theta$. We may interpret $T_{i,y}(q)$ as the posterior distribution, if q is the prior distribution and $y \in Y$ is observed from the distribution belonging to the set $K_i$. The following formula is easily verified:

2.29 $$Q_{n+1} = \sum_{i \in I} 1_{K_i}(X_n, A_n) T_{i,Y_{n+1}}(Q_n).$$

For $q \in W$ we define the functions $q_0, q_1, q_2, \ldots$ recursively:

2.30 (i) $q_0 : W \to W$, $q_0(q) := q$.

(ii) $q_n : W \times (X \times A \times Y)^n \to W$, $n \in \mathbb{N}^*$

such that

$$q_n(q,x_0,a_0,\ldots,y_n) := \sum_{i\in I} 1_{K_i}(x_{n-1},a_{n-1})\; T_{i,y_n}(q_{n-1}(q,x_0,a_0,\ldots,y_{n-1}))$$

for $q \in W$, $x_\ell \in X$, $y_\ell \in Y$ and $a_\ell \in A$, $\ell \in \mathbb{N}$.

It is straightforward to verify that, for $B \in \mathcal{T}$

$$2.31 \qquad q_n(q,x_0,a_0,\ldots,y_n)(B) = \frac{\int_B \prod_{j=0}^{n-1} \{\sum_{i\in I} 1_{K_i}(x_j,a_j) p_i(y_{j+1}|\theta_i)\} q(d\theta)}{\int_\Theta \prod_{j=0}^{n-1} \{\sum_{i\in I} 1_{K_i}(x_j,a_j) p_i(y_{j+1}|\theta_i)\} q(d\theta)}$$

provided that the denominator is positive.

Using the notations above, we may write:

$$Q_n = q_n(Q_0,X_0,A_0,\ldots,Y_n) \quad , \quad \text{on } \Omega \ .$$

We conclude this section with a few remarks:

Remarks.

(i) If $\xi$ and $\eta$ are independent random variables, defined on some probability space, and $\mathcal{B}$ is a sub-$\sigma$-field, then in general

$$\mathbb{E}[\xi\cdot\eta|\mathcal{B}] \neq \mathbb{E}[\xi|\mathcal{B}]\mathbb{E}[\eta|\mathcal{B}] \ .$$

However in th. 2.1(ii) we proved that equality holds $\mathbb{P}$-a.s., if $\xi := f(Z_i)$, $\eta := g(Z_j)$ and $\mathcal{B} := \mathcal{F}_n$ for $i \neq j$, $i,j, \in I$, $f$ and $g$ non-negative measurable functions on $\theta_i$ and $\theta_j$ respectively, and if $q = \bigotimes_{i\in I} q_i$.

(ii) Instead of defining the posterior distributions $Q_n$ by 2.25 we could define them directly as conditional distributions (cf. th. 2.1(i)). However the conditional distribution $\mathbb{P}[Z \in \cdot|\mathcal{F}_n]$ is undetermined on some set with $\mathbb{P}$-measure zero.

(iii) If the prior distribution $q$ is concentrated on a set of finitely many points then all posterior distributions are concentrated on this set.

## 2.3 *Limit behaviour of the posterior distributions*

The main result of this section is the convergence of the posterior distributions $Q_n$ of $Z$ (cf. 2.16 and 2.25) to a degenerate distribution, i.e. $Q_n(B)$ converges almost surely to $1_B(Z)$ for all $B \in \mathcal{T}$, provided that the

strategy, by which the system is controlled, ensures that the number of visits to the set $K_i \subset X \times A$ is infinite, with probability one. Our proof of this statement is similar to the proof of a theorem in [Doob (1949)]. In fact, our result extends Doob's result to a more abstract setting.

Another result of this section concerns the expected differences of the successive posterior distributions. In the proofs of the above mentioned results, elementary martingale theory plays an important role. As a corollary of the next lemma we shall show that given Z, the sequence $\{Y_{\tau(i,n)},\ n \in \mathbb{N}^*\}$ forms a sequence of *conditionally independent and identically distributed* random variables (cf. [Neveu (1965) page 129]), provided that $\mathbb{P}[\tau(i,n) < \infty] = 1$ for all $n \in \mathbb{N}^*$. This lemma is also used in chapter 6 (remember that $\rho \in P(X)$, $q \in W$ and $\pi \in \Pi$ are fixed).

Lemma 2.2

Let f be a nonnegative measurable function on $Y^n$. Then

$$\mathbb{E}[f(Y_{\tau(i,1)},\ldots,Y_{\tau(i,n)})] \leq \int q_i(d\theta_i)\{\int\ldots\int \nu(dy_1)\ldots\nu(dy_n) \prod_{j=1}^{n} p_i(y_j|\theta_i) f(y_1,..,y_n)\}$$

with equality if

$$\mathbb{P}[\{\tau(i,n) < \infty\}] = 1$$

(we use convention 2.23, note that $\tau(i,k) < \tau(i,n)$, $k < n$).

Proof.

It suffices to consider functions f of the form:

$$f(y_1,\ldots,y_n) = \prod_{j=1}^{n} 1_{E_j}(y_j)\ , \qquad E_j \in Y\ .$$

It is easy to verify that for $E \in Y$ we have $\mathbb{P}$-a.s. (cf. th. 1.4):

$$(*) \qquad \mathbb{E}[1_E(Y_k)|Z,X_0,A_0,\ldots,Y_{k-1},X_{k-1},A_{k-1}] =$$

$$= \int \nu(dy) \sum_{i\in I} 1_{K_i}(X_{k-1},A_{k-1}) p_i(y|z_i) 1_E(y)\ .$$

Note that $\{\tau(i,n) = k\} \in F_{k-1} \subset \sigma(Z,X_0,A_0,\ldots,Y_{k-1},X_{k-1},A_{k-1})$ .

Let h be a nonnegative measurable function on $\Theta_i$. Consider

$$\mathbb{E}[h(Z_i)\prod_{j=1}^{n} 1_{E_j}(Y_{\tau(i,j)})] = \sum_{k=n}^{\infty} \mathbb{E}[h(Z_i)\prod_{j=1}^{n-1} 1_{E_j}(Y_{\tau(i,j)})1_{\{\tau(i,n)=k\}}1_{E_n}(Y_k)] =$$

$$= \sum_{k=n}^{\infty} \mathbb{E}[h(Z_i)\prod_{j=1}^{n-1} 1_{E_j}(Y_{\tau(i,j)})1_{\{\tau(i,n)=k\}}\mathbb{E}[1_{E_n}(Y_k)|Z,X_0,A_0,\ldots,Y_{k-1},X_{k-1},A_{k-1}]] =$$

$$= \sum_{k=n}^{\infty} \mathbb{E}[h(Z_i)\prod_{j=1}^{n-1} 1_{E_j}(Y_{\tau(i,j)})1_{\{\tau(i,n)=k\}}\int \nu(dy)p_i(y|Z_i)1_{E_n}(y)] \leq$$

$$\leq \mathbb{E}[h(Z_i)\prod_{j=1}^{n-1} 1_{E_j}(Y_{\tau(i,j)})\int \nu(dy)p_i(y|Z_i)1_{E_n}(y)]$$

with equality if $\mathbb{P}[\{\tau(i,n) < \infty\}] = 1$. Note that the third equality is a consequence of (*) and the fact that $(X_{k-1},A_{k-1}) \in K_i$ on $\{\tau(i,n) = k\}$. Now we may repeat the argument for $\tau(i,n-1)$ with the function $\tilde{h}(Z_i) := h(Z_i)\int \nu(dy)p_i(y|Z_i)1_{E_n}(y)$. So we find putting h = 1

$$\mathbb{E}[\prod_{j=1}^{n} 1_{E_j}(Y_{\tau(i,j)})] \leq \mathbb{E}[\prod_{j=1}^{n}\int \nu(dy_j)p_i(y_j|Z_i)1_{E_j}(y_j)]$$

which proves the lemma. □

The following corollary is immediate.

Corollary 2.3

Let $\mathbb{P}[\tau(i,n) < \infty] = 1$. Then for $E_1,\ldots,E_n \in \mathcal{Y}$ we have q-a.s.

$$\mathbb{P}^{\pi}_{\rho,\theta}[Y_{\tau(i,1)} \in E_1,\ldots,Y_{\tau(i,n)} \in E_n] = \prod_{j=1}^{n}\int_{E_j} p_i(y|\theta_i)\nu(dy)$$

or similarly as functions on $\Omega$ we have $\mathbb{P}$-a.s.

$$\mathbb{P}^{\pi}_{\rho,Z}[Y_{\tau(i,1)} \in E_1,\ldots,Y_{\tau(i,n)} \in E_n] = \prod_{j=1}^{n}\int_{E_j} p_i(y|Z_i)\nu(dy) \ .$$

Hence

$Y_{\tau(i,1)},\ldots,Y_{\tau(i,n)}$ are, conditionally given Z, i.i.d.

Theorem 2.4

If $\mathbb{P}[\bigcap_{n\in\mathbb{N}^*} \{\tau(i,n) < \infty\}] = 1$, then for all bounded real-valued measurable functions f on $\theta_i$:

$$\lim_{n\to\infty} \int f(\theta_i) Q_n(d\theta) = f(Z_i) \qquad \mathbb{P}\text{-a.s.}$$

Remark.

This holds in particular for all bounded continuous functions f. Hence $Q_n$ converges *weakly* to the distribution which is degenerate in $Z_i$, $\mathbb{P}$-a.s.

Proof. The proof is divided into five parts.

a) We first reduce the problem.
It is easy to verify that $\int f(\theta_i) Q_n(d\theta) = \mathbb{E}[f(Z_i)|F_n]$ $\mathbb{P}$-a.s. (cf. th. 2.1) by considering indicator functions first. Since $\{F_n, n \in \mathbb{N}\}$ is an increasing sequence of $\sigma$-fields with limit $F_\infty$, and since the sequence $\{\mathbb{E}[f(Z_i)|F_n], n \in \mathbb{N}\}$ is a martingale with respect to $\{F_n, n \in \mathbb{N}\}$, we have (cf.[Neveu (1972) th. II-2-11]):

$$\lim_{n\to\infty} \int f(\theta_i) Q_n(d\theta) = \mathbb{E}[f(Z_i)|F_\infty], \quad \mathbb{P}\text{-a.s.}$$

Let $F_\infty^*$ be the completion of $F_\infty$ in $H$, i.e.
$F_\infty^* := \{A \Delta N | A \in F_\infty, N \in H, \mathbb{P}[N] = 0\}$. Obviously, it is sufficient to prove that $f(Z_i)$ is $F_\infty^*$-measurable, since $\mathbb{E}[f(Z_i)|F_\infty] = \mathbb{E}[f(Z_i)|F_\infty^*]$, $\mathbb{P}$-a.s.

So we proceed with proving that $f(Z_i)$ is $F_\infty^*$-measurable.

b) We define on $\Omega^* := \bigcap_{n\in\mathbb{N}} \{\tau(i,n) < \infty\}$ the *empirical distributions* on $(Y,\mathcal{Y})$:

$$F_n(E) := \frac{1}{n} \sum_{j=1}^{n} 1_E(Y_{\tau(i,j)}), \qquad E \in \mathcal{Y} \quad \text{(note that } \Omega^* \in F_\infty\text{)}.$$

First we verify that $F_n(\cdot)$ is a measurable function from $\Omega^*$ to $P(\mathcal{Y})$, where $P(\mathcal{Y})$ is endowed with the $\sigma$-field generated by the weak topology.

This $\sigma$-field is generated by sets of the form (cf. lemma 1.5)

$$\{\mu \in P(Y) \mid \mu(E) \leq a\} , \qquad E \in Y, \quad a \in \mathbb{R} .$$

Since

$$\{\omega \in \Omega^* \mid F_n(E) \in \{\mu \in P(Y) \mid \mu(E) \leq a\}\} = \{\omega \in \Omega^* \mid F_n(E) \leq a\} \in \mathcal{F}_\infty ,$$

the measurability of $F_n(\cdot)$ has been verified.

c) Next we consider the limit behaviour of $F_n$ for $n \to \infty$, as functions on $\Omega^*$. Let $\tilde{\mathcal{F}}_\infty$ be the restriction of $\mathcal{F}_\infty$ on $\Omega^*$, i.e. $\tilde{\mathcal{F}}_\infty := \{A \cap \Omega^* \mid A \in \mathcal{F}_\infty\}$. Since $\Omega^* \in \mathcal{F}_\infty$ we have $\tilde{\mathcal{F}}_\infty \subset \mathcal{F}_\infty$ . By corollary 2.3 we have, for q-almost all $\theta$: $Y_{\tau(i,1)}, \ldots, Y_{\tau(i,n)}$ are i.i.d. on $(\Omega^*, \tilde{\mathcal{F}}_\infty, \mathbb{P}^\pi_{\rho,\theta})$ for $n \in \mathbb{N}^*$ with common distribution $P_{\theta_i}$, where $P : \theta_i \to P(Y)$ is defined by

$$P_{\theta_i}(E) := \int_E p_i(y \mid \theta_i)\nu(dy) , \qquad \text{for } E \in Y .$$

Let $C(Y)$ be the set of all bounded continuous functions on Y. Since $(Y, Y)$ is a Borel space we may use a generalization of the Glivenko-Cantelli lemma (cf. [Parthasarathy (1967) th. 7.1 page 53]). This lemma states that $F_n$ converges weakly to $P_{\theta_i}$, $\mathbb{P}^\pi_{\rho,\theta}$-a.s.
Hence, for q-almost all $\theta$, we have

$$\mathbb{P}^\pi_{\rho,\theta}[\{\omega \in \Omega^* \mid \lim_{n\to\infty} \int g(y)F_n(dy) = \int g(y)P_{\theta_i}(dy) \text{ for all } g \in C(Y)\}] = 1 .$$

Since $\mathbb{P}^\pi_{\rho,\theta}[Z_i = \theta_i] = 1$, we have for

$$\Omega^{**} := \{\omega \in \Omega^* \mid \lim_{n\to\infty} \int g(y)F_n(dy) = \int g(y)P_{Z_i}(dy) \quad \text{for all } g \in C(Y)\}$$

that $\mathbb{P}^\pi_{\rho,\theta}[\Omega^{**}] = 1$ and using 2.10 we find $\mathbb{P}[\Omega^{**}] = 1$ .
Hence, since $\mathbb{P}[\Omega^{**}] = 1$ we have $\Omega^{**} \in \mathcal{F}^*_\infty$.

d) Further we prove that the function $P_{Z_i}$ from $\Omega$ to $P(Y)$ is $\mathcal{F}^*_\infty$-measurable. Remember that the $\sigma$-field on $P(Y)$ is also generated by sets of the form (cf. lemma 1.5):

$$\{\mu \in P(Y) \mid \int g(y)\mu(dy) \leq a\} , \qquad g \in C(Y), \; a \in \mathbb{R} .$$

Since $\Omega^{**} \in \mathcal{F}^*_\infty$ we have

$$\{\omega \in \Omega^{**} \mid \int g(y) P_{Z_i}(dy) \leq a\} = \{\omega \in \Omega^{**} \mid \lim_{n\to\infty} \int g(y) F_n(dy) \leq a\} \in \mathcal{F}_\infty^* .$$

Hence $\{\omega \in \Omega \mid \int g(y) P_{Z_i}(dy) \leq a\} \in \mathcal{F}_\infty^*$ and therefore

$P_{Z_i} : \Omega \to \mathcal{P}(Y)$ is $\mathcal{F}_\infty^*$-measurable.

e) Finally we show that $Z_i$ is $\mathcal{F}_\infty^*$-measurable.
By the separation property (cf. 2.1(j)) we have $P : \theta_i \to \mathcal{P}(Y)$ is a one-one map into $\mathcal{P}(Y)$. Since this mapping is measurable we have by Kuratowski's theorem (cf. A7) that $P^{-1}$ is also measurable. Hence, since the function $P^{-1}(P_{Z_i}) : \Omega \to \theta_i$ is $\mathcal{F}_\infty^*$-measurable and since $P^{-1}(P_{Z_i}) = Z_i$ we have that $Z_i$ is $\mathcal{F}_\infty^*$-measurable on $\Omega^{**}$. Therefore, by part (a), the theorem is proved. □

Corollary 2.5

Let $i_1, i_2, \ldots, i_k \in I$ and let $f$ be a bounded measurable function on $\theta_{i_1} \times \ldots \times \theta_{i_k}$, $k \in \overline{\mathbb{N}}^*$. Then:

$$(*) \qquad \mathbb{P}\Big[\bigcap_{j=1}^{k} \bigcap_{n\in\mathbb{N}^*} \{\tau(i_j,n) < \infty\}\Big] = 1$$

implies

$$\lim_{n\to\infty} \int f(\theta_{i_1},\ldots,\theta_{i_k}) Q_n(d\theta) = f(Z_{i_1},\ldots,Z_{i_k}) \quad \mathbb{P}\text{-a.s.}$$

Proof.

We extend $f$ to a function on $\theta$ by defining $\tilde{f}(\theta) := f(\theta_{i_1},\ldots,\theta_{i_k})$ for $\theta \in \theta$ with $\theta_{i_j}$ as $i_j$-coordinate. Let (*) hold.
As in part (a) of the proof of th. 2.4 we have:

$$\lim_{n\to\infty} \int \tilde{f}(\theta) Q_n(d\theta) = \mathbb{E}[\tilde{f}(Z) \mid \mathcal{F}_\infty] = \mathbb{E}[\tilde{f}(Z) \mid \mathcal{F}_\infty^*] \qquad \mathbb{P}\text{-a.s.}$$

It suffices to consider functions f of the form

(**) $f(\theta_{i_1},\ldots,\theta_{i_k}) = \prod_{j=1}^{m} 1_{E_j}(\theta_{i_j})$ where $E_j \in \mathcal{T}_{i_j}$, $m \le k$, $m \in \mathbb{N}^*$.

In the proof of th. 2.4 we showed that (*) implies that $Z_{i_j}$ is $\mathcal{F}_\infty^*$-measurable, $1 \le j \le k$.

Hence $\tilde{f}(Z)$ is $\mathcal{F}_\infty^*$-measurable here, which proves the statement. □

Corollary 2.6

Let $q$ be concentrated on a countable subset of $\Theta$. Then $\mathbb{P}[\bigcap_{n\in\mathbb{N}^*}\{\tau(i,n)<\infty\}]=1$ implies $\lim_{n\to\infty}\int f(\theta_i)Q_n(d\theta) = f(\theta_i)$, $\mathbb{P}^\pi_{\rho,\theta}$-a.s. for $\theta \in \Theta$ with $q(\{\theta\}) > 0$ and for any bounded measurable function $f$ on $\theta_i$.

To prove this, note that for $B = \Theta \times C$, $C \in \mathcal{H}_\infty$ : $\mathbb{P}^\pi_{\rho,q}[B] = 1$ implies $\mathbb{P}^\pi_{\rho,\theta}[B]=1$ for all $\theta \in \Theta$ with $q(\{\theta\}) > 0$ (cf. 2.11).
The countability condition in corollary 2.6 is essential. In [Freedman (1963),(1965)] and [Fabius (1964)] this problem is studied for the situation of real-valued i.i.d. random variables.

In th. 2.7 we consider a slight extension of th. 2.4, to be used in section 4.2.

Theorem 2.7

Let $i_1,\ldots,i_k \in I$, with $k \in \bar{\mathbb{N}}^*$ and let $\{\sigma_n, n \in \mathbb{N}\}$ be a sequence of stopping times, such that for $m \in \mathbb{N}$, $\{\sigma_n = m\} \in \mathcal{F}_m$ and $\sigma_0 := 0$, $\sigma_{n+1} > \sigma_n$. Let the $\sigma$-field $\mathcal{F}_{\sigma_n}$ be defined as in 2.19.

Assume: $\mathbb{P}[\bigcap_{j=1}^{k}\bigcap_{n=1}^{\infty}\{\tau(i_j,n) < \infty\}] = 1$. Then, for all bounded measurable functions $f$ on $\theta_{i_1} \times \ldots \times \theta_{i_k}$ we have, on $\bigcap_{n\in\mathbb{N}}\{\sigma_n < \infty\}$ :

$$\lim_{n\to\infty} \int Q_{\sigma_n}(d\theta) f(\theta_{i_1},\ldots,\theta_{i_k}) = f(Z_{i_1},\ldots,Z_{i_k}) \qquad \mathbb{P}\text{-a.s.}$$

(by convention $Q_\infty$ is the zero-measure).

Proof.

a) We first consider the $\sigma$-fields $\mathcal{F}_{\sigma_n}$ in more detail. Let $B \in \mathcal{F}_n$, then $B \cap \{\sigma_n = k\} = \emptyset$ if $k < n$ since $\{\sigma_n = k\} = \emptyset$. If $k \geq n$ then $B \cap \{\sigma_n = k\} \in \mathcal{F}_k$. Hence $\mathcal{F}_n \subset \mathcal{F}_{\sigma_n}$, $n \in \mathbb{N}$. Now, let $B \in \mathcal{F}_{\sigma_n}$. Then

$$B \cap \{\sigma_{n+1} = k\} = B \cap \Big(\bigcup_{n\leq \ell<k} \{\sigma_n = \ell\}\Big) \cap \{\sigma_{n+1} = k\} \in \mathcal{F}_k \text{, since}$$

$$B \cap \Big(\bigcup_{n\leq \ell<k} \{\sigma_n = \ell\}\Big) \in \mathcal{F}_{k-1}.$$

Hence the sequence $\{\mathcal{F}_{\sigma_n}, n \in \mathbb{N}\}$ is increasing. Since $\mathcal{F}_n \subset \mathcal{F}_{\sigma_n} \subset \mathcal{F}_\infty$, $n \in \mathbb{N}$ we have $\bigcup_{n\in\mathbb{N}} \mathcal{F}_n \subset \bigcup_{n\in\mathbb{N}} \mathcal{F}_{\sigma_n} \subset \mathcal{F}_\infty$. Therefore $\mathcal{F}_\infty$ is the smallest $\sigma$-field containing $\bigcup_{n\in\mathbb{N}} \mathcal{F}_{\sigma_n}$.

b) Further we consider $\int Q_{\sigma_n}(d\theta) f(\theta)$ where $f$ is a bounded measurable function on $\theta$. Let $B \in \mathcal{T}$. Then:

$$\int Q_{\sigma_n}(d\theta) 1_B(\theta) = Q_{\sigma_n}(B) = \sum_{m=n}^{\infty} 1_{\{\sigma_n=m\}} Q_m(B) =$$

$$= \sum_{m=n}^{\infty} 1_{\{\sigma_n=m\}} \mathbb{P}[Z \in B | \mathcal{F}_m] = \mathbb{P}[Z \in B | \mathcal{F}_{\sigma_n}] 1_{\{\sigma_n<\infty\}} \quad , \quad \mathbb{P}\text{-a.s.}$$

(for the last equality cf. [Neveu (1972) prop. II 1-3]).

Hence, using standard arguments, we find for each bounded measurable function $f$:

$$(*) \quad \int Q_{\sigma_n}(d\theta) f(\theta) = \mathbb{E}[f(Z) | \mathcal{F}_{\sigma_n}] 1_{\{\sigma_n<\infty\}} .$$

c) Note that, by the conclusion of part (a): $\lim_{n\to\infty} \mathbb{E}[f(Z) | \mathcal{F}_{\sigma_n}] = \mathbb{E}[f(Z) | \mathcal{F}_\infty]$ $\mathbb{P}$-a.s.

Since $\{\sigma_n < \infty\}$, $n \in \mathbb{N}$ is a nonincreasing sequence with limit $\bigcap_{n\in\mathbb{N}} \{\sigma_n < \infty\}$ we have, $\mathbb{P}$-a.s.:

$$\lim_{n\to\infty} \mathbb{E}[f(Z)\,|\,\mathcal{F}_{\sigma_n}]1_{\{\sigma_n<\infty\}} = \mathbb{E}[f(Z)\,|\,\mathcal{F}_\infty]1_{\bigcap_{n\in\mathbb{N}}\{\sigma_n<\infty\}}\ .$$

In exactly the same way as in the proof of corollary 2.5 we find $\mathbb{E}[f(Z)\,|\,\mathcal{F}_\infty] = f(Z)$, $\mathbb{P}$-a.s., which proves the theorem. □

We conclude this section with a theorem concerning the expected quadratic differences of successive posterior distributions. This result, which might be used to obtain approximations of the value function (cf. chapters 6 and 7) is of some interest in its own right.

Theorem 2.8

Let $B_1, B_2, \ldots$ be a measurable partition of $\Theta_i$, let $q = \bigotimes_{i\in I} q_i$ and assume $\mathbb{P}[\bigcap_{n\in\mathbb{N}^*}\{\tau(i,n) < \infty\}] = 1$.
Then

$$\sum_{n=m}^{\infty} \mathbb{E}\Big[\sum_{j=1}^{\infty}\{Q_{i,n+1}(B_j) - Q_{i,n}(B_j)\}^2\,|\,\mathcal{F}_m\Big] = 1 - \sum_{j=1}^{\infty} Q_{i,m}^2(B_j)$$

in particular, for m = 0

$$\sum_{n=0}^{\infty} \mathbb{E}\Big[\sum_{j=1}^{\infty}\{Q_{i,n+1}(B_j) - Q_{i,n}(B_j)\}^2\Big] = 1 - \sum_{j=1}^{\infty} q_i^2(B_j)$$

Proof.

According to th. 2.1 $Q_{i,n}(B_j) = \mathbb{P}[Z_i \in B_j\,|\,\mathcal{F}_n]$, $\mathbb{P}$-a.s. For $n \geq m \geq 0$:

$\mathbb{E}[\{Q_{i,n+1}(B_j) - Q_{i,n}(B_j)\}^2\,|\,\mathcal{F}_m] =$

$\mathbb{E}[Q_{i,n+1}^2(B_j) + Q_{i,n}^2(B_j) - 2\mathbb{E}[Q_{i,n+1}(B_j)Q_{i,n}(B_j)\,|\,\mathcal{F}_n]\,|\,\mathcal{F}_m] =$

$\mathbb{E}[Q_{i,n+1}^2(B_j) - Q_{i,n}^2(B_j)\,|\,\mathcal{F}_m]$, since $\mathbb{E}[Q_{i,n+1}(B_j)\,|\,\mathcal{F}_n] = Q_{i,n}(B_j)$

$\mathbb{P}$-a.s. (cf. th. 2.1 (vi)). Hence

$$\sum_{n=m}^{N} \mathbb{E}[\{Q_{i,n+1}(B_j) - Q_{i,n}(B_j)\}^2\,|\,\mathcal{F}_m] = \mathbb{E}[Q_{i,N+1}^2(B_j)\,|\,\mathcal{F}_m] - Q_{i,m}^2(B_j)\ .$$

By th. 2.4 we have $\lim_{n\to\infty} Q_{i,n}(B_j) = 1_{B_j}(Z_i)$, $\mathbb{P}$-a.s. Hence by the dominated

convergence theorem for conditional expectations

$$\lim_{n\to\infty} \mathbb{E}[Q^2_{i,n}(B_j)\,|\,F_m] = \mathbb{E}[1_{B_j}(Z_i)\,|\,F_m] = Q_{i,m}(B_j)\ .$$

Consequently, by changing the order of summation

$$\sum_{n=m}^{\infty} \mathbb{E}\Big[\sum_{j=1}^{\infty}\{Q_{i,n+1}(B_j) - Q_{i,n}(B_j)\}^2\,|\,F_m\Big] = \sum_{j=1}^{\infty}\{Q_{i,m}(B_j) - Q^2_{i,m}(B_j)\}$$

which proves the theorem. □

Remark.

The quantity $\sum_{j=1}^{\infty} q_i^2(B_j)$ is a measure of degeneration for the distribution $q_i$.
In fact,

$$2\{1 - \sum_{j=1}^{\infty} q_i^2(B_j)\}$$

is the *parabolic entropy* of $q_i$ with respect to the partition $B_1, B_2, \ldots$ if only a finite number of sets $B_k$ are non-empty, see [Behara and Nath (1973)]. It is easy to verify that if N is the number of non-empty sets in the partition, then $1 - \sum_{j=1}^{\infty} q_i^2(B_j) \le 1 - \frac{1}{N}$ with equality if $q_i(B_j) = \frac{1}{N}$ for the non-empty sets $B_j$, and $1 - \sum_{j=1}^{\infty} q_i(B_j) = 0$ if $q_i$ is concentrated on one.

# 3. *THE EQUIVALENT DYNAMIC PROGRAM AND OPTIMAL REWARD OPERATORS*

In section 3.1 we show that the Bayesian control model is equivalent to a dynamic program with a state space that is the Cartesian product of the original state space X and the set W of all distributions on the parameter set. This property is used frequently in the remaining chapters. In section 3.2 we study a class of optimal reward operators based on stopping times, as introduced by Wessels (cf. [Van Nunen and Wessels (1977)]). Here we consider general dynamic programs and therefore the section may be of some independent interest. However in section 3.3 we return to the Bayesian control model and we specialize the results of section 3.2 for the equivalent dynamic program. Further we give some useful properties of the value function (cf. 2.12). We note that the results of section 3.2 are used in chapters 6 and 7.

## 3.1 *Transformation into a dynamic program*

Before discussing in detail the methods and results of this section, we define the dynamic program, that turns out to be equivalent to the Bayesian control model.

model 2: *Equivalent dynamic program*

The model is defined in terms of the objects of model 1 (cf. 2.1).

3.1 (a) $X \times W$ is the *state space* endowed with the $\sigma$-field $\mathcal{X} \otimes \mathcal{W}$.

(b) $A$ is the *action space* endowed with the $\sigma$-field $\mathcal{A}$.

(c) $\tilde{D}$ is a function from $X \times W$ to the non-empty subsets of $A$ such that $\tilde{D}((x,q)) := D(x)$, $x \in X$, $q \in W$ (the sets of *admissible actions*).

(d) $\tilde{P}$ is a transition probability from $X \times W \times A$ to $X \times W$ such that

$$\tilde{P}(B \times C|x,q,a) := \sum_{i \in I} 1_{K_i}(x,a) \int_{\{y \in Y | T_{i,y}(q) \in C\}} \nu(dy) p_i(y,q) \int_B P(dx'|x,a,y)$$

$(x \in X,\ q \in W,\ a \in A,\ B \in \mathcal{X},\ C \in \mathcal{W})$ (cf. 2.28).

(e) the *reward function* $\tilde{r} : X \times W \times A \to \mathbb{R}$ is defined by

$$\tilde{r}(x,q,a) := \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y,q) r(x,a,y) \ .$$

As in section 1.2 we define the sets of histories, the strategies, the random variables and the probabilities on the sample space.

3.2 The set of *histories* at stage n, is defined by:

(i) $\tilde{H}_0 := X \times W$, $\tilde{H}_n := (X \times W \times A)^n \times X \times W$, $n \in \mathbb{N}^*$,

$\tilde{H}_\infty := \tilde{\Omega} := (X \times W \times A)^{\mathbb{N}}$.

(ii) $\tilde{\mathcal{H}}_n$ is the $\sigma$-field on $\tilde{H}_n$ induced by $X$, $W$ and $A$, $n \in \overline{\mathbb{N}}$. Let $\tilde{\mathcal{H}} := \tilde{\mathcal{H}}_\infty$.

3.3 A *strategy* $\tilde{\pi}$ is a sequence $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \tilde{\pi}_2, \ldots)$, where $\tilde{\pi}_n$ is a transition probability from $\tilde{H}_n$ to A such that $\pi_n(\cdot | x_0, q_0, a_0, \ldots, x_n, q_n)$ is concentrated on $\tilde{D}(x_n, q_n)$. The set of all strategies is denoted by $\tilde{\Pi}$.

3.4 On $\tilde{\Omega}$ we define the *coordinate functions* or *random variables* $\tilde{X}_n$, $\tilde{Q}_n$ and $\tilde{A}_n$ by $\tilde{X}_n(\omega) := x_n$, $\tilde{Q}_n(\omega) := q_n$, $\tilde{A}_n(\omega) := a_n$ where

$\omega = (x_0, q_0, a_0, x_1, q_1, a_1, \ldots) \in \tilde{\Omega}$.

3.5 For each $\rho \in P(X)$, each $q \in W$ and each $\tilde{\pi} \in \tilde{\Pi}$ there is a probability $\tilde{\mathbb{P}}^{\tilde{\pi}}_{\rho,q}$ on $\tilde{\mathcal{H}}$ determined by (cf. th. 1.4):

$$\tilde{\mathbb{P}}^{\tilde{\pi}}_{\rho,q}[\tilde{X}_0 \in B_0, \tilde{Q}_0 \in C_0, \tilde{A}_0 \in E_0, \ldots, \tilde{X}_n \in B_n, \tilde{Q}_n \in C_n, \tilde{A}_n \in E_n] :=$$

$$1_{C_0}(q_0) \int_{B_0} \rho(dx_0) \int_{E_0} \tilde{\pi}_0(da_0 | x_0, q_0) \ldots$$

$$\ldots \int_{B_n \times C_n} \tilde{P}(d(x_n, q_n) | x_{n-1}, q_{n-1}, a_{n-1}) \int_{E_n} \tilde{\pi}_n(da_n | x_0, q_0, a_0, \ldots, x_n, q_n)$$

for $B_i \in X$, $C_i \in W$ and $E_i \in A$, $i \in \mathbb{N}$ and $q_0 := q$.

We introduce a sequence of transformations $t_n : W \times H_n \to \tilde{H}_n$ which relate histories for model 1 to histories for model 2.

3.6 $t_n(q, x_0, a_0, y_1, x_1, \ldots, y_n, x_n) := (x_0, q, a_0, x_1, q_1, \ldots, x_n, q_n)$

where $q_i := q_i(q, x_0, a_0, \ldots, y_n)$ (cf. 2.30(ii)), for $i = 1, \ldots, n$.

Hence, if $q \in W$ is the prior distribution of Z, then $t_n(q, h_n)$ is the history at time n if we only observe the states $x_i$, the actions $a_i$ and the posterior distributions $q_i$ ($i \leq n$, $h_n \in H_n$).

Further we define the subset $\Pi_0$ of $\Pi$ by:

3.7 $\pi \in \Pi_0$ iff there is for each $q \in W$ a $\tilde{\pi} \in \tilde{\Pi}$ such that for all $h_n \in H_n$, $n \in \mathbb{N}$:

$$\pi_n(\cdot|h_n) = \tilde{\pi}_n(\cdot|t_n(q,h_n)) \ .$$

The strategy $\tilde{\pi}$ is called the *corresponding strategy* of $\pi$ (with respect to q).

Notice that the strategies $\pi \in \Pi_0$ base the choice of the action at time n only on $X_0,Q_0,A_0,\ldots,X_n,Q_n$.
In this section we show that we may restrict our attention to the subset of strategies $\Pi_0$ when we are looking for "good" strategies for model 1, (cf. th. 3.4).
i.e. $\sup_{\pi\in\Pi_0} v(x,q,\pi) = v(x,q)$ for all $x \in X$ and $q \in W$.
Moreover we shall show that for each $\pi \in \Pi_0$ and its corresponding $\tilde{\pi} \in \tilde{\Pi}$ the following equality is valid for all $\rho \in P(X)$ and $q \in W$:

$$\mathbb{E}^{\pi}_{\rho,q}[r(X_n,A_n,Y_{n+1})] = \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[\tilde{r}(\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n)] \quad \text{for all } n \in \mathbb{N} .$$

This implies that we may apply techniques of dynamic programming to the equivalent dynamic program in order to determine or to approximate the value function v and nearly optimal strategies.
Transformations of this type are well-known, see for example [Martin (1967)], [Wessels (1968)], [Hinderer (1970)], [Furukawa (1970)], [Yushkevich (1976)].
As far as we know the most general result is proved in [Rieder (1975)].
Translated to our situation Rieder's result implies that the Bayesian control model is equivalent to a dynamic program with state space $X \times Y \times W$, in other words that the process $\{(X_n,Y_n,Q_n,A_n),\ n \in \mathbb{N}\}$ is a Markov decision process.
To this end Rieder transforms the Bayesian equivalent model into a so-called non-Markovian decision model, as defined in [Hinderer (1970)] and afterwards he shows, using Hinderer's concept of sufficiency, that the non-Markovian decision model is equivalent to the dynamic program with state space $X \times Y \times W$.
However, we need the equivalence of the Bayesian control model to model 2, a dynamic program with state space $X \times W$. Therefore we prefer a direct proof.
Our approach employs the same idea in [Strauch (1966) th. 4.1], which is also the basis of Hinderer's sufficiency concept.

We start with some preliminaries.
Note that, according to th. 1.4, we have a "natural" regular conditional distribution $\mathbb{P}^{\pi}_{\rho,q}[\cdot|Z,X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n]$, $\rho \in P(X)$, $q \in W$, $\pi \in \Pi$.
We always choose this version without comment. For real-valued measurable functions f on $\Omega$ that are bounded form above we always define $\mathbb{E}^{\pi}_{\rho,q}[f|Z,X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n]$ as in th. 1.4(ii).

Note that we are working in model 1 until th. 3.5. Recall that in 2.24 we defined the mapping $Q_n : \Omega \to W$. On W we have the Borel $\sigma$-field $\mathcal{W}$, generated by the weak topology, which is the smallest $\sigma$-field on W such that the maps $B \to \mu(B)$ are measurable (cf. lemma 1.5) for every $B \in \mathcal{T}$. Since for every $B \in \mathcal{T}$ the mapping $\omega \to Q_n(B)(\omega)$ is measurable it follows that $\mathbb{E}^{\pi}_{\rho,q}[Q_n(B) \mid Q_n] = Q_n(B)$, $\mathbb{P}^{\pi}_{\rho,q}$-a.s.

Lemma 3.1

Fix $\rho \in \mathcal{P}(X)$, $q \in W$, $\pi \in \Pi$ and $n \in \mathbb{N}$.

(i) Let f be a real-valued measurable function on $\Theta \times X \times W \times A$, bounded from above. Then

$$\mathbb{E}^{\pi}_{\rho,q}[f(Z,X_n,Q_n,A_n) \mid X_n,Q_n,A_n] = \int f(\theta,X_n,Q_n,A_n)Q_n(d\theta) \quad , \quad \mathbb{P}^{\pi}_{\rho,q}\text{-a.s.}$$

(ii) Let f be a real-valued measurable function on $\Theta \times (X \times W \times A)^{n+1}$, bounded above. Then, for $V := (X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n)$, we have

$$\mathbb{E}^{\pi}_{\rho,q}[f(Z,V) \mid V] = \int f(\theta,V)Q_n(d\theta) \quad , \quad \mathbb{P}^{\pi}_{\rho,q}\text{-a.s.}$$

Proof.

We start with assertion (i). All equalities hold $\mathbb{P}^{\pi}_{\rho,q}$-a.s. We may suppose $f(\theta,x_n,q_n,a_n) := 1_B(\theta)1_C(x_n,q_n,a_n)$ with $B \in \mathcal{T}$ and $C \in X \otimes W \otimes A$

$$\mathbb{E}^{\pi}_{\rho,q}[f(Z,X_n,Q_n,A_n) \mid X_n,Q_n,A_n] = 1_C(X_n,Q_n,A_n)\mathbb{P}^{\pi}_{\rho,q}[Z \in B \mid X_n,Q_n,A_n].$$

Note that $\sigma(X_n,Q_n,A_n) \subset \mathcal{F}_n$. Hence

$$\mathbb{P}^{\pi}_{\rho,q}[Z \in B \mid X_n,Q_n,A_n] = \mathbb{E}^{\pi}_{\rho,q}[\mathbb{P}^{\pi}_{\rho,q}[Z \in B \mid \mathcal{F}_n] \mid X_n,Q_n,A_n] =$$

$$= \mathbb{E}^{\pi}_{\rho,q}[Q_n(B) \mid X_n,Q_n,A_n] = Q_n(B) \ .$$

Therefore we have

$$\mathbb{E}^{\pi}_{\rho,q}[f(Z,X_n,Q_n,A_n) \mid X_n,Q_n,A_n] = 1_C(X_n,Q_n,A_n)Q_n(B) =$$

$$= \int 1_B(\theta)1_C(X_n,Q_n,A_n)Q_n(d\theta) = \int f(\theta,X_n,Q_n,A_n)Q_n(d\theta) \ .$$

The $\sigma(X_n,Q_n,A_n)$-measurability of $\int f(\theta,X_n,Q_n,A_n)Q_n(d\theta)$ follows from lemma 1.6(iii). The proof of assertion (ii) is analogous. □

In lemma 3.2 we employ symbols we used before, but here we do not use their interpretation.

Lemma 3.2

Let $(\Omega,\mathcal{H})$, $(U,\mathcal{X})$ and $(V,\Gamma)$ be Borel spaces, let $X : \Omega \to U$ and $Y : \Omega \to V$ be measurable. Let $\mathbb{P}$ be a probability on $\mathcal{H}$. Further let $\mathbb{P}[\cdot|X = x]$ be a regular conditional probability given $X = x$ (cf. corollary 1.3) and let $f$ be a real-valued measurable function on $U \times V$, that is bounded above.
Define $m_x(C) := \mathbb{P}[Y \in C|X = x]$ for all $x \in U$ and $C \in \Gamma$.
Then $\int f(X,y)m_X(dy)$ is a version of $\mathbb{E}[f(X,Y)|X]$.

Proof.

First let $f(x,y) := 1_A(x)1_B(y)$ with $A \in \mathcal{X}$, $B \in \Gamma$. By corollary 1.3 and by 1.21 we have $\mathbb{E}[1_B(Y)|X] = \int 1_B(y)m_X(dy)$, $\mathbb{P}$-a.s. Hence

$$\mathbb{E}[f(X,Y)|X] = 1_A(X)\int 1_B(y)m_X(dy) = \int f(X,y)m_X(dy) \ .$$

By standard arguments the statement can be proved in general. □

In th. 3.4 we shall show that we may restrict attention to the subset of strategies $\Pi_0 \subset \Pi$, defined in 3.7. In fact we shall show more: the only interesting strategies are those, where for all $n \in \mathbb{N}$ the choice of the distribution of the action at time n depends only on the values of $X_n$ and $Q_n$. Here we use the same construction as in [Strauch (1966) th. 4.1]. The idea of this construction can also be found in [Derman and Strauch (1966)], [Wessels (1968) th. 7.4 and th. 7.5] and in [Hinderer (1970) th. 18.1]. We start with a lemma where this construction is carried out.

Lemma 3.3

Fix $\rho \in P(X)$, $q \in W$ and $\pi \in \Pi$. For all $n \in \mathbb{N}$ we fix a regular conditional probability $\mathbb{P}^{\pi}_{\rho,q}[\cdot|X_n = x_n, Q_n = q_n]$, such that
$\mathbb{P}^{\pi}_{\rho,q}[A_n \in D(x_n)|X_n = x_n, Q_n = q_n] = 1$ for $x_n \in X$ and $q_n \in W$. We define

the strategy $\pi^* \in \Pi$ by

3.8 $\pi_n^*(B|x_0,a_0,y_1,x_1,a_1,\ldots,y_n,x_n) := \mathbb{P}^{\pi}_{\rho,q}[A_n \in B|X_n = x_n,\ Q_n = q_n]$ ,

$B \in A$, where $q_n := q_n(q,x_0,a_0,y_1,\ldots,y_n)$ (cf. 2.30).

Then for each real-valued measurable function f on $X \times W \times A \times X \times W$ that is bounded above, we have simultaneously for all $n \in \mathbb{N}$:

$$\mathbb{E}^{\pi}_{\rho,q}[f(X_n,Q_n,A_n,X_{n+1},Q_{n+1})] = \mathbb{E}^{\pi^*}_{\rho,q}[f(X_n,Q_n,A_n,X_{n+1},Q_{n+1})] .$$

Proof.

The existence of a regular conditional probability $\mathbb{P}^{\pi}_{\rho,q}[\cdot|X_n = x_n,\ Q_n = q_n]$ with the desired property can be proved as in [Hinderer (1970) th. 18.1 and corollary 12.7]. We proceed by induction on n. Remember $Q_0 = q$ on $\Omega$. Hence $\sigma(X_0) = \sigma(X_0,Q_0)$. Hence for all $B \in A$ and $x_0 \in X$:

$$\pi^*(B|x_0) = \mathbb{P}^{\pi}_{\rho,q}[A_0 \in B|X_0 = x_0] = \pi(B|x_0) .$$

Hence the statement is proved for n = 0. Assume it holds for n - 1 and for all functions f satisfying the assumptions of the lemma. Define for notational convenience the function F on $\Omega$ by $F := f(X_n,Q_n,A_n,X_{n+1},Q_{n+1})$.
First we show that for all $\pi' \in \Pi$ we have $\mathbb{P}^{\pi'}_{\rho,q}$-a.s.:

(a) $$\mathbb{E}^{\pi'}_{\rho,q}[F|X_n,Q_n,A_n] = \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy)$$

$$\cdot \int P(dx|X_n,A_n,y)p_i(y,Q_n)f(X_n,Q_n,A_n,x,T_{i,y}(Q_n)) =: g(X_n,Q_n,A_n) .$$

To prove this note that (according to th. 1.4):

$$\mathbb{E}^{\pi'}_{\rho,q}[F|Z,X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n] = \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy)$$

$$\cdot \int P(dx|X_n,A_n,y)p_i(y|Z_i)f(X_n,Q_n,A_n,x,T_{i,y}(Q_n)) =: h(Z,X_n,Q_n,A_n) .$$

Hence, $\mathbb{P}^{\pi'}_{\rho,q}$-a.s.

$$\mathbb{E}^{\pi'}_{\rho,q}[F|X_n,Q_n,A_n] = \mathbb{E}^{\pi'}_{\rho,q}[h(Z,X_n,Q_n,A_n)|X_n,Q_n,A_n] =$$

$$= \int h(\theta,X_n,Q_n,A_n)Q_n(d\theta) ,$$

where the last equality follows from lemma 3.1. This proves (a). Next we prove that there are versions of the conditional expectations such that

(b) $$\mathbb{E}^{\pi}_{\rho,q}[F|X_n,Q_n] = \mathbb{E}^{\pi^*}_{\rho,q}[F|X_n,Q_n] \ .$$

Note that, by (a), $\mathbb{P}^{\pi}_{\rho,q}$-a.s.

$$\mathbb{E}^{\pi}_{\rho,q}[F|X_n,Q_n] = \mathbb{E}^{\pi}_{\rho,q}[g(X_n,Q_n,A_n)|X_n,Q_n] \ .$$

Let $m_{x_n,q_n}(B) := \mathbb{P}^{\pi}_{\rho,q}[A_n \in B|X_n = x_n,\ Q_n = q_n]$ .
Hence, by lemma 3.2, we have $\mathbb{P}^{\pi}_{\rho,q}$-a.s.

$$\mathbb{E}^{\pi}_{\rho,q}[F|X_n,Q_n] = \int g(X_n,Q_n,a)m_{X_n,Q_n}(da) \ .$$

Remember that by definition we have for all $B \in \mathcal{B}$

$$\mathbb{P}^{\pi^*}_{\rho,q}[A_n \in B|F_n] = \pi^*_n(B|X_0,A_0,Y_1,\ldots,Y_n,X_n) = m_{X_n,Q_n}(B) \ .$$

Hence $\mathbb{P}^{\pi^*}_{\rho,q}$-a.s., by lemma 3.2:

$$\mathbb{E}^{\pi^*}_{\rho,q}[F|X_n,Q_n] = \int g(X_n,Q_n,a)m_{X_n,Q_n}(da) \ .$$

This proves that $\tilde{g}(X_n,Q_n) := \int g(X_n,Q_n,a)m_{X_n,Q_n}(da)$ is a version of $\mathbb{E}^{\pi}_{\rho,q}[F|X_n,Q_n]$ and also of $\mathbb{E}^{\pi^*}_{\rho,q}[F|X_n,Q_n]$. We proceed with the final step.

$$\mathbb{E}^{\pi}_{\rho,q}[F] = \mathbb{E}^{\pi}_{\rho,q}[\mathbb{E}^{\pi}_{\rho,q}[F|X_n,Q_n]] = \mathbb{E}^{\pi}_{\rho,q}[\tilde{g}(X_n,Q_n)] =$$

$$= \mathbb{E}^{\pi^*}_{\rho,q}[\tilde{g}(X_n,Q_n)] = \mathbb{E}^{\pi^*}_{\rho,q}[\mathbb{E}^{\pi^*}_{\rho,q}[F|X_n,Q_n]] = \mathbb{E}^{\pi^*}_{\rho,q}[F]$$

where the third equality follows from the induction assumption, if we define $\tilde{f}(X_{n-1},Q_{n-1},A_{n-1},X_n,Q_n) := \tilde{g}(X_n,Q_n)$ . □

Theorem 3.4

Let $\rho \in P(X)$, $q \in W$, $\pi \in \Pi$, $n \in \mathbb{N}$ and let $\pi^*$ be as in 3.8. Then

$$\mathbb{E}^{\pi}_{\rho,q}[r(X_n,A_n,Y_{n+1})] = \mathbb{E}^{\pi}_{\rho,q}[\tilde{r}(X_n,Q_n,A_n)] = \mathbb{E}^{\pi^*}_{\rho,q}[\tilde{r}(X_n,Q_n,A_n)]$$

($\tilde{r}$ is defined in 3.1(e)).

Proof.

The second equality follows directly from lemma 3.3. We proceed with the proof of the first equality. According to th. 1.4 we have

$$\mathbb{E}^{\pi}_{\rho,q}[r(X_n,A_n,Y_{n+1})|Z,X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n] = \sum_{i\in I} 1_{K_i}(X_n,A_n)$$

$$\cdot \int \nu(dy)\, r(X_n,A_n,y)\, p_i(y|z_i) =: f(Z,X_n,A_n) \ .$$

By taking conditional expectations with respect to $\sigma(X_n,Q_n,A_n)$ we obtain $\mathbb{P}^{\pi}_{\rho,q}$-a.s.

$$(*) \qquad \mathbb{E}^{\pi}_{\rho,q}[r(X_n,A_n,Y_{n+1})|X_n,Q_n,A_n] = \mathbb{E}^{\pi}_{\rho,q}[f(Z,X_n,A_n)|X_n,Q_n,A_n] =$$

$$= \int f(\theta,X_n,A_n)\, Q_n(d\theta) = \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy)\, r(X_n,A_n,y)\, p_i(y,Q_n) =$$
$$= \tilde{r}(X_n,Q_n,A_n) \ .$$

Hence integration of the first and the last member of (*) with respect to $\mathbb{P}^{\pi}_{\rho,q}$ yields the desired result. □

In th. 3.5 we prove the announced correspondence between the strategies of $\Pi_0$ for model 1 and the strategies of $\tilde{\Pi}$ for model 2.
Note that according to th. 1.4, we have for model 2 a "natural" regular conditional probability $\mathbb{P}^{\tilde{\pi}}_{\rho,q}[\cdot|\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n]$.

Theorem 3.5

Let $\rho \in P(X)$, $q \in W$, $\pi \in \Pi_0$ and let $\tilde{\pi} \in \tilde{\Pi}$ be the corresponding strategy (cf. 3.7). Then, for all $n \in \mathbb{N}$ and all measurable functions $f : (X \times W \times A)^{n+1} \to \mathbb{R}$ that are bounded from above, we have

$$\mathbb{E}^{\pi}_{\rho,q}[f(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n)] = \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[f(\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n)] \ .$$

Proof.

Let $n = 0$. In this case the statement is valid, since $\pi_0(\cdot|x) = \tilde{\pi}_0(\cdot|x,q)$ for all $x \in X$. Assume the statement is valid for $n$ and all admissible functions $f$. It is straightforward to verify that

$$\mathbb{E}^{\pi}_{\rho,q}[f(X_0,Q_0,A_0,\ldots,X_{n+1},Q_{n+1},A_{n+1})|Z,X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n] =$$

$$= \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy) \int P(dx|X_n,A_n,y) \int \pi_{n+1}(da|X_0,A_0,Y_1,\ldots,Y_n,X_n,A_n,y,x)$$

$$\cdot p_i(y|Z_i) f(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n,x,T_{i,y}(Q_n)) =$$

$$= \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy) \int P(dx|X_n,A_n,y)$$

$$\cdot \int \tilde{\pi}_{n+1}(da|X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n,x,T_{i,y}(Q_n))$$

$$\cdot p_i(y|Z_i) f(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n,x,T_{i,y}(Q_n)) =: h(Z,X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n)$$

where the second equality is a consequence of 3.7.
By lemma 3.1(ii) we have $\mathbb{P}^{\pi}_{\rho,q}$-a.s.

$$\mathbb{E}^{\pi}_{\rho,q}[f(X_0,Q_0,A_0,\ldots,X_{n+1},Q_{n+1},A_{n+1})|X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n] =$$

$$= \sum_{i\in I} 1_{K_i}(X_n,A_n) \int \nu(dy) \int P(dx|X_n,A_n,y)$$

$$\cdot \int \tilde{\pi}_{n+1}(da|X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n,x,T_{i,y}(Q_n))$$

$$\cdot p_i(y,Q_n) f(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n,x,T_{i,y}(Q_n)) =: g(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n) .$$

For model 2 we have, according to th. 1.4:

$$\tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[f(\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_{n+1},\tilde{Q}_{n+1},\tilde{A}_{n+1})|\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n] =$$

$$= g(\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n) .$$

Hence, using the induction hypothesis we have

$$\mathbb{E}^{\pi}_{\rho,q}[f(X_0,Q_0,A_0,\ldots,X_{n+1},Q_{n+1},A_{n+1})] = \mathbb{E}^{\pi}_{\rho,q}[g(X_0,Q_0,A_0,\ldots,X_n,Q_n,A_n)] =$$

$$= \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[g(\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n)] = \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[f(\tilde{X}_0,\tilde{Q}_0,\tilde{A}_0,\ldots,\tilde{X}_{n+1},\tilde{Q}_{n+1},\tilde{A}_{n+1})] . \square$$

**Remark.**

In fact we proved by th. 3.5 the following result.
Let $q \in W$ and let $F : \Omega \to \tilde{\Omega}$ be defined by

$$F(\theta,x_0,a_0,y_1,x_1,a_1,\ldots) := (x_0,q_0,a_1,x_1,q_1,a_1,\ldots)$$

where $q_n := q_n(q,x_0,a_0,y_1,\ldots,y_n)$.
Then for all $B \in \tilde{H}$ :

$$\mathbb{P}^{\pi}_{\rho,q}[\{\omega \mid F(\omega) \in B\}] = \tilde{\mathbb{P}}^{\tilde{\pi}}_{\rho,q}[B] \ .$$

The following corollary is an immediate consequence of theorems 3.4 and 3.5.

Corollary 3.6

Let $\rho \in P(X)$ and $q \in W$. Then for all $\pi \in \Pi_0$ and its corresponding $\tilde{\pi}$ (cf. 3.7)

(i) $$\int \rho(dx)v(x,q,\pi) = \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[\sum_{n=0}^{\infty} \beta^n \tilde{r}(\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n)] \qquad \text{(cf. 2.12)}$$

and

(ii) $$\liminf_{N\to\infty} \frac{1}{N} \mathbb{E}^{\pi}_{\rho,q}[\sum_{n=0}^{N-1} r(X_n,A_n,Y_{n+1})] = \liminf_{N\to\infty} \frac{1}{N} \tilde{\mathbb{E}}^{\tilde{\pi}}_{\rho,q}[\sum_{n=0}^{N-1} \tilde{r}(\tilde{X}_n,\tilde{Q}_n,\tilde{A}_n)].$$

Moreover, the supremum over all $\pi \in \Pi_0$ on the left hand side equals the supremum over all $\tilde{\pi} \in \tilde{\Pi}$ on the right hand side, in (i) and also in (ii). (To verify this, note that for each $\tilde{\pi} \in \tilde{\Pi}$ there is a $\pi \in \Pi_0$ such that $\tilde{\pi}$ is the corresponding strategy for $\pi$).

Remark.

In case $q \in W$ is concentrated at $\theta \in \Theta$ then all posterior distributions $Q_n$ are degenerate in $\theta$ (cf. the remarks at the end of section 2.2). Hence, in this situation the Bayesian control model is equivalent to a dynamic program with state space $X^* := \{(x,\theta) \mid x \in X\}$. So we have shown here that observation of the supplementary state variables $Y_1,Y_2,Y_3,\ldots$ of the system is superfluous, in case the transition law is completely known, i.e. all information needed to control the system is contained in the state variables $X_0,X_1,X_2,\ldots$

*Since model 1 and model 2 are equivalent we shall omit the tilde in the notations for model 2 and we shall switch between these models without comment.*

We conclude this section with the introduction of some terminology.
In the class of strategies $\Pi_0$ we shall consider two nested subsets, that are of special interest in the remaining chapters.

3.9 Each measurable function $f : X \to A$ such that $f(x) \in D(x)$ for all $x \in X$, is called a *Markov policy* and the strategy $\pi \in \Pi_0$, defined by $\pi_n(\{a\}|x_0,a_0,y_1,x_1,\ldots,y_n,x_n) := 1$ if $f(x_n) = a$, $n \in \mathbb{N}$ is called a *stationary strategy*.

3.10 Each measurable function $f : X \times W \to A$ such that $f(x,q) \in D(x)$ for all $x \in X$ and $q \in W$, is called a *Bayesian Markov policy* and the strategy $\pi \in \Pi_0$ defined by $\pi_n(\{a\}|x_0,a_0,y_1,x_1,\ldots,y_n,x_n) := 1$ if $f(x_n,q_n) = a$, $n \in \mathbb{N}$ where $q_i = q_i(q_0,x_0,a_0,y_1,\ldots,y_i)$, is called a *Bayesian stationary strategy*.

Remarks.

(i) Each stationary strategy is also a Bayesian stationary strategy.

(ii) It is easy to verify that, under each Bayesian stationary strategy, the process $\{(X_n,Q_n), n \in \mathbb{N}\}$ forms a Markov chain. This is well-known if we are considering model 2, however for model 1 it is a consequence of the equivalence between the two models.

(iii) As a consequence of the equivalence between the models 1 and 2 we may apply the numerous results for dynamic programs to model 1. We only mention one of these results: if r is bounded then the supremum over all Bayesian stationary strategies of the Bayesian discounted total return equals the optimal value (cf. [Blackwell (1965)]). In other words it suffices to consider only the Bayesian stationary strategies.

(iv) If the action space A is a finite set then any Bayesian Markov policy f such that, $f(x,q)$ is a maximizer in the set $D(x)$ of

$$a \to \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y,q) \{r(x,a,y) + \beta \int P(dx'|x,a,y) v(x', T_{i,y}(q))\}$$

for $(x,q) \in X \times W$ is optimal (cf. [Blackwell (1965) th. 7]).

## 3.2 *A class of optimal reward operators*

In this section we study optimal reward operators for dynamic programs with complete separable metric state and action spaces. These operators are based on stopping times. They generalize the well-known optimal reward operator introduced in [Blackwell (1965)]. In [Wessels (1974)] these operators have been studied for dynamic programs with finite state and action spaces and they have been generalized for models with a countable state space and an

arbitrary action space in [Van Nunen and Wessels (1977)]. Van Nunen and Wessels show that a number of well-known approximation methods for the value function in discounted dynamic programming, such as the Gauss - Seidel iteration are equivalent to successive applications of an optimal reward operator corresponding to a suitable stopping time. We prove some new results on these optimal reward operators. First we show that if such an operator is applied to a function on the state space that is upper semi-analytic and bounded from above, then the result of the operation is again an upper semi-analytic function which is bounded from above. This generalizes a rather theoretical result in [Blackwell, Freedman and Orkin (1974)] and [Shreve (1977)] for the optimal reward operator introduced by Blackwell, to a similar result for all optimal reward operators of the class we consider.
Further we show that successive applications of two of these operators, possibly for different stopping times, have the same result as one application of the optimal reward operator which belongs to the *composed stopping time* (cf. 3.14 for a definition). This property has some interesting consequences, one of which is that we can generalize results proved by Van Nunen and Wessels for discounted dynamic programs, using the fixed point theorem for contraction mappings, to more general models.
In chapter 6 we use another consequence of this property for the equivalent dynamic program (model 2). There we study the optimal reward operator, corresponding to the entrance time in a subset of the state space.
Using this operator is equivalent to transforming the model into a dynamic program with this subset as a state space.
Since we are dealing with a general dynamic program here, we have to introduce some new notations. (Symbols used in this section do not have the interpretation, given in the foregoing part of the monograph).

model 3: *General dynamic program.*

3.11 (a) $(S,\mathcal{S})$ is a Borel space, called the state space.
(b) $(A,\mathcal{A})$ is a Borel space, called the action space.
(c) D is a function from S to the non-empty subsets of A such that $K := \{(s,a) \mid s \in S,\ a \in D(s)\}$ is an element of $\mathcal{S} \otimes \mathcal{A}$, and it is assumed that K contains the graph of some measurable function. from S to A.
(d) P is a transition probability from $S \times A$ to S.
(e) r is a real-valued measurable function on $S \times A$, that is bounded from above. $\beta \in [0,1)$ is the discount factor.

The sets of histories, the strategies, the random variables and the probabilities on the sample space are defined analogous to model 2 (cf. 3.2-3.5). We even use the same notations, with the exception that we omit the tilde and that the coordinate functions on the state space are denoted by $S_n$ for $n \in \mathbb{N}$.

We start with some definitions:

3.12 A *stopping time* $\tau$ is a measurable function from $\Omega$ to $\overline{\mathbb{N}}$ such that $\{\tau = n\} \in H_n$.

3.13 The *shift operator* $\psi$ is a function from $\Omega$ to $\Omega$ such that $\psi(s_0,a_0,s_1,a_1,\ldots) := (s_1,a_1,s_2,a_2,\ldots)$ for $(s_0,a_0,s_1,a_1,\ldots) \in \Omega$. The iterates of $\psi$ are defined by: $\psi^0(\omega) := \omega$ and $\psi^n(\omega) := \psi(\psi^{n-1}(\omega))$ for $\omega \in \Omega$ and $n \in \mathbb{N}^*$.

3.14 The set of all stopping times is denoted by $\Sigma$, and on $\Sigma$ we define the operation: $\circ$ by

$$(\tau_1 \circ \tau_2)(\omega) := \tau_1(\omega) + \tau_2(\psi^{\tau_1(\omega)}(\omega)) \quad \text{if} \quad \tau_1(\omega) < \infty$$

$$:= \infty \text{ if } \tau_1(\omega) = \infty, \quad \text{for } \omega \in \Omega \text{ and } \tau_1,\tau_2 \in \Sigma .$$

The function $\tau_1 \circ \tau_2$ on $\Omega$ is called the *composed stopping time*.

It is easy to verify that $\tau_1 \circ \tau_2 \in \Sigma$ (cf. [Revuz (1976) page 22]).

3.15 (i) $\mathcal{B}_m(S)$ is the set of real-valued measurable functions on S, which are bounded from above.

(ii) $\mathcal{B}_a(S)$ is the set of *upper semi-analytic* (u.s.a.) functions on S, which are bounded from above.

In appendix A we give a definition of u.s.a. functions and there we also collect some useful properties of these functions. Note that $\mathcal{B}_m(S) \subset \mathcal{B}_a(S)$. Finally we define for each $\tau \in \Sigma$ the corresponding optimal reward operator.

3.16 The *optimal reward operator* $U_\tau$ is defined for functions $b \in \mathcal{B}_a(S)$ by:

$$(U_\tau b)(s) := \sup_{\pi \in \Pi} \mathbb{E}_s^\pi \Big[ \sum_{n=0}^{\tau-1} \beta^n r(S_n,A_n) + \beta^\tau b(S_\tau) \Big]$$

(we use the convention $b(S_\tau) = 0$ on $\{\tau = \infty\}$ (cf. 2.23)).

The usual optimal reward operator U, introduced by Blackwell, can be defined by

3.17 $U := U_1$ where 1 is the stopping time identically one on $\Omega$.

Note that, $(Ub)(s) = \sup_{a \in D(s)} \{r(s,a) + \beta \int P(ds'|s,a) b(s')\}$, for $b \in \mathcal{B}_a(S)$.

It is well-known that Ub need not to be an element of $\mathcal{B}_m(S)$ if $b \in \mathcal{B}_m(S)$ (cf. [Blackwell (1965)]). However in [Strauch (1966)] it is shown that the value function is u.s.a. and in [Blackwell, Freedman and Orkin (1974)] it has been proved that $Ub \in \mathcal{B}_a(S)$ if $b \in \mathcal{B}_a(S)$. In [Shreve (1977)] the same result is obtained. We show this property for all operators $U_\tau$, $\tau \in \Sigma$.

**Theorem 3.7**

Let $\tau \in \Sigma$ and $b \in \mathcal{B}_a(S)$. Then $(U_\tau b) \in \mathcal{B}_a(S)$.

**Proof.**

Define on $\Omega$ $Y := \sum_{n=0}^{\tau-1} \beta^n r(S_n, A_n) + \beta^\tau b(S_\tau)$. It is straightforward to verify that $Y \leq M(1-\beta)^{-1}$, where $M := \sup_{(s,a) \in S \times A} r(s,a) + \sup_{s \in S} b(s)$.

(We divide the proof into three parts.)

(a) We first show that Y is u.s.a. on $\Omega$. Define for $c \in \mathbb{R}$:
$E_c := \{\omega \in \Omega | \beta^\tau b(S_\tau) > c\}$. Let $c > 0$. Then:
$E_c = \bigcup_{n \in \mathbb{N}} (\{\tau = n\} \cap \{b(S_n) > c\beta^{-n}\})$. Note that $b(S_n)$ is u.s.a. (cf. A9). Consequently $E_c$ is analytic (cf. A2). Let $c \leq 0$. Then:
$E_c = \bigcup_{n \in \mathbb{N}} (\{\tau = n\} \cap \{b(S_n) > c\beta^{-n}\}) \cup \{\tau = \infty\}$. Hence in this case $E_c$ is also analytic. Therefore $\beta^\tau b(S_\tau)$ is u.s.a. on $\Omega$ and so Y is u.s.a. on $\Omega$ (cf. A 8).

(b) Consider the function on $P((A \times S)^{\mathbb{N}}) \times S$ defined by:
$(\mathbb{P}, s) \to \int Y(s,\omega') \mathbb{P}(d\omega')$, $(\omega' \in (A \times S)^{\mathbb{N}})$. We show that this function is u.s.a. (Note that $P((A \times S)^{\mathbb{N}})$ is endowed with the topology of weak convergence.) To this end we define the function $\tilde{Y}$ on $P((A \times S)^{\mathbb{N}}) \times \Omega$ by: $\tilde{Y}(\mathbb{P}, \omega) := Y(\omega)$. To show that $\tilde{Y}$ is u.s.a. note that, for $c \in \mathbb{R}$:

$$\{(\mathbb{P}, \omega) | \tilde{Y}(\mathbb{P}, \omega) > c\} = P((A \times S)^{\mathbb{N}}) \times \{\omega \in \Omega | Y(\omega) > c\} .$$

Since $P((A \times S)^{\mathbb{N}})$ is a measurable set and by part (a) $\{\omega \in \Omega | Y(\omega) > c\}$ is analytic we have by A2 that $\tilde{Y}$ is u.s.a. Further we define a transition probability p from $P((A \times S)^{\mathbb{N}}) \times S$ to $(A \times S)^{\mathbb{N}}$ by

$$p(d\omega' | \mathbb{P}, s) := \mathbb{P}(d\omega') .$$

To verify that p is indeed a transition probability, note that

$$\{(\mathbb{P},s)\,|\,p(B|\mathbb{P},s) \le c\} = \{\mathbb{P} \in \mathcal{P}((A \times S)^{\mathbb{N}})\,|\,\mathbb{P}(B) \le c\} \times S$$

for $B \in (\mathcal{A} \otimes \mathcal{S})^{\mathbb{N}}$ and $c \in \mathbb{R}$. Hence, by lemma 1.5(i) this set is measurable. Finally we note that, by A10 the function on $\mathcal{P}((A \times S)^{\mathbb{N}}) \times S$:

$$(\mathbb{P},s) \to \int \tilde{Y}(\mathbb{P},s,\omega')p(d\omega'|\mathbb{P},s) = \int Y(s,\omega')\mathbb{P}(d\omega')$$

is u.s.a.

(c) Introduce the set $\Delta := \{(\mathbb{P},s)\,|\,s \in S,\ \mathbb{P} \in \mathcal{P}((A \times S)^{\mathbb{N}})$ such that for a $\pi \in \Pi : \mathbb{P}[B] = \mathbb{P}_s^{\pi}[S \times B]$ for all $B \in (\mathcal{A} \otimes \mathcal{S})^{\mathbb{N}}\}$.
It has been proved in [Hinderer (1970) lemma 13.1] that $\Delta$ is a Borel subset of $\mathcal{P}((A \times S)^{\mathbb{N}}) \times S$. (Note that model 3 is an example of Hinderer's model.) Hence it is straightforward to verify that

$$(\mathbb{P},s) \to F(\mathbb{P},s) := \int Y(s,\omega')\mathbb{P}(d\omega')1_{\Delta}(\mathbb{P},s) - 1_{\Delta^c}(\mathbb{P},s)\cdot\infty$$

is u.s.a. on $\mathcal{P}((A \times S)^{\mathbb{N}}) \times S$.
Finally we remark that $(U_\tau b)(s) = \sup_{\mathbb{P}} F(\mathbb{P},s)$, where the supremum is taken over all $\mathbb{P} \in \mathcal{P}((A \times S)^{\mathbb{N}})$.
Hence, since $\{s \in S\,|\,(U_\tau b)(s) > c\} = \mathrm{proj}_S\{(\mathbb{P},s)\,|\,F(\mathbb{P},s) > c\}$ for $c \in \mathbb{R}$, we have by A4 : $U_\tau b$ is u.s.a. □

It has been shown in [Blackwell (1965) example 1] that even if $b \in \mathcal{B}_m(S)$ it is not necessarily true that for every $\varepsilon > 0$ there is a strategy $\pi \in \Pi$ such that

$$\mathbb{E}_s^{\pi}[r(S_0,A_0) + \beta b(S_1)] \ge (Ub)(s) - \varepsilon \quad \text{for all } s \in S\ .$$

However, in [Blackwell, Freedman and Orkin (1974)] it has been shown that there always exists a *"universally measurable strategy"* $\pi$ with this property (i.e. $\pi_n(B|\cdot)$ is universally measurable for all $B \in \mathcal{A}$, see appendix A). Moreover in [Shreve (1977)] the same property is proved for the stopping time that is identically infinite. It should be possbile to establish a similar result for arbitrary stopping times in $\Sigma$. However, we do not need such a result, as we have the following lemma.

Lemma 3.8

If $b \in \mathcal{B}_m(S)$, $\tau \in \Sigma$, $\rho \in \mathcal{P}(S)$ and $\varepsilon > 0$, then there is a strategy $\pi \in \Pi$ such that

3.18 $$\rho(\{\mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^{\tau} b(S_{\tau})] \geq (U_{\tau}b)(s) - \varepsilon\}) = 1 \ .$$

Proof.

This statement is a simple consequence of th. 14.1 in [Hinderer (1970)], which is a generalization of th. 8.1 in [Strauch (1966)] to non-stationary models. To verify this, we note that the function $1_{\{\tau=n\}}$ on $\Omega$ is $H_n$-measurable. Hence by lemma 1.1 there are real-valued measurable functions $f_n$ and $\ell_n$ on $(S \times A)^n \times S$ such that:

3.19 (i) $f_n(S_0,A_0,\ldots,S_n) = 1_{\{\tau=n\}}$ on $\Omega$.

(ii) $\ell_n(s_0,a_0,\ldots,s_n) = 1$ if $\Sigma_{m=0}^{n} f_m(s_0,a_0,\ldots,s_m) = 0$

$= 0$ otherwise,

$s_i \in S$, $a_i \in A$ and $i = 0,\ldots,n$ .

Hence $\ell_n(S_0,A_0,\ldots,S_n) = 1$ if and only if $\tau > n$. Further we define, for $n \in \mathbb{N}$:

$$r_n(s_0,a_0,\ldots,s_n,a_n) := \beta^n r(s_n,a_n)\ell_n(s_0,a_0,\ldots,s_n) + \beta^n b(s_n) f_n(s_0,a_0,\ldots,s_n)$$

for $s_i \in S$, $a_i \in A$ and $i = 0,\ldots,n$. It is straightforward to verify that

$$\sum_{n=0}^{\infty} r_n(S_0,A_0,\ldots,S_n,A_n) = \sum_{n=0}^{\tau-1} \beta^n r(S_n,A_n) + \beta^{\tau} b(S_{\tau}) \ .$$

Hence we are dealing with a total-return model in the sense of Hinderer and the assertion follows from the above mentioned result of Hinderer. □

The main result of this section is th. 3.11 which states that for each pair of stopping times $\tau_1,\tau_2 \in \Sigma$ and each function $b \in \mathcal{B}_m(S)$ the following identity is valid:

$$U_{\tau_1 \circ \tau_2} b = U_{\tau_1}(U_{\tau_2} b) \ .$$

To prove this we need some preparations.

3.20 For any pair of strategies $\pi^{(1)},\pi^{(2)} \in \Pi$ and any $\tau \in \Sigma$ we define a new strategy $\pi^{\tau} \in \Pi$ by:

$$\pi_n^\tau(\cdot|s_0,a_0,\ldots,s_n) := \sum_{k=0}^{n} \pi_{n-k}^{(2)}(\cdot|s_k,a_k,\ldots,s_n) f_k(s_0,a_0,\ldots,s_k) +$$

$$+ \pi_n^{(1)}(\cdot|s_0,a_0,\ldots,s_n) \ell_n(s_0,a_0,\ldots,s_n)$$

for $s_i \in S$, $a_i \in A$ and $i \in \mathbb{N}$ ($f_k$ and $\ell_n$ are defined in 3.19).

Note that $\pi^\tau$ uses $\pi^{(1)}$ until time $\tau$ and $\pi^{(2)}$ afterwards. It is easy to verify that indeed $\pi^\tau \in \Pi$, since $\pi_n^\tau(B|\cdot)$ is $(S \otimes A)^{n-1} \otimes S$-measurable for all $B \in A$.

3.21 Let $\tau \in \Sigma$. The $\sigma$-field $H_\tau$ is defined as usual by:

$$H_\tau := \{B \in H \mid B \cap \{\tau = n\} \in H_n\} .$$

Lemma 3.9

Let f be a real-valued measurable function on $\Omega$, which is bounded from above. Let $\pi^{(1)},\pi^{(2)} \in \Pi$, $\tau \in \Sigma$ and let $\pi^\tau$ be defined by 3.20. Then we have on $\{\tau < \infty\}$ :

$$\mathbb{E}_s^{\pi^\tau}[f(\psi^\tau)|H_\tau] = \mathbb{E}_{S_\tau}^{\pi^{(2)}}[f] , \qquad \mathbb{P}_s^{\pi^{(1)}}\text{-a.s.}$$

(By convention the both sides vanish on $\{\tau = \infty\}$.)

Proof.

Let $n \in \mathbb{N}$, $B_1 \in H_n$ and $B_2 \in H$. For the stopping time n, which is identically equal to n, it is straightforward to verify, using th. 1.4:

$$(*) \qquad \mathbb{P}_s^{\pi^n}[B_1 \cap \{\psi^n \in B_2\}] = \int_{B_1} \mathbb{P}_{S_n}^{\pi^{(2)}}[B_2] d\mathbb{P}_s^{\pi^n} .$$

Let $f := 1_{B_2}$ and let $B_1 \in H_\tau$ .
Then, since $\mathbb{P}_s^{\pi^\tau}[B] = \mathbb{P}_s^{\pi^n}[B]$ if $B \subset \{\tau = n\}$ , $B \in H$ we have

$$(**) \qquad \int_{B_1} \mathbb{E}_{S_\tau}^{\pi^{(2)}}[f] d\mathbb{P}_s^{\pi^\tau} = \sum_{n=0}^{\infty} \int_{B_1 \cap \{\tau=n\}} \mathbb{P}_{S_n}^{\pi^{(2)}}[B_2] d\mathbb{P}_s^{\pi^n} , \quad \text{for } s \in S .$$

On the other hand, by the definition of conditional expectations (cf. 1.20(i))

$$(***) \qquad \int_{B_1} \mathbb{E}_s^{\pi^\tau}[f(\psi^\tau)|H_\tau] d\mathbb{P}_s^{\pi^\tau} = \sum_{n=0}^{\infty} \int_{B_1 \cap \{\tau=n\}} 1_{B_2}(\psi^n) d\mathbb{P}_s^{\pi^n} .$$

Hence, using (*) we conclude that the left-hand sides of (**) and (***) are identical. By standard arguments the assertion is proved in general. □

Remark.

In fact we proved here the strong Markov property for a special stopping time and a special non-Markovian process.

3.22 For each $\pi \in \Pi$ and each $(s_0,a_0,\ldots,s_k,a_k) \in (S \times A)^{k+1}$ we define a new strategy $\tilde{\pi}(s_0,a_0,\ldots,s_k,a_k) = (\tilde{\pi}_0,\tilde{\pi}_1,\ldots)$ by:

$$\tilde{\pi}_n(\cdot|h_n) := \pi_{n+k+1}(\cdot|s_0,a_0,\ldots,s_k,a_k,h_n), \quad h_n \in H_n, \quad n \in \mathbb{N}.$$

Note that $\tilde{\pi}(s_0,a_0,\ldots,s_k,a_k) \in \Pi$. This strategy $\tilde{\pi}(s_0,a_0,\ldots,s_k,a_k)$ acts like the strategy $\pi$ if the process has a "prehistory" $s_0,a_0,\ldots s_k,a_k$.

Lemma 3.10

Let $\pi \in \Pi$ and let $\tilde{\pi}(s_0,a_0,\ldots,s_k,a_k)$ be defined in 3.22 for each $(s_0,a_0,\ldots,s_k,a_k) \in (S \times A)^{k+1}$. Further let $f$ be a real-valued measurable function on $\Omega$, which is bounded from above, and let $\tau \in \Sigma$. Then

$$\mathbb{E}_s^{\pi}[f(\psi^{\tau})|H_{\tau}] = \mathbb{E}_{S_{\tau}}^{\tilde{\pi}(S_0,A_0,\ldots,A_{\tau-1})}[f], \quad \mathbb{P}_s^{\pi}\text{-a.s. on } \{\tau < \infty\}.$$

(By convention both expressions vanish on $\{\tau = \infty\}$.)

Proof.

It is easy to verify (cf. th. 1.4) that $\mathbb{E}_{S_k}^{\tilde{\pi}(S_0,A_0,\ldots,A_{k-1})}[f]$ is $H_k$-measurable, $k \in \mathbb{N}$. Let $B_1 \in H_n$ and $B_2 \in H$. Again using th. 1.4 one easily verifies that

$$(*) \qquad \mathbb{P}_s^{\pi}[B_1 \cap \{\psi^n \in B_2\}] = \int_{B_1} \mathbb{P}_{S_n}^{\tilde{\pi}(S_0,A_0,\ldots,A_{n-1})}[B_2]d\mathbb{P}_s^{\pi}.$$

Now we let $B_1 \in H_{\tau}$ and $f := 1_{B_2}$. Then

$$(**) \qquad \int_{B_1} \mathbb{E}_{S_{\tau}}^{\tilde{\pi}(S_0,A_0,\ldots,A_{\tau-1})}[f]d\mathbb{P}_s^{\pi} = \sum_{n=0}^{\infty} \int_{B_1 \cap \{\tau=n\}} \mathbb{P}_{S_n}^{\tilde{\pi}(S_0,A_0,\ldots,A_{n-1})}[B_2]d\mathbb{P}_s^{\pi},$$

and on the other hand (cf. 1.20(i)):

$$(***)\qquad \int_{B_1} \mathbb{E}_s^\pi[f(\psi^\tau)\,|\,H_\tau]\,d\mathbb{P}_s^\pi = \sum_{n=0}^{\infty} \int_{B_1\cap\{\tau=n\}} 1_{B_2}(\psi^n)\,d\mathbb{P}_s^\pi =$$

$$\sum_{n=0}^{\infty} \mathbb{P}_s^\pi[B_1 \cap \{\tau = n\} \cap \{\psi^n \in B_2\}] .$$

From (*) we conclude that (**) and (***) are identical. Hence the assertion has been verified for indicator functions, and it can be proved in general by standard arguments. □

Now we are ready to prove th. 3.11. Note that $U_\sigma b \in \mathcal{B}_a(S)$ if $\sigma \in \Sigma$ and $b \in \mathcal{B}_m(S)$.

Theorem 3.11

Let $\tau,\sigma \in \Sigma$ and let $b \in \mathcal{B}_m(S)$. Then we have

$$U_{\tau\circ\sigma} = U_\tau(U_\sigma b) .$$

Proof.

(a) Fix $\varepsilon > 0$ and let $s_0 \in S$, $\pi \in \Pi$ and $b \in \mathcal{B}_m(S)$. First we assume: $\mathbb{P}_{s_0}^\pi[\tau = \infty] < 1$. Then it is easily verified that the set-function $\rho$ on $S$, defined by

$$\rho(B) := \mathbb{E}_{s_0}^\pi[\beta^\tau 1_B(S_\tau)]\{\mathbb{E}_{s_0}^\pi[\beta^\tau]\}^{-1}$$

is a probability on $\rho$. By lemma 3.8 there exists a strategy $\pi^* \in \Pi$ such that

$$(*)\qquad \mathbb{E}_s^{\pi^*}[\sum_{n=0}^{\sigma-1} \beta^n r(S_n,A_n) + \beta^\sigma b(S_\sigma)] \geq (U_\sigma b)(s) - \varepsilon,\ \rho\text{-a.s. on } S.$$

Define: $\pi^{(1)} := \pi$, $\pi^{(2)} := \pi^*$. Let $\pi^\tau$ be defined by 3.20. Then we have

$$Y(s,\pi) := \mathbb{E}_s^{\pi^\tau}[\sum_{n=0}^{\tau\circ\sigma-1} \beta^n r(S_n,A_n) + \beta^{\tau\circ\sigma} b(S_{\tau\circ\sigma})] =$$

$$= \mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \mathbb{E}_s^{\pi}[\sum_{n=\tau}^{\tau\circ\sigma-1}\beta^n r(S_n,A_n) + \beta^{\tau\circ\sigma}b(S_{\tau\circ\sigma})|H_\tau]]$$

Note that $S_{\tau\circ\sigma} = S_{\tau+\sigma(\psi^\tau)} = S_{\sigma(\psi^\tau)}(\psi^\tau)$ on $\{\tau < \infty, \sigma < \infty\}$.

Using lemma 3.9 we find

$$(**) \qquad Y(s,\pi) = \mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^\tau \mathbb{E}_{S_\tau}^{\pi^*}[\sum_{n=0}^{\sigma-1}\beta^n r(S_n,A_n) + \beta^\sigma b(S_\sigma)]] \quad .$$

Note that $\mathbb{P}_s^{\pi^\tau}[B] = \mathbb{P}_s^{\pi}[B]$ for $B \in H_\tau$, $s \in S$. (To verify this note that $\mathbb{P}_s^{\pi^\tau}[B] = \sum_{n=0}^{\infty}\mathbb{P}_s^{\pi^\tau}[B \cap \{\tau = n\}] = \sum_{n=0}^{\infty}\mathbb{P}_s^{\pi}[B \cap \{\tau = n\}] = \mathbb{P}_s^{\pi}[B]$ since $B \cap \{\tau = n\} \in H_n$). Therefore we also have $\mathbb{E}_s^{\pi^\tau}[f] = \mathbb{E}_s^{\pi}[f]$ for a real-valued $H_\tau$-measurable function f, which is bounded from above. Using (*), (**) and the definition of $\rho$ we find:

$$Y(s_0,\pi) \geq \mathbb{E}_{s_0}^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^\tau (U_\sigma b)(S_\tau)] - \varepsilon\mathbb{E}_{s_0}^{\pi}[\beta^\tau] \; .$$

Now we assume $\mathbb{P}_{s_0}^{\pi}[\tau = \infty] = 1$. Then

$$Y(s_0,\pi) = \mathbb{E}_{s_0}^{\pi}[\sum_{n=0}^{\infty}\beta^n r(S_n,A_n)] = \mathbb{E}_{s_0}^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^\tau (U_\sigma b)(S_\tau)].$$

Hence

$$\sup_{\pi\in\Pi} Y(s_0,\pi) \geq (U_\tau(U_\sigma b))(s_0) - \varepsilon$$

and since $s_0 \in S$ and $\varepsilon$ are arbitrary we conclude:

$$(***) \qquad (U_{\tau\circ\sigma}b)(s) \geq \sup_{\pi\in\Pi} Y(s,\pi) \geq (U_\tau(U_\sigma b))(s) \; , \qquad s \in S \; .$$

(b) We show that (***) is valid with $\leq$ instead of $\geq$.

Let $\pi \in \Pi$ and let $\tilde{\pi}(s_0,a_0,\ldots,s_k,a_k)$ be defined by 3.22 for $(s_0,a_0,\ldots,s_k,a_k) \in (S \times A)^{k+1}$. Note that $S_n = S_{n-\tau}(\psi^\tau)$ on $\{\tau \leq n\}$. Consider for $s \in S$:

$$\mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau\circ\sigma-1}\beta^n r(S_n,A_n) + \beta^{\tau\circ\sigma}b(S_{\tau\circ\sigma})] =$$

$$= \mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \mathbb{E}_s^{\pi}[\sum_{n=\tau}^{\tau\circ\sigma-1}\beta^n r(S_n,A_n) + \beta^{\tau\circ\sigma}b(S_{\tau\circ\sigma})|H_\tau]] =$$

$$= \mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^{\tau}\mathbb{E}_{S_\tau}^{\tilde{\pi}(S_0,A_0,\ldots,A_{\tau-1})}[\sum_{n=0}^{\sigma-1} r(S_n,A_n) + \beta^{\sigma}b(S_\sigma)]] \leq$$

$$\leq \mathbb{E}_s^{\pi}[\sum_{n=0}^{\tau-1}\beta^n r(S_n,A_n) + \beta^{\tau}(U_\sigma b)(S_\tau)] \leq (U_\tau(U_\sigma b))(s) \ .$$

The second equality is justified by lemma 3.10 and the inequality by the definition of $U_\sigma$. This proves the theorem. □

To be able to apply th. 3.11 we need the following lemma.

Lemma 3.12

The operation $\circ$ on $\Sigma$ is associative.

Proof.

Let $\tau,\sigma,\rho \in \Sigma$. By 3.14 we have: $\tau \circ \sigma = \tau + \sigma(\psi^\tau)$ and so

$$(\tau \circ \sigma) \circ \rho = \tau + \sigma(\psi^\tau) + \rho(\psi^{\tau+\sigma(\psi^\tau)}) \ .$$

On the other hand

$$\tau \circ (\sigma \circ \rho) = \tau + (\sigma \circ \rho)(\psi^\tau) = \tau + \sigma(\psi^\tau) + \rho(\psi^{\sigma(\psi^\tau)}(\psi^\tau)) \ .$$

Since $\psi^k(\psi^\ell) = \psi^{k+\ell}$ we have $\psi^{\tau+\sigma(\psi^\tau)} = \psi^{\sigma(\psi^\tau)}(\psi^\tau)$ which proves the lemma.□

Hence $\tau_1 \circ \tau_2 \circ \ldots \circ \tau_n$ is defined now. As a consequence of th. 3.11 and lemma 3.12 we find:

Corollary 3.13

Let $b \in \mathcal{B}_m(S)$ and let $\tau_1,\tau_2,\ldots,\tau_n \in \Sigma$. Then

$$U_{\tau_1\circ\tau_2\circ\ldots\circ\tau_n} b = U_{\tau_1}(U_{\tau_2}(\ldots U_{\tau_n} b)\ldots) \ .$$

To prove this, note that $\tau_1 \circ \tau_2 \circ \ldots \circ \tau_n = \tau_1 \circ (\tau_2 \circ \ldots \circ \tau_n)$. Hence $U_{\tau_1\circ\tau_2\circ\ldots\circ\tau_n} b = U_{\tau_1}(U_{\tau_2\ldots\tau_n} b)$, and the assertion follows by iteration.

We continue with two definitions:

3.23 Let $\tau \in \Sigma$. The stopping time $\tau^n$, $n \in \mathbb{N}^*$ is defined by:
$\tau^1 := \tau$, $\tau^n := \tau^{n-1} \circ \tau$, $n = 2,3,\ldots$ .

3.24 Let $\tau \in \Sigma$. Then $U_\tau^n$ is defined on $\mathcal{B}_a(S)$ by:
$U_\tau^1 b := U_\tau b$ for all $b \in \mathcal{B}_a(S)$ and $U_\tau^n b := U_\tau(U_\tau^{n-1} b)$ for all $b \in \mathcal{B}_a(S)$ and $n = 2,3,\ldots$ .

In th. 3.14 we collect some consequences of th. 3.11.

Theorem 3.14

Let $b \in \mathcal{B}_m(S)$, $\tau \in \Sigma$. Then:

(i) $U_{\tau^n} b = U_\tau^n b$ (cf. 3.23).

(ii) The value function v for model 3 is u.s.a. and satisfies the optimality equation:

$$U_\tau v = v, \text{ on } S .$$

(iii) If the reward function r and b are bounded and if $\tau(\omega) \geq 1$ for all $\omega \in \Omega$ then $\lim_{n\to\infty} U_\tau^n b = v$, on S.

Proof.

Note that (i) is an immediate consequence of 3.23, 3.24 and corollary 3.13. We proceed with (ii). Note that $v(s) = (U_\infty v)(s)$, $s \in S$ by definition, where ($\infty$ represents the stopping time that is $\infty$ with probability one). Hence by th. 3.7 v is u.s.a. By th. 3.11 we find, for all $b \in \mathcal{B}_m(S)$

$$v = U_{\tau\circ\infty} b = U_\tau(U_\infty b) = U_\tau v .$$

Finally we prove (iii). First let $0 \leq r \leq M$ and $0 \leq b \leq M$ for $M \in \mathbb{R}$. Note that $\tau^n \geq n$ on $\Omega$. Then we have, for $s \in S$:

$$(*) \qquad (U_1^n 0)(s) = (U_n 0)(s) \leq (U_{\tau^n} b)(s) \leq (U_{\tau^n} 0)(s) + \beta^n M \leq v(s) + \beta^n M.$$

It is well-known that $\lim_{n\to\infty} (U_1^n 0)(s) = v(s)$ (see e.g. [Hinderer (1970), th. 14.5]). Hence we have, by (*):

$$\lim_{n\to\infty} (U_\tau^n b)(s) = v(s), \quad s \in S .$$

Further let $-M \leq r \leq M$ and $-M \leq b \leq M$ for $M \in \mathbb{R}$, and define $\tilde{r} := r + M$ and $\tilde{b} := b + M$. Let $\tilde{U}_\tau$ be the operator for the model with reward function $\tilde{r}$ instead of r, and let $\tilde{v}$ be the value function in this case. Hence we have

$$(U_\tau{}^n b)(s) + M\frac{1-\beta^{n+1}}{1-\beta} \le (\tilde{U}_\tau{}^n \tilde{b})(s) \le (U_\tau{}^n b)(s) + \frac{M}{1-\beta}$$

and we have also $\tilde{v}(s) = v(s) + \frac{M}{1-\beta}$.

Hence we find:

$$\lim_{n\to\infty} (U_\tau{}^n b)(s) = \lim_{n\to\infty} (\tilde{U}_\tau{}^n \tilde{b})(s) - \frac{M}{1-\beta} = \tilde{v}(s) - \frac{M}{1-\beta} = v(s),\ s \in S\ .\ \square$$

We conclude this section with some remarks.

Remarks.

(i) The theorems th. 3.11 and th. 3.14 can easily be generalized to models with weaker conditions on the reward functions. In fact th. 3.14(iii) is valid for models where a *strong convergence condition* is satisfied (cf. [Van Hee, Hordijk and Van der Wal (1977)]).

(ii) If there is a (nonempty) subset $B(S)$ of $B_m(S)$ such that for all $\tau \in \Sigma$ $U_\tau b \in B(S)$ if $b \in B(S)$ then $\{U_\tau, \tau \in \Sigma\}$ is a *semi-group* of operators on $B(S)$. If S is countable then the set of all real-valued functions on S that are bounded from above will do for $B(S)$. In section 3.3 we show that there is such a set $B(S)$ for the equivalent dynamic program (model 2).

(iii) There need not be, for each $b \in B_a(S)$, $\rho \in P(S)$ and $\varepsilon > 0$ a strategy $\pi \in \Pi$ such that:

$$(*)\qquad \rho(\{s \in S \mid E_s^\pi[\sum_{n=0}^{\tau-1} \beta^n r(S_n,A_n) + \beta^\tau b(S_\tau)] \ge (U_\tau b)(s) - \varepsilon\}) = 1\ .$$

However, if $b = U_\sigma \tilde{b}$ for some $\tilde{b} \in B_m(S)$ and $\sigma \in \Sigma$ then for each $\rho \in P(S)$ and each $\varepsilon > 0$ there is a $\pi \in \Pi$ such that (*) holds. To verify this, note that by th. 3.11 $U_\tau b = U_{\tau\circ\sigma}\tilde{b}$. Hence by lemma 3.9 we have the desired property.

(iv) Th. 3.14(ii) and (iii) are also proved in [Van Nunen and Wessels (1976)] by use of the fixed point theorem for contraction mappings, for dynamic programs with countable state space.

## 3.3 *Miscellaneous results for the Bayesian control model*

In this section we first study the optimal reward operators for the Bayesian control model (model 1). We show that these operators applied to functions that are lower semi-continuous (l.s.c.) in the second coordinate, i.e. l.s.c.

on W, yield functions that are again l.s.c. in the second coordinate. In the rest of this section we consider the value function (cf. 2.12) in more detail. We show that the value function v is convex in the second coordinate. Finally, we consider another consequence of the convexity of v, namely an upperbound on the value function. We note that the second part of this section is independent of the first part.
(Remember that the symbols used in section 3.2 have a local meaning only.)

3.25 The set of stopping times $\Sigma$ for the Bayesian control model consists of all measurable functions $\tau$ from $\Omega$ to $\mathbf{N}$ such that

$$\{\tau = n\} \in \sigma(X_0, A_0, Y_1, X_1, \ldots, Y_n, X_n), \; n \in \mathbf{N} .$$

3.26 $\mathcal{B}_\ell(X \times W)$ is the set of real-valued bounded measurable functions on $X \times W$, which are lower semi-continuous (l.s.c.) on W (cf. appendix A).

3.27 For each $\tau \in \Sigma$ we define the optimal reward operator $U_\tau$ on $\mathcal{B}_\ell(X \times W)$ by (cf. 3.16)

$$(U_\tau b)(x,q) := \sup_{\pi \in \Pi_0} E^{\pi}_{x,q}\Big[\sum_{n=0}^{\tau-1} \beta^n r(X_n, A_n, Y_{n+1}) + \beta^\tau b(X_\tau, Q_\tau)\Big]$$

where $(x,q) \in X \times W$ and $b \in \mathcal{B}_\ell(X \times W)$.

Note that, if $\tau \in \Sigma$ satisfies the property:

$$\{\tau = n\} \in \sigma(X_0, A_0, X_1, A_1, \ldots, X_n), \; n \in \mathbf{N}$$

then there is a stopping time $\tilde{\tau}$ for the equivalent dynamic program (model 2), such that for all $x_i \in X$, $y_i \in Y$, $a_i \in A$, $q_i \in W$ and $i \in \mathbf{N}$:

$$\tilde{\tau}(x_0, q_0, a_0, x_1, q_1, a_1, \ldots) = \tau(x_0, a_0, y_1, x_1, a_1, \ldots) .$$

Then the optimal reward operator $\tilde{U}_{\tilde{\tau}}$ for the equivalent dynamic program, defined by 3.16, is *equivalent* to $U_\tau$ in the following sense:

$$(U_\tau b)(x,q) = (\tilde{U}_{\tilde{\tau}} b)(x,q) \quad \text{for all } x \in X, \; q \in W \text{ and } b \in \mathcal{B}_\ell(X \times W).$$

Theorem 3.15

Let r be bounded and let $\theta_i \to p_i(y|\theta_i)$ be bounded and continuous for all $i \in I$ and $y \in Y$. Further let $\tau \in \Sigma$. Then $b \in \mathcal{B}_\ell(X \times W)$ implies $U_\tau b \in \mathcal{B}_\ell(X \times W)$.

Proof.

(a) Fix the prior distribution $q \in W$, and let $Q_n$ be the posterior distribution at stage n. Then we have: $r(X_n,A_n,Y_{n+1})1_{\{\tau>n\}} + b(X_n,Q_n)1_{\{\tau=n\}}$ is measurable with respect to the $\sigma$-field $\sigma(X_0,A_0,Y_1,\ldots,A_n,Y_{n+1})$. Hence there are measurable functions $F_n: W \times (X \times A \times Y)^{n+1} \to \mathbb{R}$ such that on $\Omega$:

$$F_n(q,X_0,A_0,Y_1,\ldots,X_n,A_n,Y_{n+1}) = r(X_n,A_n,Y_{n+1})1_{\{\tau>n\}} + b(X_n,Q_n)1_{\{\tau=n\}}$$

(cf. lemma 1.1). Hence we have

$$\sum_{n=0}^{\tau-1} \beta^n r(X_n,A_n,Y_{n+1}) + \beta^\tau b(X_\tau,Q_\tau) = \sum_{n=0}^{\infty} \beta^n F_n(q,X_0,A_0,Y_1,\ldots,X_n,A_n,Y_{n+1}).$$

(b) Further we show that $F_n$ is l.s.c. in the first coordinate. To this end we first prove that $Q_n$ is a continuous function of q in the sence of weak convergence. Let the sequence $\{q_k,\ k \in \mathbb{N}\} \subset W$ converge weakly to $q \in W$ (cf. 1.22) (notation: $q_k \xrightarrow{W} q$). Fix $(x_0,a_0,Y_1,\ldots,Y_n,x_n) \in X \times A \times (Y \times X \times A)^{n-1} \times Y \times X$ and let:

$$g_n(q) := q_n(q,x_0,a_0,Y_1,\ldots,Y_n) \quad \text{(cf. 2.30(i))}.$$

We have to show that $g_n(q_k) \xrightarrow{W} g_n(q)$. Let f be a bounded and continuous function on $\theta$. Notice that, by 2.30:

$$\int f(\theta)g_n(q_k)(d\theta) = \int f(\theta) \prod_{j=0}^{n-1} \sum_{i\in I} 1_{K_i}(x_j,a_j)p_i(Y_{j+1}|\theta_i)q_k(d\theta)$$

$$\cdot \{\Delta_n(q_k,x_0,a_0,Y_1,\ldots,Y_n)\}^{-1}$$

where

$$\Delta_n(q_k,x_0,a_0,Y_1,\ldots,Y_n) := \int \prod_{j=0}^{n-1} \sum_{i\in I} 1_{K_i}(x_j,a_j)p_i(Y_{j+1}|\theta_i)q_k(d\theta).$$

Hence, since

$$\theta \to \prod_{j=0}^{n-1} \sum_{i\in I} 1_{K_i}(x_j,a_j)p_i(Y_{j+1}|\theta_i)$$

is bounded and continuous, we have $\int f(\theta)g_n(q_k)(d\theta)$ tends to $\int f(\theta)g_n(q)(d\theta)$ if k tends to infinity, provided that $\Delta_n(q,x_0,a_0,Y_1,\ldots,Y_n) > 0$.
Hence $q \to b(x_n,g_n(q))$ is l.s.c. (cf. A 15) and so
$q \to F_n(q,x_0,a_0,Y_1,\ldots,x_n,a_n,Y_{n+1})$ is l.s.c. if $\Delta_n(q,x_0,a_0,Y_1,\ldots,Y_n) > 0$.
Consequently:

$$q \to F_n(q,x_0,a_0,y_1,\ldots,x_n,a_n,y_{n+1})\Delta_{n+1}(q,x_0,a_0,y_1,\ldots,y_{n+1})$$

is l.s.c. since $\Delta_n(q,x_0,a_\theta,y_1,\ldots,y_n) = 0$ implies
$\Delta_{n+1}(q,x_0,a_0,y_1,\ldots,y_{n+1}) = 0$.

(c) Next we show that

$$q \to \mathbf{E}^{\pi}_{x,q}[F_n(q,X_0,A_0,Y_1,\ldots,Y_{n+1})] \text{ is l.s.c.}$$

Note that

$$(*)\qquad \mathbf{E}^{\pi}_{x,q}[F_n(q,X_0,A_0,Y_1,\ldots,Y_{n+1})] =$$

$$= \int \pi_0(da_0|x_0)\int \nu(dy_1)\int P(dx_1|x_0,a_0,y_1)\ldots\int P(dx_n|x_{n-1},a_{n-1},y_n)$$

$$\cdot\int \pi_n(da_n|x_0,a_0,y_1,\ldots,y_n,x_n)\int \nu(dy_{n+1})F_n(q,x_0,a_0,y_1,\ldots,y_{n+1})$$

$$\cdot\ \Delta_{n+1}(q,x_0,a_0,y_1,\ldots,y_{n+1}).$$

Since $F_n$ is bounded, it follows from Fatou's lemma, applied in (*) that

$$(**)\qquad \liminf_{k\to\infty}\mathbf{E}^{\pi}_{x,q_k}[F_n(q_k,X_0,A_0,Y_1,\ldots,Y_{n+1})] \geq \mathbf{E}^{\pi}_{x,q}[F_n(q,X_0,A_0,Y_1,\ldots,Y_{n+1})].$$

(d) Finally we consider $q \to (U_\tau b)(x,q)$. Let the sequence $\{q_k,\ k\in\mathbb{N}\}$ converge weakly to $q \in W$. Again by Fatou's lemma we have:

$$\liminf_{k\to\infty}\sum_{n=0}^{\infty}\beta^n\mathbf{E}^{\pi}_{x,q_k}[F_n(q_k,X_0,A_0,Y_1,\ldots,Y_{n+1})] \geq$$

$$\geq \sum_{n=0}^{\infty}\beta^n\liminf_{k\to\infty}\mathbf{E}^{\pi}_{x,q_k}[F_n(q_k,X_0,A_0,Y_1,\ldots,Y_{n+1})] \geq$$

$$\geq \sum_{n=0}^{\infty}\beta^n\mathbf{E}^{\pi}_{x,q}[F_n(q,X_0,A_0,Y_1,\ldots,Y_{n+1})]\ .$$

Note that

$$(U_\tau b)(x,q) = \sup_{\pi\in\Pi_0}\mathbf{E}^{\pi}_{x,q}[\sum_{n=0}^{\infty}\beta^n F_n(q,X_0,A_0,Y_1,\ldots,Y_{n+1})]\ .$$

Hence, by A16, we have $q \to (U_\tau b)(x,q)$ is l.s.c. □

It is an immediate consequence of th. 3.15 that the function

$$q \to v(x,q)$$

is l.s.c. if r is bounded.

We conclude this section with two convexity properties of the value function.

**Theorem** 3.16

The value function of the Bayesian control model (cf. 2.12) is convex on W, i.e.

3.28 $$v(x,\lambda q_1 + (1-\lambda)q_2) \le \lambda v(x,q_1) + (1-\lambda)v(x,q_2),\ x \in X,\ q_1,q_2 \in W$$

and $\lambda \in (0,1)$. $((\lambda q_1 + (1-\lambda)q_2)(B) := \lambda q_1(B) + (1-\lambda)q_2(B)$ for $B \in \mathcal{T}$.)
Further v satisfies the inequality:

3.29 $$v(x,q) \le \int q(d\theta)v(x,\theta) \quad \text{for } x \in X,\ q \in W .$$

**Proof.**

Fix $\lambda \in (0,1)$ and $q_1,q_2 \in W$. Then we have for all $x \in X$ (cf. 2.12):

$$v(x,\lambda q_1 + (1-\lambda)q_2) = \sup_{\pi \in \Pi} \int (\lambda q_1 + (1-\lambda)q_2)(d\theta)v(x,\theta,\pi) \le$$

$$\le \lambda \sup_{\pi \in \Pi} \int q_1(d\theta)v(x,\theta,\pi) + (1-\lambda)\sup_{\pi \in \Pi} \int q_2(d\theta)v(x,\theta,\pi) =$$

$$= \lambda v(x,q_1) + (1-\lambda)v(x,q_2) .$$

We proceed with 3.29. Let $q \in W$ and $x \in X$. Then

$$v(x,q) = \sup_{\pi \in \Pi} \int q(d\theta)v(x,\theta,\pi) \le \int q(d\theta) \sup_{\pi \in \Pi} v(x,\theta,\pi) = \int q(d\theta)v(x,\theta). \square$$

**Remark.**

The inequality 3.29 is a direct consequence of the convexity of $q \to v(x,q)$ in case this function is continuous.
Namely, if F is a continuous and convex function on W which is bounded from above then the following inequality can be proved:

$$F(q) \le \int F(\theta)q(d\theta)$$

(remember that we have embedded $\Theta$ in W).

# 4. *BAYESIAN EQUIVALENT RULES AND THE AVERAGE-RETURN CRITERION*

In section 4.1 we consider procedures to construct good strategies. Special attention is given to strategies that are generated by so-called Bayesian equivalent rules. In section 4.2 we consider the average-return criterion and we give sufficient conditions for the existence of optimal strategies, based on such rules. In chapter 5 these rules will be considered in connection with the total-return criterion.

## 4.1 *Bayesian equivalent rules and other approaches*

We first consider the total discounted return criterion for models with finite sets X, Y and A. If the parameter value $\theta$ is known the usual technique to determine an optimal strategy is solving the optimality equation $v(x,\theta) = (Uv)(x,\theta)$ for all $x \in X$. Then each Markov policy for which the maximum in this equation is attained, is optimal (cf. [Blackwell (1965) th. 7]).It seldom happens that an analytic solution of the optimality equation can be found and that the value function $\theta \to v(x,\theta)$ , $x \in X$ is found in an explicit form. In chapter 1 we noted that even if $\Theta$ is a finite set, the equivalent dynamic program (model 2) has an uncountable state space $X \times W$ and for each starting state $(x,q) \in X \times W$ there is a countable subset of $X \times W$ that can be reached in the long run. Hence it is even impossible to determine the value function v for all $(x,q) \in X \times W$. However there are rather complicated algorithms to determine $v(x,q)$ for any fixed pair $(x,q) \in X \times W$ (cf. chapters 6 and 7).

Hence it is possible to determine in each state $(x,q) \in X \times W$ the action $f(x,q)$, corresponding to an optimal Bayesian Markov policy f in the following way. First determine $v(x',T_{i,y}(q))$ for all $x' \in X$, $y \in Y$ and all $i \in I$ for which there is an $a \in D(x)$ with $(x,a) \in K_i$. Then $f(x,q)$ is maximizing the function

$$4.1 \quad a \to \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y,q) \{ r(x,a,y) + \beta \int P(dx' \mid x,a,y) v(x',T_{i,y}(q)) \}$$

on the set $D(x)$, $(x,q) \in X \times W$ (cf. the remarks at the end of section 3.1).

Since this is in general a very complicated procedure, it would be preferable to have a simple recipe to determine in each state $(x,q) \in X \times W$ an action that corresponds to a good, not necessarily optimal, strategy.

For example, in practice one often uses the following recipe:

4.2 *At each stage estimate the unknown parameter* $\theta$ *using the available data, by* $\hat{\theta}$. *Then compute an (nearly) optimal strategy for the model where the parameter is known and equal to* $\hat{\theta}$. *Then use the action corresponding to this strategy in the actual state. Repeat this procedure at the next stage.*

The computation of an optimal strategy for a fixed parameter value is carried out much faster than the determination of $v(x', T_{i,y}(q))$ for $x' \in X$, $y \in Y$ and the relevant $i \in I$. Hence the recipe 4.2 is simpler than the procedure given by 4.1. However, the strategy specified in 4.2 is not optimal in general. Under some conditions it is optimal when we are dealing with the average-return criterion (cf. section 4.2).

We consider the following method to construct simple recipes. Now we consider the average-return criterion too. If $\theta \in \Theta$ is known, an action corresponding to an (nearly) optimal strategy is often found by maximizing some real-valued function F on $X \times \Theta \times A$, over all available actions $a \in D(x)$. For example, if A is finite and r is bounded then we may define F by

4.3 a $\quad F(x,\theta,a) :=$

$$\sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) + \beta \int P(dx'|x,a,y) v(x',\theta)\} \quad ,$$

in case we work with the total discounted return criterion (cf. 4.1). Sometimes there exist bounded measurable functions h and g such that

4.3 b $\quad h(x,\theta) + g(\theta) =$

$$\max_{a\in D(x)} \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) + \int P(dx'|x,a,y) h(x',\theta)\}$$

(cf. section 4.2).

Then each strategy that chooses in state x a maximizing action in the equation 4.3 b is optimal with respect to the average-return criterion (cf. section 4.2). Hence in this situation we have

$$F(x,\theta,a) := \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i) \{r(x,a,y) + \int P(dx'|x,a,y) h(x',\theta)\} .$$

We now assume that such a function F is known. In the two examples above F can be computed by standard methods, if X,Y,A and $\theta$ are finite sets (cf. chapter 7). Using this function F we construct a Bayesian Markov policy f such that for some $\varepsilon > 0$:

$$4.4 \qquad \int q(d\theta) F(x,\theta,f(x,q)) \geq \sup_{a\in D(x)} \int q(d\theta) F(x,\theta,a) - \varepsilon ,$$

for all $(x,q) \in X \times W$ .

We call such a Bayesian Markov policy a *Bayesian equivalent rule* since we are maximizing the "Bayesian equivalent" of the function we have to maximize in case the parameter is known. Note that we may choose $\varepsilon = 0$ if there is a maximizer of $a \to \int q(d\theta) F(x,\theta,a)$ for all $(x,q) \in X \times W$ (cf. the informal definition in section 1.2).

If $q \in W$ is degenerate, then the Bayesian equivalent rule is (nearly) optimal. But Bayesian equivalent rules are not optimal in general. However in section 4.2 we give sufficient conditions for optimality in case we are considering the average-return criterion, and in chapter 5 we consider examples of the Bayesian control model where a Bayesian equivalent rule is optimal for the total-return criterion.

Consider again the model with finite action space A and bounded reward function with respect to the total discounted return criterion. Then we may define a Bayesian equivalent rule using the function F defined in 4.3 a. This rule has the following interpretation. Consider a modified model where the decision-maker is told the true parameter value after one transition. It is easy to verify that this rule would be optimal in that situation.

In th. 6.4 (chapter 6) we give a lower bound on the Bayesian discounted total return of this strategy. In th. 6.3 we consider another simple recipe to construct a good strategy for the total discounted return criterion.

We conclude this section with an overview of procedures suggested by other authors for the average-return criterion.

In [Mandl (1974), (1976)] the strategy described in 4.2 is studied. Mandl used *minimum contrast estimators* and in his model the parameter structure ensures the consistency of these estimators, under each strategy. Mandl considers the following average-return criterion: a strategy $\pi$ is optimal if for all $\theta \in \Theta$ and $x \in X$

$$\liminf_{N\to\infty} \mathbb{E}^{\pi}_{x,\theta}\left[\frac{1}{N}\sum_{n=0}^{N-1} r(X_n, A_n, Y_{n+1})\right]$$

is maximal. Note that this criterion is stronger than ours and not depending on the choice of a prior distribution. Mandl shows that the strategy described in 4.2 is optimal in the model where X is a finite set, A a compact subset of a Euclidean space and where for each stationary strategy (cf. 3.9) the resulting Markov chain is irreducible. We show a similar result in section 4.2 for a Bayesian equivalent rule. In [Fox and Rolph (1973)] an optimal strategy is constructed for Markov renewal programs where also the recipe 4.2 is used. However, in their situation they have to ensure the consistency of the estimators for the unknown parameter. This problem is solved by so-called *forced choice actions*. These actions do not necessarily agree with the recipe of 4.2, but they are performed to get information. Fox and Rolph also use the stronger optimality criterion discussed above. In [Rose (1975)] another strategy is proposed. Rose assumes that for each parameter value an optimal Markov policy is known. At each stage an action is selected by randomizing over the actions belonging to some Markov policy that is optimal for some parameter value, according to the current posterior distribution. Rose also needs forced choice actions to ensure degeneration of the posterior distributions.

## 4.2 *Optimal strategies for the average-return criterion*

In this section we construct optimal strategies for the average-return criterion. Bayesian equivalent rules play an important role. We first consider an example showing that, even in case of finite state and action spaces, there need not be an optimal strategy.

**Example 4.1**

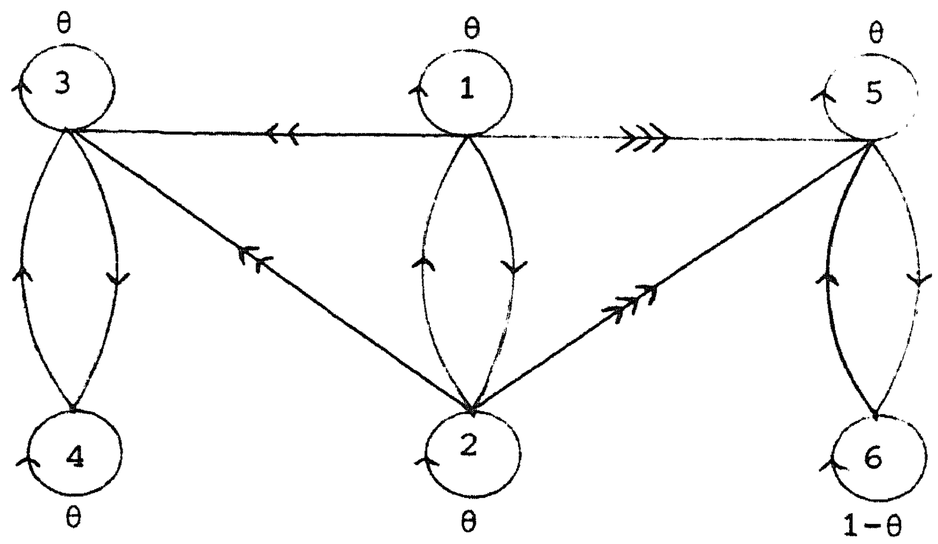
Consider the following model: $X = \{1,2,3,4,5,6\}$, $A = D(1) = D(2) = \{1,2,3\}$ $D(x) = \{1\}$ for $x \in \{3,4,5,6\}$. The transition probabilities $p(x'|x,a)$ from x

to x' if action a is chosen are:

$p(3|3,1) = p(4|4,1) = p(5|5,1) = p(5|6,1) = p(1|1,1) = p(2|2,1) = \theta$

$p(4|3,1) = p(3|4,1) = p(6|6,1) = p(6|5,1) = p(2|1,1) = p(1|2,1) = 1 - \theta$

$p(3|1,2) = p(5|1,3) = p(3|2,2) = p(5|2,3) = 1$




Only in the states 3,4,5 and 6 a reward is obtained: $r(3) = r(5) = 7$ and $r(4) = r(6) = 3$. Let $\Theta = (0,1)$. It is easy to transform this example into the framework of the Bayesian decision model.

The average return in the subchain $\{3,4\}$ is: $\frac{1}{2}(7 + 3) = 5$ and in the subchain $\{5,6\}$: $7\theta + 3(1 - \theta) = 4\theta + 3$. Consider a starting state $x \in \{1,2\}$. For fixed $\theta \in \Theta$ the optimal action is a maximizer of $5\delta(2,a) + \{4\theta + 3\}\delta(3,a)$, $a \in \{2,3\}$. Hence,the corresponding Bayesian equivalent rule for the distribution q on (0,1) is the maximizer of $5\delta(2,a) + \{4\int\theta q(d\theta) + 3\}\delta(3,a)$, $a \in \{2,3\}$. It is easy to verify that if we have to choose one of the actions 2 or 3 and if q is the prior distribution, then this Bayesian equivalent rule is the best one. Let $\pi^n$ be the strategy that chooses action 1 the first n times and in states 1 and 2 the maximizer of $5\delta(2,a) + \{4\int \theta Q_n(d\theta) + 3\}\delta(3,a)$, $a \in \{2,3\}$ thereafter, where $Q_n$ is the posterior distribution at time n, if the system starts in state 1 with prior $q \in W$. Then the Bayesian average return in states 1 and 2 is:

$$E_q[\max\{5,\ 4\int\theta Q_n(d\theta) + 3\}]$$

(note that this expression does not depend on the starting state and the strategy). Note that:

$$\mathbb{E}_q[\max\{5,\ 4\int\theta\ Q_{n+1}(d\theta) + 3\}|Q_1,\ldots,Q_n] \geq$$

$$\geq \max\{5,4\ \mathbb{E}_q[\int\theta Q_{n+1}(d\theta)|Q_1,\ldots,Q_n] + 3\} = \max\{5,4\int\theta Q_n(d\theta) + 3\}$$

with equality if and only if $5 \geq 4\int\theta\ Q_{n+1}(d\theta) + 3$, $\mathbb{P}_q$-a.s. However if q gives positive mass to the set $\{\theta \in \Theta|\theta > \frac{1}{2}\}$ then equality never holds. Hence in this case the strategy $\pi^n$ is worse than $\pi^{n+1}$ and consequently there is no optimal strategy.

In this section we need the following assumptions:

Assumptions

4.5 (i) r is bounded on $X \times A \times Y$.

(ii) $D(x) = A$ for all $x \in X$.

4.6 There are bounded measurable functions g and h on $\Theta$ and $X \times \Theta$ respectively such that

$$h(x,\theta) + g(\theta) = \sup_{a\in A} L(x,\theta,a)$$

where

$$L(x,\theta,a) := \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy)p_i(y|\theta_i)\{r(x,a,y) +$$

$$+ \int P(dx'|x,a,y)h(x',\theta)\} \text{ for } x \in X,\ a \in A \text{ and } \theta \in \Theta\ .$$

4.7 For all $\varepsilon > 0$ there is a Bayesian Markov policy f such that

$$\int q(d\theta)L(x,\theta,f(x,q)) \geq \sup_{a\in A} \int q(d\theta)L(x,\theta,a) - \varepsilon \quad \text{for } (x,q) \in X \times W\ .$$

Note that assumption 4.6 is identical to 4.3b. The assumption 4.5(ii) is not essential, but it makes things more transparant. Further it seems possible to weaken assumption 4.5(i). The only serious assumption is 4.6. For models with known parameter value $\theta$ and finite action space A, assumption 4.6 guarantees the existence of a stationary optimal strategy. This has been proved for finite X in [Derman (1966)] and for arbitrary X in [Ross (1968)].

In fact the strategy that chooses the maximizer of $a \to L(x,\theta,a)$ at each stage is optimal with average return $g(\theta)$. In [Ross (1968)] several situations are given where, for fixed parameter value $\theta$ a solution of 4.6 exists. For instance, if X and A are finite and for each stationary strategy the process is an irreducible Markov chain, then 4.6 is valid. The results of Derman and Ross follow from th. 4.1 below. Assumption 4.7 is a regularity condition to guarantee a measurable selection. In lemmas 4.4 and 4.5 we give some sufficient conditions guaranteeing 4.7.
Note that a Bayesian Markov policy satisfying 4.7 is a Bayesian equivalent rule.
In th. 4.1 we derive a sufficient condition for a strategy to be optimal. In the rest of this section we consider model assumptions which guarantee this condition for strategies generated by the Bayesian equivalent rules of the form 4.7.
Remember that the functions g and h are easy to compute by standard methods (cf. [Derman (1970)]) if 4.6 holds and X and A are finite sets.

First we introduce some notations:

4.8 (i) $\phi(x,\theta,a) := L(x,\theta,a) - h(x,\theta) - g(\theta)$ , $x \in X$, $\theta \in \Theta$ and $a \in A$.

(ii) $\phi(x,q,a) := \int q(d\theta)\phi(x,\theta,a)$ , $x \in X$, $q \in W$ and $a \in A$.

4.9 (i) $h(x,q) := \int q(d\theta)h(x,\theta)$ , $x \in X$ and $q \in W$ .

(ii) $g(q) := \int q(d\theta)g(\theta)$ , $q \in W$ .

The definitions 4.8(i) and (ii) are consistent, since we embedded $\Theta$ in W, similarly the definitions of h and g in 4.9 are consistent.

Theorem 4.1

Assume 4.5, 4.6 and the existence of a strategy $\pi^* \in \Pi_0$ such that

4.10 $$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi^*}_{x,q}[\phi(X_n,Q_n,A_n)] = 0 .$$

Then:

$$\sup_{\pi\in\Pi_0} \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,q}[\, r(X_n,A_n,Y_{n+1})] = g(q)$$

and $\pi^*$ is optimal.

Proof.

First note that

$$\int q(d\theta)\{ \sum_{i\in I} 1_{K_i}(x,a)\, p_i(y|\theta_i) h(x',\theta)\} = \sum_{i\in I} 1_{K_i}(x,a)\, p_i(y,q) h(x', T_{i,y}(q)) =$$

$$= \sum_{i\in I} 1_{K_i}(x,a)\, p_i(y,q) h(x', \sum_{i\in I} 1_{K_i}(x,a)\, T_{i,y}(q))$$

since by 4.9(i) and 2.28

$$h(x', T_{i,y}(q)) = \{ \int q(d\theta)\, p_i(y|\theta_i) h(x',\theta)\}\{p_i(y,q)\}^{-1} \quad , \quad \text{if } p_i(y,q) > 0.$$

Hence, by definitions 4.8(ii) and 3.1(e) we have for all $\pi \in \Pi_0$:

$$\phi(X_n,Q_n,A_n) = \tilde{r}(X_n,Q_n,A_n) + \mathbb{E}^{\pi}_{x,q}[h(X_{n+1},Q_{n+1}) \,|\, X_n,Q_n,A_n] - h(X_n,Q_n) - g(Q_n).$$

Therefore, by first conditioning on $\sigma(X_n,Q_n,A_n)$, we have for all $N \in \mathbb{N}^*$

$$\sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,q}[\tilde{r}(X_n,Q_n,A_n) + h(X_{n+1},Q_{n+1}) - h(X_n,Q_n) - g(Q_n) - \phi(X_n,Q_n,A_n)] = 0$$

Since $g(Q_n) = \int Q_n(d\theta) g(\theta)$, we have (cf. th. 2.1)

$$\mathbb{E}^{\pi}_{x,q}[g(Q_n)] = g(q) \ .$$

Using the boundedness of h we find for $\pi \in \Pi_0$:

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,q}[\tilde{r}(X_n,Q_n,A_n)] = g(q) + \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,q}[\phi(X_n,Q_n,A_n)] \ .$$

Note that $\phi(x,\theta,a) \le 0$ by 4.6. Hence, for all $\pi \in \Pi_0$, $g(q)$ is an upperbound for the Bayesian average return. On the other hand, if 4.10 holds, then $g(q)$ is the optimal value and $\pi^*$ is optimal. □

**Remarks.**

(i) In [Mandl (1974) th. 3] a similar result has been obtained.
In fact, Mandl's result, formulated in our terminology, reads:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} r(X_n, A_n, Y_{n+1}) = g(\theta) \quad , \qquad \mathbb{P}^{\pi}_{x,\theta}\text{-a.s.}$$

if and only if

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \phi(X_n, \theta, A_n) = 0 \quad , \qquad \mathbb{P}^{\pi}_{x,\theta}\text{-a.s.}$$

(note that one limit exists iff the other exists).

(ii) We conclude from th. 4.1 that if there is a $\varepsilon > 0$ such that for all $\pi \in \Pi_0$

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi}_{x,q}[\phi(X_n, Q_n, A_n)] \geq \varepsilon \ ,$$

then the optimal Bayesian average return is at most equal to $g(q) - \varepsilon$, in state $(x,q) \in X \times W$.

(iii) According to corollary 3.6 we may replace $\Pi_0$ by $\Pi$ in th. 4.1.

We need the following obvious lemma.

**Lemma 4.2**

Let $\{\varepsilon_n, n \in \mathbb{N}\}$ be a sequence of bounded real numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$ then: $\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \varepsilon_n = 0$.

The following corollary to th. 4.1 includes the already mentioned results of Derman and Ross.

**Corollary 4.3**

Let $\{\varepsilon_n, n \in \mathbb{N}\}$ be a non-increasing sequence of positive numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$, and let $f_n$ be a Markov policy such that for fixed $\theta \in \Theta$:

$$L(x,\theta,f_n(x)) \geq \sup_{a\in A} L(x,\theta,a) - \varepsilon_n \ , \qquad n \in \mathbb{N} \ .$$

Then the strategy $\pi^*$ that uses Markov policy $f_n$ at stage $n$, $n \in \mathbb{N}$, is optimal for this parameter value $\theta$.

Proof.

Note that, if we start with a prior distribution which is degenerate at $\theta$, then $Q_n$ is degenerate in $\theta$ for $n \in \mathbb{N}$. Hence, using $\pi^*$ we find: $\phi(X_n,Q_n,A_n) = L(X_n,\theta,f_n(X_n)) - h(X_n,\theta) - g(\theta) \leq \varepsilon_n$ for $n \in \mathbb{N}$. Therefore, using lemma 4.2

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi^*}_{x,\theta}[\phi(X_n,Q_n,A_n)] \leq \liminf_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \varepsilon_n = 0 \ .$$

Now th. 4.1 applies. □

We continue with some conditions guaranteeing 4.7. Note that if X is countable then 4.7 is fulfilled.

Lemma 4.4

If A is countable then 4.7 is valid.

Proof.

Fix $\varepsilon > 0$. Let $a_1,a_2,\ldots$ be an enumeration of A and define

$$B_1 := \bigcap_{a\in A} \{(x,q) \in X \times W \mid \int q(d\theta)L(x,\theta,a_1) \geq \int q(d\theta)L(x,\theta,a) - \varepsilon\}$$

and, for $k = 2,3,\ldots$

$$B_k := \bigcap_{a\in A} \{(x,q) \in X \times W \mid (x,q) \notin \bigcup_{i=1}^{k-1} B_i, \int q(d\theta)L(x,\theta,a_k) \geq \int q(d\theta)L(x,\theta,a) - \varepsilon\}.$$

Note that $B_k$ is measurable, for $k \in \mathbb{N}^*$ and $B_k \cap B_\ell = \emptyset$ if $k \neq \ell$. Further note that for each $(x,q)$ there is at least one $k \in \mathbb{N}^*$ such that $(x,q) \in B_k$. Hence the function $f : X \times W \to A$ defined by $f(x,q) := a_k$ iff $(x,q) \in B_k$ is a Bayesian Markov policy satisfying 4.7. □

Lemma 4.5

Let the following assumptions hold:

4.11 (i) A is compact.

(ii) $a \to P(\cdot|x,a,y)$ is a continuous mapping from A to $P(X)$, where $P(X)$ is endowed with the weak topology, for all $x \in X$ and $y \in Y$.

(iii) $a \to r(x,a,y)$ is continuous for all $x \in X$ and $y \in Y$.

(iv) $x \to h(x,\theta)$ is continuous for all $\theta \in \Theta$.

(v) $a \to 1_{K_i}(x,a)$ is continuous for all $x \in X$ and $i \in I$.

Then there is a Bayesian Markov policy f such that

$$\int q(d\theta)L(x,\theta,f(x,q)) = \sup_{a\in A} \int q(d\theta)L(x,\theta,a) \quad , \quad \text{for } (x,q) \in X \times W \ .$$

Proof.

Since h and r are bounded, we have $(x,\theta,a) \to L(x,\theta,a)$ is bounded. And, since this mapping is measurable, we have $(x,q,a) \to \int q(d\theta)L(x,\theta,a)$ is bounded and measurable.

Further, since $a \to \int \nu(dy)p_i(y|\theta_i)r(x,a,y)$ and $a \to \int P(dx'|x,a,y)h(x',\theta)$ are continuous, we have $a \to \int q(d\theta)L(x,\theta,a)$ is continuous. Hence all conditions for Schäl's, selection theorem (cf. A17) are satisfied, which proves the lemma. □

Remark.

The condition 4.11(v) is fulfilled in the following situation:

4.12 $A := \cup_{k=1}^{n} N_k$, where $N_k$ is compact with $N_k \cap N_\ell = \emptyset$ if $k \neq \ell$ ,

$X := \cup_{k=1}^{m} M_k$, where $M_k$ is measurable and $M_k \cap M_\ell = \emptyset$ if $k \neq \ell$ ,

and $K_{(i,j)} := M_i \times N_j$ and $I := \{(i,j) \mid i = 1,\ldots,m,\ j = 1,\ldots,n\}$ .

If A is finite then 4.12 is valid, and therefore 4.11(i), (ii) and (iii). Another example of 4.11(v) is the situation where A is compact and $K_i = M_i \times A$, $i \in \mathbb{N}^*$ where $M_1, M_2, \ldots$ is a measurable partition of X.

Theorem 4.6

Assume 4.5, 4.6 and 4.7. Let $\{\varepsilon_n, n \in \mathbb{N}\}$ be a nonincreasing sequence of positive numbers such that $\lim_{n\to\infty} \varepsilon_n = 0$. Let $f_n$ be a Bayesian Markov policy for $n \in \mathbb{N}$, such that for $(x,q) \in X \times W$:

4.13 $$\int q(d\theta)L(x,\theta,f_n(x,q)) \geq \sup_{a\in A} \int q(d\theta)L(x,\theta,a) - \varepsilon_n \quad \text{(cf. 4.7)}.$$

Further let $\pi^*$ be the strategy that uses $f_n$ at stage n, $n \in \mathbb{N}$. Assume for fixed $(x_0,q_0) \in X \times W$:

4.14 $$\mathbb{P}^{\pi^*}_{x_0,q_0}[\bigcap_{i\in I}\bigcap_{n\in\mathbb{N}} {}_*\{\tau(i,n) < \infty\}] = 1 .$$

Finally assume the existence of a finite set F of Markov policies such that

$\theta \to \inf_{x\in X} \phi(x,\theta,f(x))$ is measurable for $f \in F$ and

4.15 $$\max_{f\in F} \inf_{x\in X} \phi(x,\theta,f(x)) = 0 , \qquad \text{for all } \theta \in \Theta .$$

Then condition 4.10 holds and therefore $\pi^*$ is optimal with Bayesian average return $g(q_0)$, in $(x_0,q_0)$ (cf. 2.13).

Proof.

Note that $\pi^* \in \Pi_0$. Under the strategy $\pi^*$ we have $A_n = f_n(X_n,Q_n)$ and therefore, by 4.6 and 4.8:

$$0 \geq \phi(X_n,Q_n,A_n) \geq \sup_{a\in A} \int Q_n(d\theta)\phi(X_n,\theta,a) - \varepsilon_n \geq$$

$$\geq \max_{f\in F} \int Q_n(d\theta)\phi(X_n,\theta,f(X_n)) - \varepsilon_n \geq \max_{f\in F} \int Q_n(d\theta) \inf_{x\in X} \phi(x,\theta,f(x)) - \varepsilon_n .$$

Using 4.14, the boundedness of $\phi$ and corollary 2.5 we find

$$\lim_{n\to\infty} \int Q_n(d\theta) \inf_{x\in X} \phi(x,\theta,f(x)) = \inf_{x\in X} \phi(x,Z,f(x)) \quad , \ \mathbb{P}^{\pi^*}_{x_0,q_0}\text{-a.s.}$$

And since F is a finite set we have by 4.15 $\mathbb{P}^{\pi^*}_{x_0,q_0}$-a.s.

$$\lim_{n\to\infty} \max_{f\in F} \int Q_n(d\theta) \inf_{x\in X} \phi(X,\theta,f(x)) = \max_{f\in F} \inf_{x\in X} \phi(x,Z,f(x)) = 0 .$$

Since $\lim_{n\to\infty} \varepsilon_n = 0$ we finally have $\lim_{n\to\infty} \phi(X_n,Q_n,A_n) = 0$, $\mathbb{P}^{\pi^*}_{x_0,q_0}$-a.s. And therefore, by the boundedness of $\phi$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{E}^{\pi^*}_{x_0,q_0}[\phi(X_n,Q_n,A_n)] = 0 \ .$$

This proves the theorem. □

Although, at first glance the number of assumptions in th. 4.6 is overwhelming, only 4.14 and 4.15 are serious restrictions on the applicability of the theorem. In 4.14 it is required that the strategy $\pi^*$ guarantees that we obtain enough "information" concerning the "true" parameter. If there is a finite collection of Markov policies, which contains an optimal one for all models with known parameter value, then 4.15 is fulfilled. In th. 4.8 we consider more appealing conditions guaranteeing all requirements of th. 4.6. We start with a lemma, the truth of which is intuitively clear.

Lemma 4.7

Let X and A be finite sets and assume that, for all $\theta \in \Theta$ and each stationary strategy, the resulting Markov chain $\{X_n, n \in \mathbb{N}\}$ is irreducible. Then, for all $\theta \in \Theta$, $x \in X$ and $\pi \in \Pi$, the number of visits to each state $x' \in X$ is infinite, $\mathbb{P}^{\pi}_{x,\theta}$-a.s.

Proof.

Fix $\theta \in \Theta$ and $x' \in X$.

a) We first prove there is at least one visit, $\mathbb{P}^{\pi}_{x,\theta}$-a.s. To show this, transform the transition law in such a way that x' becomes absorbing: i.e. $P(\{x'\}|x',a,y) := 1$ for all $a \in A$, $y \in Y$. Further consider the reward function:

$$r(x,a) := \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) P(\{x'\}|x,a,y) p_i(y|\theta_i), \quad \text{if } x \neq x'$$

$$:= 0 \quad \text{if } x = x'$$

In this model the total expected return is defined as usual:

$$v(x,\theta,\pi) := \sum_{n=0}^{\infty} \mathbb{E}^{\pi}_{x,\theta}[r(X_n,A_n)] \ , \qquad \pi \in \Pi \text{ and } x \in X \ .$$

According to th. 3.4 there is a strategy $\pi^* \in \Pi_0$ such that $v(x,\theta,\pi) = v(x,\theta,\pi^*)$. However since we start with a degenerate prior distribution, all posterior distributions are degenerate. Hence $\Pi_0$ is the

set of strategies such that the action at time n only depends on $X_0,A_0,\ldots,X_n$ (cf. the remark following corollary 3.6). Now we restrict ourselves to strategies in $\Pi_0$. Note that $\mathbb{P}^{\pi}_{x,\theta}$-a.s.

$$r(X_n,A_n) = \mathbb{P}^{\pi}_{x,\theta}[X_{n+1} = x' \,|\, X_n,A_n] \quad , \quad \text{if } X_n \neq x' \ .$$

Hence, for $x \neq x'$, $\pi \in \Pi_0$:

$$v(x,\theta,\pi) = \sum_{n=0}^{\infty} \mathbb{E}^{\pi}_{x,\theta}[\mathbb{P}^{\pi}_{x,\theta}[X_{n+1} = x' \,|\, X_n,A_n] 1_{X\setminus\{x'\}}(X_n)] =$$

$$= \sum_{n=0}^{\infty} \mathbb{P}^{\pi}_{x,\theta}[X_{n+1} = x',\ X_n \neq x'] = \mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ for some } n].$$

Since this last probability does not depend on the transition law in $x'$, the probability $\mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ for some } n]$ is not affected by the transformation of the model. We return to the transformed model. Minimizing $v(x,\theta,\pi)$ over all $\pi \in \Pi_0$ is a *negative dynamic programming* problem with state space X and action space A. Therefore we have, by [Strauch (1966) th. 9.1] for all $\pi \in \Pi_0$ $v(x,\theta,\pi) \geq \min_{\bar{\pi}} v(x,\theta,\bar{\pi})$, where the minimum has to be taken over all stationary strategies $\bar{\pi}$ in $\Pi_0$.

Since by assumption the Markov chain is irreducible under each stationary strategy, this minimum equals one, if $x \neq x'$. Therefore $v(x,\theta,\pi) = 1$ for all $\pi \in \Pi$ and $x \neq x'$. So we have in the original model for $x \neq x'$:

(*) $\mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ for some } n > 0] = 1$ for all $\pi \in \Pi$ .

It is easy to verify that (*) is also valid for $x = x'$ in the original model.

b) Consider the original model. By conditioning on the first visit to $x'$ we obtain:

$$\mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ for at least two numbers } n > 0] =$$

$$= \sum_{k=1}^{\infty} \sum_{\substack{x_1,\ldots,x_{k-1} \\ x_j \neq x'}} \sum_{a_0,\ldots,a_{k-1}} \int \nu(dy_1) \ldots \int \nu(dy_k) \prod_{j=0}^{k-1} \{P(\{x_{j+1}\} \,|\, x_j,a_j,y_{j+1})$$

$$\cdot \sum_{i\in I} 1_{K_i}(x_j,a_j) p_i(y_{j+1} \,|\, \theta_i) \pi_j(\{a_j\} \,|\, x_0,a_0,y_1,x_1,\ldots,y_j,x_j)\}$$

$$\cdot \mathbb{P}^{\pi}_{x,\theta}[X_{n+k} = x' \text{ for some } n > 0 \,|\, X_0=x,\ A_0=a_0,\ Y_1=y_1,\ X_1=x_1,\ldots,Y_k=y_k,\ X_k=x'].$$

Analogous to the construction in 3.22, there is for each

$(a_0,y_1,x_1,\ldots,y_k) \in (A \times Y \times X)^{k-1} \times A \times Y$ a strategy $\tilde{\pi} \in \Pi$ such that

$$\mathbb{P}^{\pi}_{x,\theta}[X_{n+k}=x' \text{ for some } n>0 \mid X_0=x,\ A_0=a_0,\ Y_1=y_1,\ X_1=x_1,\ldots,Y_k=y_k,\ X_k=x'] =$$

$$= \mathbb{P}^{\tilde{\pi}}_{x',\theta}[X_n=x' \text{ for some } n>0] \ .$$

This last term is equal to 1 , by part (a). Therefore there are, $\mathbb{P}^{\pi}_{x,\theta}$-a.s., at least two visits to x'. Repeating the argument yields

$\mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ for at least k numbers } n > 0] = 1$ and therefore

$\mathbb{P}^{\pi}_{x,\theta}[X_n = x' \text{ infinitely often}] = 1$ . □

Theorem 4.8

Let X and A be finite sets and let the Markov chain $\{X_n,\ n \in \mathbb{N}\}$ be irreducible for each stationary strategy and all $\theta \in \Theta$. Let $M_1,\ldots,M_m$ be a partition of X and $K_i := M_i \times A$ , $i = 1,\ldots,m$.
Then the strategy $\pi^*$, defined in th. 4.6, is optimal.

Proof.

We only have to verify the assumptions of th. 4.6. As already noted, it has been proved in [Ross (1970) corollary 6.20] that 4.6 is true, here. Further 4.7 is a consequence of lemma 4.4. Since for each known parameter value $\theta \in \Theta$ the Bayesian control model reduces to a dynamic program with state space X and action space A (cf. the proof of lemma 4.7) we have for each $\theta \in \Theta$ an optimal stationary strategy (cf. corollary 4.3).
Since there are finitely many of these strategies 4.15 is fulfilled.
Finally, by lemma 4.7, we have $\mathbb{P}^{\pi}_{x,\theta}[X_n \in M_i \text{ infinitely often}] = 1$. Hence 4.14 holds. □

Remarks.

i) The conditions of th. 4.8 are satisfied in particular if, for $X = \{x_1,\ldots,x_m\}$ we have $M_i = \{x_i\}$ , $i = 1,\ldots,m$ and if for all $\theta \in \Theta$ , $a \in A$ and $x_i, x_j \in X$:

$$\int \nu(dy) P(\{x_j\} \mid x_i, a, y) p_i(y \mid \theta_i) > 0 \ .$$

ii) In the situation of th. 4.8 we may use at each stage a Bayesian equivalent rule maximizing $a \to \int q(d\theta) L(x,\theta,a)$ in $(x,q) \in X \times W$, since A is finite.

In th. 4.8 we assumed that, if we are in state $x \in M_i$, then the information we get after the next transition does not depend on the action chosen. In th. 4.9 we relax this assumption. We shall assume there, besides 4.6 and 4.7:

4.16 (i) $A$ is a finite set and $N_1,\ldots,N_n$ is a partition of $A$.

(ii) $M_1,\ldots,M_m$ is a measurable partition of $X$ and $K_{i,j} = M_i \times N_j$, $(i,j) \in I = \{(i,j) \mid i = 1,\ldots,m,\ j = 1,\ldots,n\}$.

(iii) For each $x \in X$, $\theta \in \Theta$ and $\pi \in \Pi_0$
$$\mathbb{P}^{\pi}_{x,\theta}[X_n \in M_i \text{ infinitely often}] = 1 , \quad i = 1,\ldots,m .$$

(iv) There is a finite set $F$ of Markov policies such that
$\theta \to \inf_{x \in X} \phi(x,\theta,f(x))$ is measurable, for $f \in F$ and
$\max_{f \in F} \inf_{x \in X} \phi(x,\theta,f(x)) = 0$ (cf. 4.15) .

First, we discuss the form of a reasonable strategy for this situation. Then, in th. 4.9, we prove the optimality of such a strategy and afterwards, in th. 4.10 we consider a practical situation where 4.16 is satisfied in a natural way.

Although 4.16(iii) guarantees that we return to $M_i$, $i = 1,\ldots,m$, infinitely often under the strategy $\pi^*$ defined in th. 4.6, it is not sure that we return to each set $K_{i,j}$ infinitely often. Hence we have to modify the strategy $\pi^*$ of th. 4.6.

The idea for the modification is found in [Mallows and Robbins (1964)]. In [Fox and Rolph (1973)] and in [Rose (1975)] this idea is worked out for Markov renewal programs and Markov decision processes respectively, in a way similar to our approach here. The idea is, that we make *forced choice actions* to guarantee that we return to all sets $K_{i,j}$ infinitely often. However, we do this with a frequency that is so low as not to influence the Bayesian average return.

We start with some preparations.

We define a (double) sequence of stopping times $\{\sigma(i,t) \mid t \in \mathbb{N},\ i = 1,\ldots,m\}$:

4.17 $\sigma(i,0)(\omega) := 0$ , $\sigma(i,t)(\omega) := \inf\{k > \sigma(i,t-1)(\omega) \mid X_k(\omega) \in M_i\}$
for $\omega \in \Omega$, $i = 1,\ldots,m$, $t \in \mathbb{N}^*$ .

Hence $\sigma(i,t)$ is the time of the $t$-th visit to set $M_i$, after stage zero.

4.18 An increasing sequence $S = (s_1,s_2,s_3,\ldots)$ of positive integers is said to be of *density zero* if

$$\lim_{k\to\infty} \sup \frac{1}{k} \max\{i \in \mathbb{N}^* | s_i \leq k\} = 0 \ .$$

Examples of such sequences are: $s_i = 2^i$, $i \in \mathbb{N}^*$, since

$$\frac{1}{k} \max\{i \in \mathbb{N}^* | 2^i \leq k\} \leq \frac{^2\log k}{k} \ ,$$

and $s_i = i^2$, since

$$\frac{1}{k} \max\{i \in \mathbb{N}^* | i^2 \leq k\} \leq k^{-\frac{1}{2}} \ .$$

We define the strategy $\pi^{**}$, which will be considered in th. 4.9, in an informal way.

4.19 Fix in each set $N_j$ an action $a_j$ for $j = 1,\ldots,n$ and fix some sequence $S = (s_1,s_2,s_3,\ldots)$ of density zero. If for $t \in \mathbb{N}$ there are $i \in \{1,\ldots,m\}$ and $k \in \mathbb{N}^*$ such that $t = \sigma(i,s_k)$ then determine $\ell \in \mathbb{N}$ such that $k = bm + \ell$ with $1 \leq \ell \leq m$, $b \in \mathbb{N}$.
In that case action $a_\ell$ is chosen at stage t.
If $t \neq \sigma(i,s_k)$ at stage t, for all $i = 1,\ldots,m$ and all $k \in \mathbb{N}^*$ then the Bayesian Markov policy $f_t$, defined in 4.13 is used to select an action.

It is straightforward to make this definition more rigorous in a way similar to definition 3.20.
Note that $\pi^{**}$ tries the actions $a_1,\ldots,a_n$ successively at the *forced choice stages:* $\sigma(i,t)$, $i = 1,\ldots,m$, $t \in S$.
Since it is assumed in 4.16(iii), that the process visits each set $M_i$ infinitely often, almost surely under every strategy, we have almost surely: $\sigma(i,s_k) < \infty$ for all $i = 1,\ldots,m$ and $k \in \mathbb{N}^*$. Hence each set $K_{i,j}$ is visited infinitely often almost surely, under all strategies.

Theorem 4.9

Assume 4.5, 4.6 and 4.16. Then the strategy $\pi^{**}$ defined in 4.19 is optimal with Bayesian average return $g(q)$ in each starting state $(x,q) \in X \times W$.

Proof.

Fix $(x,q) \in X \times W$. For notational convenience we write $\mathbb{P}$ instead of $\mathbb{P}^{\pi^{**}}_{x,q}$.
a) As noted above, we have by 4.16(iii) and definition 4.19:

$$\mathbb{P}\Big[\bigcap_{i,j\in I}\ \bigcap_{k\in\mathbb{N}^*}\{\tau((i,j),k) < \infty\}\Big] = 1 \ .$$

Using th. 2.7 for the stopping times $\sigma(i,0)$, $\sigma(i,1)$,... we find in exactly the same way as in the proof of th. 4.6 that $\mathbb{P}$-a.s.

$$(*)\qquad \lim_{\substack{k\to\infty\\ k\notin S}} \phi(X_{\sigma(i,k)},Q_{\sigma(i,k)},A_{\sigma(i,k)}) = 0 \ , \qquad \text{for } i = 1,\dots,m \ .$$

b) For notational convenience we write $B(i,k)$ instead of $\phi(X_{\sigma(i,k)},Q_{\sigma(i,k)},A_{\sigma(i,k)})$. The following assertion is easy to verify (cf. 4.17). For all $t \in \mathbb{N}^*$ there is exactly one pair $i,k$ such that $\sigma(i,k) = t$, $i \in \{1,\dots,m\}$, $k \in \{1,\dots,t\}$.

Hence it is easy to verify that on $\Omega$:

$$(**)\qquad \sum_{t=1}^{N} \phi(X_t,Q_t,A_t) = \sum_{i=1}^{m}\sum_{k=1}^{N} B(i,k)\,1_{\{\sigma(i,k)\le N\}} =$$

$$= \sum_{i=1}^{m}\sum_{k=1}^{N} B(i,k)\,1_{\{\sigma(i,k)\le N,\ k\in S\}} + \sum_{i=1}^{m}\sum_{k=1}^{N} B(i,k)\,1_{\{\sigma(i,k)\le N,\ k\notin S\}} .$$

Notice that $\sigma(i,k) \ge k$ for $k \in \mathbb{N}^*$. Hence we have

$$\#\{k \in S \mid \sigma(i,k) \le N\} \le \#\{k \in S \mid k \le N\} \ .$$

Let $M := \inf_{x,\theta,a} \phi(x,\theta,a)$. Then $-\infty < M \le 0$, since $r$ is bounded and by 4.5.
Therefore we have for $i = 1,\dots,m$:

$$\frac{1}{N}\sum_{k=1}^{N} B(i,k)\,1_{\{\sigma(i,k)\le N,\ k\in S\}} \ge \frac{M}{N}\#\{k \in S \mid k \le N\}$$

which tends to zero as $N$ tends to infinity by the definition of $S$ (cf. 4.18).
Now we consider the last term of (**). Since, by (*), $B(i,k)$ tends $\mathbb{P}$-a.s. to zero, if $k \notin S$ and if $k$ tends to infinity, we have by lemma 4.2:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} B(i,k)\,1_{\{\sigma(i,k)\le N,k\notin S\}} = 0 \ , \qquad \mathbb{P}\text{-a.s.}$$

Finally we conclude from (**):

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} \phi(X_k,Q_k,A_k) = 0 \ , \qquad \mathbb{P}\text{-a.s.}$$

And therefore 4.10 is satisfied. Hence by th. 4.1 the theorem is proved. □

In the following theorem we give more appealing conditions, which imply all conditions of th. 4.9. Hence the strategy $\pi^{**}$ defined in 4.19 is optimal here.

Theorem 4.10

Let X and A be finite sets and let the Markov chain $\{X_n, n \in \mathbb{N}\}$ be irreducible for all $\theta \in \Theta$ and each stationary strategy. Further let $I = X \times A$ and let $K_{x,a} = \{(x,a)\}$. Then the strategy $\pi^{**}$ defined in 4.19 is optimal.

The proof of this theorem proceeds along the same lines as the proof of th. 4.8.

We conclude this section with some remarks.

Remarks.

i) Consider the situation of th. 4.9.
The strategy $\pi^{**}$ is easy to handle. For each set $M_i$, $i = 1,\ldots,m$ the decision maker has to keep count of the number of visits. If this number is equal to a number in the sequence S of density zero, then he has to select the next action from $\{a_1,\ldots,a_n\}$ in cyclical order (cf. 4.19). If the number of the visits does not belong to S, then the decision maker has to compute in state (x,q) an action $a^*$, such that $\phi(x,q,a^*) = \max_{a \in A} \phi(x,q,a)$.

ii) If we are dealing with a dynamic program with finite state space X and finite action space A and if for all $x,x' \in X$ and all $a \in A$: $P(x'|x,a)$ is positive but unknown, then we can transform this model into our Bayesian control model (cf. example 2.2), and by th. 4.10 the strategy $\pi^{**}$ is optimal.

iii) Th. 4.10 is more general then the results in [Rose (1975)], since we allow arbitrary prior distributions. Further, the strategy $\pi^{**}$ is easier to handle than the strategy Rose proposes, if $\Theta$ is finite.

iv) It is not clear whether all situations considered in [Mandl (1974)] are covered by th. 4.6 or not. Mandl assumes that $\theta \to \phi(x,\theta,a)$ is continuous (cf. [Mandl (1974) th. 8]), moreover he assumes the existence of minimum contrast estimators. Although we conjecture that under the assumptions of th. 4.8 minimum contrast estimators exist, it is easy to show that under the assumptions of th. 4.10 they do not exist.

# 5. BAYESIAN EQUIVALENT RULES AND THE TOTAL-RETURN CRITERION

We do not know Bayesian equivalent rules that are optimal with respect to the discounted total-return criterion, in general. Example 5.1 shows that a "natural" Bayesian equivalent rule fails to be optimal. However, in section 5.1, we prove the optimality of a Bayesian equivalent rule for the so-called independent case, and in section 5.2 for the linear system with quadratic costs. Finally, in section 5.3, we study a simple inventory model for which a Bayesian equivalent rule is sometimes optimal. Here we also study the behaviour of this rule when it is not optimal.

## 5.1 *Preliminaries and the independent case*

In the models we study in this chapter, there is only one unknown parameter, i.e. the index set I is a singleton. This implies that the decision maker obtains information about the same parameter in each state $x \in X$, regardless of the action chosen. Since I is a singleton we shall omit the subscript i in the notations $\theta_i$, $p_i(y|\theta_i)$, $p_i(y,q)$ and $T_{i,y}(q)$. Note that $\tau(1,n) = n$ (on $\Omega$), for all $n \in \mathbb{N}$ and therefore, by lemma 2.2, the distribution of $Y_1,\ldots,Y_n$, $n \in \mathbb{N}$ only depends on the prior distribution and not on the starting state or the strategy. Hence the distribution of $Q_n$ (cf. 2.24) depends only on the prior distribution and on $Y_1,\ldots,Y_n$. For that reason we shall write $\mathbb{P}_q$ and $\mathbb{E}_q$ instead of $\mathbb{P}^{\pi}_{x,q}$ and $\mathbb{E}^{\pi}_{x,q}$, when we are dealing with the random variables $Y_n$ and $Q_n$.

We start with an example. In this example the Bayesian equivalent rule, based on the function:

$$F(x,\theta,a) := \int \nu(dy)p(y|\theta)\{r(x,a,y) + \beta \int P(dx'|x,a,y)v(x',\theta)\}$$

turns out to be non-optimal.

We remark that this example has some similarity to example 4.1.

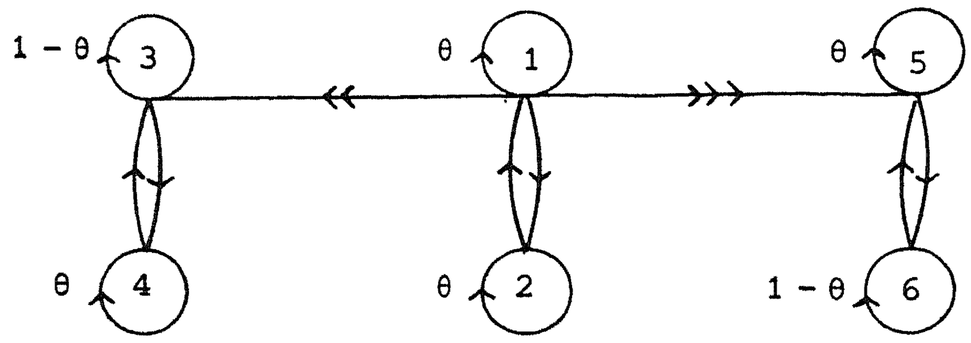
Example 5.1

Consider the following model. $X = \{1,2,\ldots,6\}$, $Y = \{0,1\}$, $D(1) = A = \{1,2,3\}$, $D(x) := \{1\}$, $x \in \{2,3,\ldots,6\}$, $\Theta := [0,1]$. The function $p(y|\theta)$ is given by: $p(1|\theta) = 1 - p(0|\theta) = \theta$, $\theta \in \Theta$. And $P(\{x'\}|x,a,y)$ is (we identify here x and $\{x\}$):

$$P(3|3,1,0) = P(3|4,1,0) = P(6|5,1,0) = P(6|6,1,0) = P(2|1,1,0) = P(1|2,1,0) = 1$$

$$P(4|3,1,1) = P(4|4,1,1) = P(5|5,1,1) = P(5|6,1,1) = P(1|1,1,1) = P(2|2,1,1) = 1$$

$P(3|1,2,y) = 1$, $P(5|1,3,y) = 1$, for $y \in Y$ .




Only in the states 3,4,5 and 6 a reward is obtained independent of $y \in Y$: $r(3) = r(5) = c$; $r(4) = r(6) = b$, $c > b \geq 0$. The prior distribution q is: $q(0) = q(1) = \frac{1}{2}$. The discounted total return $v(x,\theta)$ for discount factor $\beta$ is:

$$v(3,0) = v(5,1) = \frac{c}{1-\beta} \; ; \; v(3,1) = v(5,0) = c + \beta \frac{b}{1-\beta} \left(< \frac{c}{1-\beta}\right) .$$

Hence $v(3,\theta) = (c + \frac{\beta b}{1-\beta})\theta + \frac{c}{1-\beta}(1-\theta)$ and $v(5,\theta) = \frac{c}{1-\beta}\theta + (c + \frac{\beta b}{1-\beta})(1-\theta)$.

Further $v(2,\theta) = \frac{\beta(1-\theta)}{1-\beta\theta} v(1,\theta)$, hence

$$v(1,\theta) = \max\{\beta\theta v(1,\theta) + \beta(1-\theta)\frac{\beta(1-\theta)}{1-\beta\theta} v(1,\theta), \; \beta v(3,\theta), \; \beta v(5,\theta)\} .$$

The first term equals:

$$\beta \frac{\theta + \beta - 2\beta\theta}{1-\beta\theta} v(1,\theta) < v(1,\theta) \quad \text{for } \beta \in (0,1), \; \theta \in \Theta .$$

The Bayesian equivalent rule is based on the function F specified by:

$$F(1,\theta,1) = \beta \frac{\theta + \beta - 2\beta\theta}{1-\beta\theta} v(1,\theta), \; F(1,\theta,2) = \beta v(3,\theta)$$

and

$$F(1,\theta,3) = \beta v(5,\theta) .$$

Hence the Bayesian equivalent rule in state (1,q) chooses action 2 or 3, with equal Bayesian discounted total return: $\frac{1}{2}\beta\{c + \frac{\beta b}{1-\beta} + \frac{c}{1-\beta}\}$ . Now we consider another strategy for starting in state 1.
At stage 1 take action 1 and thereafter take the best of actions 2 and 3, in state 1. Note that under this strategy the system remains in state 1 at stage 1, or it returns to state 1 at stage 2. The discounted total return becomes: $\beta^3 \frac{c}{1-\beta}$ if $\theta = 0$ and $\beta^2 \frac{c}{1-\beta}$ if $\theta = 1$. Hence the Bayesian dis-

counted total return is: $\frac{1}{2}\frac{c}{1-\beta}\beta^2(\beta+1)$. So, if for instance $c = 2$, $b = 1$ and $\beta = 0.9$, then this strategy is better than the Bayesian equivalent rule. It is straightforward to show that the latter strategy is optimal, in this case.

We proceed with a theorem. In this theorem we show that the process of posterior distributions is a Markov chain.

Theorem 5.1

(i) Let f be a real-valued measurable function on $W \times Y$ that is bounded from above. Then, for $n \in \mathbb{N}^*$:

$$\mathbb{E}_q[f(Q_n, Y_{n+1}) \mid Y_1, \ldots, Y_n] = \int f(Q_n, y)p(y, Q_n)\nu(dy), \ \mathbb{P}_q\text{-a.s.}$$

(ii) The process $\{Q_n, n \in \mathbb{N}\}$ is a (homogeneous) Markov chain.

Proof.

(i) Define

$$q_n(d\theta) := q(d\theta)\prod_{j=1}^{n} p(y_j|\theta).\{\int q(d\theta)\prod_{j=1}^{n} p(y_j|\theta)\}^{-1}$$

if the denominator is non-zero. Let $B \in Y^n$. Using lemma 2.2 ($\tau = 1$) we have

$$(*) \qquad \int_{\{(Y_1,\ldots,Y_n)\in B\}} \mathbb{E}_q[f(Q_n, Y_{n+1}) \mid Y_1, \ldots, Y_n)]d\mathbb{P}_q =$$

$$= \int q(d\theta)\{\int \cdots \int_{(Y_1,\ldots,Y_{n+1})\in B\times Y} f(q_n, Y_{n+1})\prod_{j=1}^{n+1} p(y_j|\theta)\nu(dy_1)..\nu(dy_{n+1})\}.$$

Note that

$$\int p(Y_{n+1}|\theta)\prod_{j=1}^{n} p(y_j|\theta)q(d\theta)\{\int \prod_{j=1}^{n} p(y_j|\theta)q(d\theta)\}^{-1} = p(Y_{n+1}, q_n) \ ,$$

if the factor between braces is non-zero. Hence (*) equals:

$$\int \cdots \int_{(Y_1,\ldots,Y_{n+1})\in B\times Y} f(q_n, Y_{n+1})p(Y_{n+1}, q_n)$$

$$\cdot \{\int \prod_{j=1}^{n} p(y_j|\theta)q(d\theta)\}\nu(dy_1)\ldots\nu(dy_{n+1}) =$$

$$= \mathbb{E}_q[1_{\{(Y_1,\ldots,Y_n)\in B\}} \int f(\varrho_n,y)p(y,\varrho_n)\nu(dy)] \ ,$$

which proves part (i).

(ii) $$\mathbb{E}_q[f(\varrho_{n+1}) \mid \varrho_0,\ldots,\varrho_n] = \mathbb{E}_q[\mathbb{E}_q[f(\varrho_{n+1}) \mid Y_1,\ldots,Y_n] \mid \varrho_0,..,\varrho_n] =$$

$$= \mathbb{E}_q[\int f(T_y(\varrho_n))p(y,\varrho_n)\nu(dy) \mid \varrho_0,..,\varrho_n] = \int f(T_y(\varrho_n))p(y,\varrho_n)\nu(dy) =$$

$$= \mathbb{E}_{\varrho_n}[f(\varrho_1)] = \mathbb{E}_q[f(\varrho_{n+1}) \mid \varrho_n], \ \mathbb{P}_q\text{-a.s.}$$

The first equality follows from the fact that $\varrho_m$ is a function of q and $Y_1,\ldots,Y_m$. The second equality is a consequence of part (i) and the equality $\varrho_{n+1} = T_{Y_{n+1}}(\varrho_n)$. The other equalities are obvious. □

In th. 5.3 we prove the optimality of a Bayesian equivalent rule in the independent case. Here the reward function r is constant in the first coordinate, i.e. at each stage the reward only depends on the chosen action and the value of the supplementary state variable. Further it is assumed that all actions are available in every state, i.e. $D(x) = A$, for $x \in X$. Since, given Z, the sequence $\{Y_n, n \in \mathbb{N}^*\}$ is a sequence of i.i.d. random variables (cf. lemma 2.2) we call this case the *independent case*. It will play an important role in section 5.3. We start with a lemma.

Lemma 5.2

Let G be an upper semi-continuous (u.s.c.) function on $A \times Y$, that is bounded from above. Let A be compact. Then there is a measurable function $f: W \to A$ such that:

$$\int G(f(q),y)p(y,q)\nu(dy) = \max_{a\in A} \int G(a,y)p(y,q)\nu(dy) \ .$$

Proof.

We show that all conditions of Schäl's selection theorem (cf. A17) are satisfied. Let G be u.s.c. and bounded above by $M \in \mathbb{R}$. Then there are bounded continuous functions $G_k$ on $A \times Y$, such that the sequence $\{G_k, k \in \mathbb{N}\}$ is non-

increasing and $\lim_{k\to\infty} G_k = G$ (see A12).

Without loss of generality we may assume that $G_k \leq M$, $k \in \mathbb{N}$, since otherwise we may define $\tilde{G}_k := \min\{G_k, M\}$ and then $\{\tilde{G}_k, k \in \mathbb{N}\}$ is also a nonincreasing sequence of bounded continuous functions with limit G. Hence, by the monotone convergence theorem we have

$$\lim_{k\to\infty} \int \{-G_k(a,y) + M\}p(y,q)\nu(dy) = \int \{-G(a,y) + M\}p(y,q)\nu(dy)$$

and so $\{\int G_k(a,y)p(y,q)\nu(dy),\ k \in \mathbb{N}\}$ is a nonincreasing sequence with limit $\int G(a,y)p(y,q)\nu(dy)$.

By the dominated convergence theorem the function

$$(a,q) \to \int G_k(a,y)p(y,q\ )\nu(dy)$$

is continuous in a, and for fixed k it is bounded since $G_k$ is bounded. Using lemma 1.6 (iii) we find that this function is measurable, since

$$(a,\theta) \to \int G_k(a,y)p(y|\theta)\nu(dy)$$

is measurable. Hence we proved that $\int G_k(a,y)p(y,q)\nu(dy) \in L(W \times A)$ and therefore $\int G(a,y)p(y,q)\nu(dy) \in \hat{L}(W \times A)$ (cf. A17).

Hence all conditions of A17 are satisfied. This proves the lemma. □

Theorem 5.3

Let I be a singleton, let A be compact and let $D(x) = A$, $x \in X$. Further let $x \to r(x,a,y)$ be constant for all $a \in A$, $y \in Y$ and let $(a,y) \to r(x,a,y)$ be u.s.c. (We write $r(a,y) := r(x,a,y)$, $a \in A$, $y \in Y$.)

Then there is a strategy $\pi^* \in \Pi_0$ that choses a maximizer of $a \to \int r(a,y)p(y,q)\nu(dy)$ in each state $(x,q) \in X \times W$.

This strategy is optimal, and

$$v(x,q) = \sum_{n=0}^{\infty} \beta^n \mathbf{E}_q[e(Q_n)]$$

where,

$$5.1 \qquad e(q) := \max_{a\in A} \int r(a,y)p(y,q)\nu(dy), \quad q \in W\ .$$

Proof.

Remember that r is bounded from above. Let $x \in X$, $q \in W$ and $\pi \in \Pi_0$. We have:

$$\mathbf{E}^{\pi}_{x,q}[r(A_n,Y_{n+1})] = \mathbf{E}^{\pi}_{x,q}[\mathbf{E}^{\pi}_{x,q}[r(A_n,Y_{n+1}) \mid X_0,Q_0,A_0,\ldots,X_n,Q_n]] \ .$$

Since $\pi \in \Pi_0$, there is a corresponding strategy $\tilde{\pi} \in \tilde{\Pi}$ (cf. 3.7). Therefore we have $\mathbf{P}^{\pi}_{x,q}$-a.s.:

$$\mathbf{E}^{\pi}_{x,q}[r(A_n,Y_{n+1}) \mid X_0,Q_0,A_0,\ldots,X_n,Q_n)] =$$

$$= \int \tilde{\pi}_n(da|X_0,Q_0,A_0,\ldots,X_n,Q_n) \int \nu(dy)p(y,Q_n)r(a,y) \leq$$

$$\leq \sup_{a\in A} \int \nu(dy)p(y,Q_n)r(a,y) \ .$$

By lemma 5.2 there is a measurable function f from W to Y such that

$$\max_{a\in A} \int \nu(dy)p(y,Q_n)r(a,y) = \int \nu(dy)p(y,Q_n)r(f(Q_n),y) = e(Q_n) \ .$$

Note that the distribution of $e(Q_n)$ does not depend on x and $\pi$. Hence

$$\mathbf{E}^{\pi}_{x,q}[r(A_n,Y_{n+1})] \leq \mathbf{E}_q[e(Q_n)] \quad \text{for all } x \in X,\ \pi \in \Pi_0 \ ,$$

with equality if the strategy $\pi^*$ is used. This proves the theorem. □

The strategy $\pi^*$ defined in th. 5.3 uses a Bayesian equivalent rule. To verify this, note that an optimal strategy for the model with known parameter $\theta$ is obtained by using a maximizer of $a \to F(x,\theta,a)$ at each stage, where

$$F(x,\theta,a) := \int \nu(dy)p(y|\theta)r(a,y) \ .$$

Hence a Bayesian equivalent rule may be defined as a maximizer of $a \to \int q(d\theta)F(x,\theta,a)$, in each state $(x,q) \in X \times W$. Hence $\pi^*$ uses a Bayesian equivalent rule at each stage. Note that each maximizer of $a \to \int q(d\theta)F(x,\theta,a)$ is also a maximizer of

$$a \to \int q(d\theta)p(y|\theta)\{r(a,y) + \beta \int P(dx'|x,a,y)v(x',\theta)\}$$

since $x \to v(x,\theta)$ is constant for all $\theta \in \Theta$. Hence $\pi^*$ uses a "natural" Bayesian equivalent rule.

In th. 5.4 we give an upper and a lower bound for the value function of the model. These provide a measure for the loss of return, due to the lack of information concerning the "true" parameter value.

Theorem 5.4

Under the conditions of th. 5.3 we have:

$$5.2 \qquad \frac{e(q)}{1-\beta} \leq v(x,q) \leq \frac{\int q(d\theta)e(\theta)}{1-\beta} .$$

Proof.

The right hand inequality follows from th. 3.16. To prove the left hand inequality, note that:

$$\mathbf{E}_q[e(Q_n)] \geq \sup_{a\in A} \mathbf{E}_q[\int Q_n(d\theta)\{\int \nu(dy)p(y|\theta)r(a,y)\}] =$$

$$= \sup_{a\in A} \int q(d\theta)\{\int \nu(dy)p(y|\theta)r(a,y)\} = e(q) .$$

The first equality follows from the fact that $\mathbf{E}_q[\int f(\theta)Q_n(d\theta)] = \int f(\theta)q(d\theta)$, for real-valued measurable functions on $\theta$, which are bounded from above. □

We conclude this section with an example which has some relationship with the inventory control model we study in section 5.3. The model we consider in this example can be transformed into the model we called the independent case.

Example 5.2

Let I be a singleton, $D(x) = A$ for all $x \in X$ and let A be compact. Further let $r(x,a,y) = b(x) + c(a)$, $x \in X$ and $a \in A$, where b and c are u.s.c. and bounded from above on X and A respectively. Finally let $P(\{G(a,y)\} \mid x,a,y) = 1$ for all $x \in X$, $a \in A$ and $y \in Y$ where G is a continuous function from $A \times Y$ to X.

For each $x \in X$, $q \in W$ and $\pi \in \Pi_0$ we have

$$v(x,q,\pi) = \mathbf{E}^{\pi}_{x,q}[\sum_{n=0}^{\infty} \beta^n r(X_n,A_n,Y_{n+1})] =$$

$$= \mathbf{E}^{\pi}_{x,q}[b(X_0) + \sum_{n=0}^{\infty} \beta^n\{c(A_n) + \beta b(G(A_n,Y_{n+1})\}] =$$

$$= b(x) + \sum_{n=0}^{\infty} \beta^n \mathbf{E}^{\pi}_{x,q}[c(A_n) + \beta b(G(A_n,Y_{n+1}))] .$$

Define $\bar{r}(a,y) := c(a) + \beta b(G(a,y))$, $a \in A$, $y \in Y$. Then

$$v(x,q,\pi) = b(x) + \mathbf{E}^{\pi}_{x,q}\Big[\sum_{n=0}^{\infty} \beta^n \bar{r}(A_n, Y_{n+1})\Big] .$$

Note that $(a,y) \to \bar{r}(a,y)$ is u.s.c. (cf. A15). Hence, by th. 5.3, we find that an optimal strategy is obtained by choosing in each state $(x,q) \in X \times W$ a maximizer of

$$a \to \{c(a) + \beta \int b(G(a,y))p(y,q)\nu(dy)\}$$

and

$$v(x,q) = b(x) + \sum_{n=0}^{\infty} \beta^n \mathbf{E}_q[\tilde{e}(Q_n)] ,$$

where

$$\tilde{e}(q) := \max_{a \in A}\{c(a) + \beta \int b(G(a,y))p(y,q)\nu(dy)\} .$$

As in th. 5.4 we have the inequalities:

$$b(x) + \tilde{e}(q)(1-\beta)^{-1} \le v(x,q) \le b(x) + \int \tilde{e}(\theta)q(d\theta)(1-\beta)^{-1}, \quad (x,q) \in X \times W.$$

In the following sections we study models which have some practical relevance. As this is more natural, in these models we shall minimize costs rather than maximize rewards.

Note that all results up to here carry over if we define

5.3 (i) $\quad c(x,a) := -r(x,a)$

(ii) $$(U_\tau b)(x,q) := \inf_{\pi \in \Pi_0} \mathbf{E}^{\pi}_{x,q}\Big[\sum_{n=0}^{\tau-1} \beta^n c(X_n, A_n) + \beta^\tau b(X_\tau, Q_\tau)\Big]$$

for real-valued measurable functions b, that are bounded from below.

(iii) $\quad v(x,q) = (U_\infty 0)(x,q)$ .

## 5.2 *Linear system with quadratic costs*

In this chapter we consider a linear system with quadratic costs and with a disturbance process of i.i.d. random variables with an incompletely known distribution. We show the optimality of a Bayesian equivalent rule. In fact this rule can also be considered as a so-called *certainty equivalent rule*. We generalize results of M. Aoki on this topic in several ways: first we allow other disturbance processes than normal processes, secondly we allow general prior distributions. Finally we allow the costs to be a quadratic

function of the control variable (cf. [Aoki (1967), page 94].
The concepts and techniques we use here, are familiar in the theory of linear systems (cf. [Kushner (1971), chapter 9] and [Bertsekas (1976)]). We remark that the greater part of this section appeared in [van Hee (1976)].
We shall use symbols that we used before. However, they loose their previous interpretation here. We start with the specifications of the model. We proceed with some preliminary results, and in th. 5.9 we obtain one of the main results of this section: an explicit expression for the optimal strategy and also for the value function.
In this section x' means the transpose of x, where x is a column vector or a matrix.

Model 4: *the linear system*

5.4 (i) $X := Y := \mathbb{R}^{n_1}$, $n_1 \in \mathbb{N}^*$,

(ii) $D(x) := A := \mathbb{R}^{n_2}$, $n_2 \in \mathbb{N}^*$ for all $x \in X$,

(iii) $c(x,a) := x'Rx + a'Sa$ where R is a nonnegative definite $n_1 \times n_1$-matrix and S a positive definite $n_2 \times n_2$-matrix,

(iv) $P(\{Cx + Ba + y\} \mid x,a,y) = 1$, $x \in X$, $a \in A$, $y \in Y$ where C is a $n_1 \times n_1$-matrix and B a $n_1 \times n_2$-matrix satisfying the *controllability assumption*

$$\text{rank}[B,CB,\ldots,C^{n_1-1}B] = n_1 ,$$

(v) $\int \nu(dy) \mid y_i y_j \mid p(y|\theta)$ is bounded on $\theta$ where $y_i$ is the i-th component of $y \in Y$, for all $i,j \in \{1,\ldots,n_1\}$.

For $q \in W$ we define the vector $m_q$ and the matrices $M_q$ and $\Sigma_q$:

5.5 (i) $m_q(i) := \int y_i p(y,q)\nu(dy)$, $i \in \{1,\ldots,n_1\}$ .

(ii) $M_q(i,j) := \int m_\theta(i)m_\theta(j)q(d\theta)$, $i,j \in \{1,\ldots,n_1\}$ .

(iii) $\Sigma_q(i,j) := \int y_i y_j p(y,q)\nu(dy)$, $i,j \in \{1,\ldots,n_1\}$.

Note that $\Sigma_q - M_q$ is the *covariance matrix* of $Y_n$ averaged over $\theta$ with q.
By assumption 5.4 (v) $m_q$, $M_q$ and $\Sigma_q$ are bounded on W.

In lemma 5.5 we give some properties of $m_q$ and $M_q$.

**Lemma 5.5**

**For $q \in W$ we have**

(i) $$\int m_{T_y(q)}(i)p(y,q)\nu(dy) = m_q(i), \quad i \in \{1,\ldots,n_1\} .$$

(ii) $$\int y_j m_{T_y(q)}(i)p(y,q)\nu(dy) = M_q(i,j), \quad i,j \in \{1,\ldots,n_1\} .$$

**Proof.**

$$m_{T_y(q)}(i)p(y,q) = p(y,q)\int \tilde{y}_i\{\int \frac{p(\tilde{y}|\theta)p(y|\theta)}{p(y,q)} q(d\theta)\}\nu(d\tilde{y}) =$$

$$= \int\{\int \tilde{y}_i p(\tilde{y}|\theta)p(y|\theta)\nu(d\tilde{y})\}q(d\theta) .$$

**Hence**

$$\int m_{T_y(q)}(i)p(y,q)\nu(dy) = \int\{\int \tilde{y}_i p(\tilde{y}|\theta)\nu(d\tilde{y})\}q(d\theta) = m_q(i)$$

**and**

$$\int y_j m_{T_y(q)}(i)p(y,q)\nu(dy) = \iiint y_j\tilde{y}_i p(\tilde{y}|\theta)p(y|\theta)\nu(d\tilde{y})\nu(dy)q(d\theta) =$$

$$= \int\{\int y_j p(y|\theta)\nu(dy)\}.\{\int \tilde{y}_i p(\tilde{y}|\theta)\nu(d\tilde{y})\}q(d\theta) = M_q(i,j) .$$

Note that all changes of integration order are allowed by 5.4 (v). □

Lemma 5.6 states that the optimal reward operator U (cf. 5.3(i)) maps the set of functions f on $X \times W$ of the form given in 5.6 below, into itself. The proof proceeds in a familiar way (cf. [Kushner (1971), section 9.2.2]).

**Lemma 5.6**

Let the real-valued function f on $X \times W$ be defined by:

5.6 $$f(x,q) := x'Kx + x'Lm_q + H(q), \quad x \in X, q \in W ,$$

where K is a nonnegative definite matrix, L an arbitrary $n_1 \times n_1$-matrix and H a bounded and measurable function on W. Then:

$$(Uf)(x,q) := x'\tilde{K}x + x'\tilde{L}m_q + \tilde{H}(q), \quad x \in X, q \in W ,$$

where

5.7 (i) $$\tilde{K} := G_1(K) := R + \beta C'KC - \beta^2 C'KB(S + \beta B'KB)^{-1}B'KC .$$

(ii) $$\tilde{L} := G_2(L,K) := 2\beta C'K + \beta C'L - \beta^2 C'KB(S + \beta B'KB)^{-1}(2B'K + B'L) .$$

(iii) $\tilde{H}(q) := G_3(q,H,K,L) := -\frac{1}{4}\beta^2 m_q'(2KB + L'B)(S + \beta B'KB)^{-1}(2B'K + B'L)m_q +$

$+ \beta \int H(T_y(q))p(y,q)\nu(dy) + \beta \operatorname{trace}(K\Sigma_q) + \beta \operatorname{trace}(LM_q)$

and the minimizing action $a(x,q)$ in $(x,q)$ is:

5.8 $$a(x,q) = -\beta(S + \beta B'KB)^{-1}B'KCx - \beta(S + \beta B'KB)^{-1}(B'K + \tfrac{1}{2}B'L)m_q .$$

Further: $\tilde{K}$ is nonnegative definite and $\tilde{H}(\cdot)$ is bounded and measurable on W.

Proof.

(1) By some straightforward calculations, using lemma 5.5 we get:

$$(Uf)(x,q) = \inf_{a \in A} \{a'(S + \beta B'KB)a + (2\beta x'C'KB + 2\beta m_q'KB + \beta m_q'L'B)a\} +$$

$$+ x'(R + \beta C'KC)x + \beta x'(2C'K + C'L)m_q + \beta \int H(T_y(q))p(y,q)\nu(dy) +$$

$$+ \beta \operatorname{trace}(K\Sigma_q) + \beta \operatorname{trace}(LM_q) .$$

Since K is nonnegative definite and S is positive definite we have $S + \beta B'KB$ is positive definite and therefore $(S + \beta B'KB)^{-1}$ exists and is positive definite. Hence by standard arguments for the minimization of quadratic forms we find 5.7 and 5.8.

(2) We shall prove that $\tilde{K}$ is nonnegative definite again. Note that the value of $\tilde{K}$ does not depend on L, H or $p(y|\theta)$, $y \in Y$, $\theta \in \Theta$. Hence, to prove this, we may assume that H vanishes and that $\int |y_i y_j| p(y|\theta)\nu(dy) = 0$, for all $i,j \in \{1,\ldots,n_1\}$ and $\theta \in \Theta$. In that case $(Uf)(x,q) = x'\tilde{K}x$, since $\Sigma_q$, $M_q$ and $m_q$ contain only zeros for all $q \in W$. By the definition of $(Uf)(x,q)$ we have

$$(Uf)(x,q) = \inf_{a \in A} \{x'Rx + a'Sa + \beta \int (Cx + Ba + y)'K(Cx + Ba + y)p(y,q)\nu(dy)\}$$

and therefore $(Uf)(x,q) \geq 0$ for all $(x,q) \in X \times W$ since R, S and K are nonnegative definite. Hence $x'Kx \geq 0$ for all $x \in X$. It is easy to verify that $\tilde{K}$ is symmetric. Hence $\tilde{K}$ is nonnegative definite.

(3) Finally we consider the function $q \to \tilde{H}(q)$. Using lemma 1.6 (iii) we have $q \to m_q(i)$, $q \to M_q(i,j)$ and $q \to \Sigma_q(i,j)$ are bounded and measurable. So all terms in 5.7 (iii), except the second one, are bounded and measurable on W. We consider the second term separately. To show that

$(y,q) \to T_y(q)$ is measurable, it suffices to prove that the set $\{(y,q) \mid T_y(q)(B) \le c\}$ is measurable for $B \in \mathcal{T}$ and $c \in \mathbb{R}$ (cf. lemma 1.5). Hence it suffices to show that $(y,q) \to T_y(q)(B)$ is measurable, $B \in \mathcal{T}$. Note that $(y,q) \to \int_B p(y|\theta)q(d\theta)$ and $(y,q) \to p(y,q)$ are measurable (cf. lemma 1.6 (iii)). Hence $(y,q) \to T_y(q)(B)$ is measurable since

$$T_y(q)(B) = \int_B p(y|\theta)q(d\theta)\{p(y,q)\}^{-1} \quad \text{if } p(y,q) > 0$$

$$= q(B) \quad \text{if } p(y,q) = 0 .$$

Therefore $(y,q) \to H(T_y(q))$ is also measurable. This proves the measurability of $(y,q) \to \int H(T_y(q))p(y,q)\nu(dy)$. □

The equation $G_1(K) = K$ is called the *Riccati-equation*.
Now we shall consider the sequence of successive approximations:

$$v_n(x,q) := (U^n 0)(x,q), \quad x \in X,\ q \in W .$$

We define $n_1 \times n_1$-matrices $K_n$ and $L_n$ and a sequence of bounded measurable functions $H_n$ on $W$, for $n \in \mathbb{N}$:

5.9 (i) $K_0(i,j) := L_0(i,j) := 0,\ i,j \in \{1,\dots,n_1\};\ H_0(q) := 0,\ q \in W .$

(ii) $K_n := G_1(K_{n-1})$ , $L_n := G_2(L_{n-1}, K_{n-1})$ ,

$H_n(q) := G_3(q,H_{n-1},K_{n-1},L_{n-1}),\ q \in W,\ n \in \mathbb{N}^*$

($G_1$, $G_2$ and $G_3$ are defined in 5.7).

It is a direct consequence of lemma 5.6 that

5.10 $$v_n(x,q) = x'K_n x + x'L_n m_q + H_n(q), \quad n \in \mathbb{N} .$$

In lemma 5.7 we prove that $K_n$ and $L_n$ converge elementwise to matrices $K^*$ and $L^*$ respectively. The proof of $K_n \to K^*$ can also be found in [Kushner (1971), section 9.2.3]. In our proof we use the same arguments. In lemma 5.8 we show the pointwise convergence of $H_n$ as n tends to infinity.

Lemma 5.7

(i) $K_n$ converges, elementwise, to a nonnegative definite matrix $K^*$ satisfying the Riccati-equation ($K^* = G_1(K^*)$).

(ii) $L_n$ converges, elementwise, to a matrix $L^*$ satisfying $L^* = G_2(L^*,K^*)$.

**Proof.**

Since $K_n$ and $L_n$ do not depend on the measure $\nu$, to study the limit behaviour we may assume that $\nu$ is concentrated in a point $m^* \in \mathbb{R}^{n_1}$. Hence $m_q = m^*$ for $q \in W$, and we are dealing with a deterministic system. Let us denote the value function of this system by v and the sequence of successive approximations by $\{v_n(x), n \in \mathbb{N}\}$ (note that we omit the dependence on $q \in W$). We first show that this value function v is finite. Let $x = x_0$ be the starting state. Note that

$$x_{n_1} = C^{n_1} x_0 + \sum_{k=0}^{n_1-1} C^k B a_{n_1-1-k} + \sum_{k=0}^{n_1-1} C^k m^* ,$$

hence

$$x_{n_1} - C^{n_1} x_0 - \sum_{k=0}^{n_1-1} C^k m^* = \sum_{k=0}^{n_1-1} C^k B a_{n_1-1-k} .$$

By the controllability assumption 5.4(iv) we can find actions $a_0,\ldots,a_{n_1-1}$ such that $x_{n_1} = 0$. So there is, for the deterministic linear system, a strategy $\pi$ such that $x_{kn_1} = 0$ for $k \in \mathbb{N}^*$, and such that each cycle from $x_{kn_1} = 0$ until $x_{(k+1)n_1} = 0$ passes through the same states and actions ($k \in \mathbb{N}^*$).
Hence the discounted total costs of the system under $\pi$ is finite. Since the one-step costs are nonnegative, $v_n(x)$ is nondecreasing in n. Note that, by a simple dynamic programming argument, $v_n(x) \leq v(x)$ for all $n \in \mathbb{N}$ and $x \in X$. Hence $v_n(x)$ converges if n tends to infinity. Note that, by 5.7(iii) and the special form of $\nu$,

$$v_n(x) = x'K_n x + x'L_n m^* + H_n$$

where $H_n$ is constant on W (therefore we omit the dependence on $q \in W$).
Since $v_n(0)$ converges, we find that $\lim_{n\to\infty} H_n$ exists and is finite. Let $m^* = 0$.
So we find that $\lim_{n\to\infty} x'K_n x$ exists for all $x \in X$, since $K_n$ does not depend on the value of $m^*$. It is straightforward to show that this implies that $K_n$ converges elementwise. Consequently, for arbitrary $m^*$, $x'L_n m^*$ converges for all $x \in X$. Hence $L_n$ converges elementwise. As $K_n$ is nonnegative definite we have $x'K_n x \geq 0$ for all $x \in X$, hence $x'K^* x \geq 0$ for all $x \in X$. Since $K_n = K_n'$ we have $K^* = K^{*'}$ and therefore $K^*$ is nonnegative definite.
Finally, since $(S + \beta B'K_n B)^{-1}$ converges elementwise to $(S + \beta B'K^* B)^{-1}$ we find $K^* = G_1(K^*)$ and $L^* = G_2(L^*, K^*)$. □

**Lemma 5.8**

The sequence of functions $H_n$, defined in 5.9, converges pointwise to a bounded measurable function $H^*$ (on W), such that $H^*(\cdot) = G_3(\cdot, H^*, K^*, L^*)$.

Proof.

Let

$$5.11 \quad (a) \quad b_n(q) := -\tfrac{1}{4}\beta^2 m_q'(2K_nB + L_n'B)(S + \beta B'K_nB)^{-1}(2B'K_n + B'L_n)m_q + \\ + \beta \operatorname{trace}(K_n\Sigma_q) + \beta \operatorname{trace}(L_nM_q) .$$

It follows from lemma 5.7 that $b_n(q)$ converges if n tends to infinity.
Denote:

$$5.11 \quad (b) \quad b(q) := \lim_{n\to\infty} b_n(q), \quad q \in W .$$

By definition 5.7 we have $H_{m+1}(q) = b_m(q) + \beta \int H_m(T_y(q))p(y,q)\nu(dy)$, $m \in \mathbb{N}$. Note that $\int H_m(T_y(q))p(y,q)\nu(dy) = \mathbb{E}_q[H_m(Q_1)]$. Therefore, by th. 5.1 and the measurability of $(y,q) \to H_m(T_y(q))$ (cf. the proof of lemma 5.6):

$$(*) \qquad H_{m+1}(Q_k) = b_m(Q_k) + \beta\mathbb{E}_q[H_m(Q_{k+1}) \mid Y_1,\ldots,Y_k], \quad k \in \mathbb{N}^*, \ m \in \mathbb{N} .$$

Hence we have

$$H_{n+1}(q) = b_n(q) + \beta\mathbb{E}_q[b_{n-1}(Q_1) + \beta\mathbb{E}_q[H_{n-1}(Q_2) \mid Y_1]] = \\ = b_n(q) + \beta\mathbb{E}_q[b_{n-1}(Q_1)] + \beta^2\mathbb{E}_q[H_{n-1}(Q_2)] .$$

And by iteration, using (*) we find:

$$(**) \qquad H_{n+1}(q) = \sum_{\ell=0}^{n} \beta^\ell \mathbb{E}_q[b_{n-\ell}(Q_\ell)] + \beta^{n+1}\mathbb{E}_q[H_0(Q_{n+1})] .$$

The last term vanishes since, by definition, $H_0 = 0$. Since $K_n$ and $L_n$ are bounded in n (elementwise; see the proof of lemma 5.7), and since $q \to m_q$, $q \to M_q$ and $q \to \Sigma_q$ are bounded functions (elementwise), we have (cf. 5.11) the boundedness of $(n,q) \to b_n(q)$. Hence, for all $\varepsilon > 0$ there is a $N_\varepsilon \in \mathbb{N}$ such that:

$$\sum_{\ell=N_\varepsilon}^{\infty} \beta^\ell \mathbb{E}_q[b_{n-\ell}(Q_\ell)] \le \varepsilon .$$

By the dominated convergence theorem, for fixed $\ell$ we now have:

$$\lim_{n\to\infty} \mathbf{E}_q[b_{n-\ell}(Q_\ell)] = \mathbf{E}_q[b(Q_\ell)] \ .$$

Hence, using (**) we find:

$$5.12 \qquad H^*(q) := \lim_{n\to\infty} H_n(q) = \sum_{k=0}^{\infty} \beta^k \mathbf{E}_q[b(Q_k)] \ .$$

Since b is bounded, $H^*$ is also bounded. The measurability of $H^*$ is immediate. Finally, note that 5.12 and the Markov property of the process $\{Q_n, n\in\mathbf{N}\}$ (cf. th. 5.1) imply:

$$H^*(q) = b(q) + \beta \sum_{k=0}^{\infty} \beta^k \mathbf{E}_q[b(Q_{k+1})] = b(q) + \beta \mathbf{E}_q[\sum_{k=0}^{\infty} \beta^k \mathbf{E}_q[b(Q_{k+1})|Q_1]] =$$

$$= b(q) + \beta \mathbf{E}_q[\sum_{k=0}^{\infty} \beta^k \mathbf{E}_{Q_1}[b(Q_k)]] = b(q) + \beta \mathbf{E}_q[h(Q_1)] \ .$$

Hence $H^*(q) = G_3(q,H^*,K^*,L^*)$. □

The next theorem is one of the main results of this section. It is an immediate consequence of the foregoing lemmas and a well-known argument for *negative dynamic programming* (cf. [Strauch (1966)]).

Theorem 5.9

(i) The value function satisfies

$$v(x,q) = x'K^*x + x'L^*m_q + H^*(q) \ .$$

(ii) In state (x,q) the optimal strategy chooses the action:

$$a(x,q) = -\beta(S+\beta B'K^*B)^{-1}B'K^*Cx - \beta(S+\beta B'K^*B)^{-1}(B'K^* + \tfrac{1}{2}B'L^*)m_q$$

(where $K^*$ and $L^*$ are defined in lemma 5.7 and $H^*(\cdot)$ in 5.12).

Proof.

From the lemmas 5.6, 5.7 and 5.8 it follows that

$$v_\infty(x,q) := \lim_{n\to\infty} v_n(x,q) = x'K^*x + x'L^*m_q + H^*(q), \ x \in X, \ q \in W$$

and also that, for $x \in X$, $q \in W$:

$$(*) \qquad v_\infty(x,q) = (Uv_\infty)(x,q) = c(x,a(x,q)) + \beta \mathbf{E}^{\pi^*}_{x,q}[v_\infty(X_1,Q_1)] \ ,$$

where $\pi^*$ is the strategy that chooses in $(x,q)$ action $a(x,q)$ (defined above). Since the process $\{(X_n,Q_n),\ n \in \mathbb{N}\}$ is a Markov chain under $\pi^*$ we find by iteration of (*):

$$(**) \qquad v_\infty(x,q) = \mathbb{E}^{\pi^*}_{x,q}\Big[\sum_{n=0}^{N-1} \beta^n c(X_n, a(X_n,Q_n))\Big] + \beta^N \mathbb{E}^{\pi^*}_{x,q}[v_\infty(X_N,Q_N)] \ .$$

Note that according to a simple dynamic programming argument $v_n(x,q) \le v(x,q)$ for all $n \in \mathbb{N}$, since $c$ is a nonnegative function. Hence

$$v(x,q) \ge v_\infty(x,q) \ge \mathbb{E}^{\pi^*}_{x,q}\Big[\sum_{n=0}^{\infty} \beta^n c(X_n, a(X_n,Q_n))\Big] \ge v(x,q) \ .$$

Here the second inequality follows from (**) since $v_\infty(x,q) \ge 0$, $(x,q) \in X \times W$. □

The following theorem provides a bound for the extra costs we incur due to lack of information about the parameter value $\theta \in \Theta$.

Theorem 5.10

$$0 \le v(x,q) - \int v(x,\theta)q(d\theta) \le \frac{1}{1-\beta}\{b(q) - \int b(\theta)q(d\theta)\}, \quad (x,q) \in X \times W$$

(where $b$ is defined in 5.11 (b)).

Proof.

By th. 3.16 we have $v(x,q) \ge \int v(x,\theta)q(d\theta)$. (Remember that we are minimizing here.) Note that, by th. 5.9:

$$v(x,\theta) = x'K^*x + x'L^*m_\theta + \sum_{n=0}^{\infty} \beta^n b(\theta)$$

since all posterior distributions are concentrated in $\theta$, if the prior distribution is concentrated in $\theta \in \Theta$. Since $\int q(d\theta)x'L^*m_\theta = x'L^*m_q$ we have

$$(*) \qquad v(x,q) - \int v(x,\theta)q(d\theta) = \sum_{n=0}^{\infty} \beta^n \{\mathbb{E}_q[b(Q_n)] - \int b(\theta)q(d\theta)\} \ .$$

Note that $b(q)$, satisfies 5.11 (a), with $K_n$ and $L_n$ replaced by $K^*$ and $L^*$, respectively. Note that the matrix $E$, defined by

$$E := (2K^*B + L^{*\prime}B)(S + \beta B'K^*B)^{-1}(2B'K^* + B'L^*) \ ,$$

is positive definite, since $(S+\beta B'K^{*}B)$ is. Therefore E can be written as $E = J'\Lambda J$, where J is an orthogonal matrix and $\Lambda$ is a diagonal matrix with positive entries $\lambda_1,\ldots,\lambda_{n_1}$ on the diagonal.

Hence

$$m_q'Em_q = \sum_{i=1}^{n_1} \lambda_i \{ \sum_{n=1}^{n_1} J_{ij} m_q(j) \}^2 .$$

And, by Schwarz's inequality, we find

$$\mathbf{E}_q[m_{Q_n}' Em_{Q_n}] \geq \sum_{i=1}^{n_1} \lambda_i \{\mathbf{E}_q[ \sum_{j=1}^{n_1} J_{ij} m_{Q_n}(j)]\}^2 .$$

Since $m_{Q_n}(j) = \int Q_n(d\theta) m_\theta(j)$ we have

$$\mathbf{E}_q[ \sum_{j=1}^{n_1} J_{ij} m_{Q_n}(j)] = \sum_{j=1}^{n_1} J_{ij} m_q(j) .$$

Hence we have

$$\mathbf{E}_q[m_{Q_n}' Em_{Q_n}] \geq \sum_{i=1}^{n_1} \lambda_i \{ \sum_{j=1}^{n_1} J_{ij} m_q(j) \}^2 = m_q'Em_q .$$

Note that

$$\text{trace}(L^{*}M_q) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} L^{*}(i,j) \int m_\theta(i) m_\theta(j) q(d\theta)$$

and

$$\text{trace}(K^{*}\Sigma_q) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_1} K^{*}(i,j) \int \{\int y_i y_j p(y|\theta)\nu(dy)\} q(d\theta) .$$

Hence we find:

$$\mathbf{E}_q[\text{trace}(L^{*}M_{Q_n})] = \text{trace}(L^{*}M_q) \text{ and } \mathbf{E}_q[\text{trace}(K^{*}\Sigma_{Q_n})] = \text{trace}(K^{*}\Sigma_q).$$

So we have

$$\mathbf{E}_q[b(Q_n)] \leq b(q) .$$

This proves the theorem. □

Remarks.

(i) The linear system with (known) transition law given by:

$$\tilde{P}(D|x,a) = \int_{\{Cx+Ba+y\in D\}} p(y,q)\nu(dy), \quad D \in X$$

and the same cost structure as in model 4 has the value function $\tilde{v}_q(\cdot)$ defined by:

$$\tilde{v}_q(x) := x'K^*x + x'L^*m_q + \frac{b(q)}{1-\beta} .$$

Hence, by th. 5.10, we have

$$\int v(x,\theta)q(d\theta) \leq v(x,q) \leq \tilde{v}_q(x) .$$

(ii) The optimal strategy we found in th. 5.9 is a Bayesian equivalent rule (cf. 4.4), since the action in state $(x,q) \in X \times W$ is the minimizer of the function (cf. the proof of lemma 5.6)

$$a \to \int q(d\theta)\{a'(S+\beta B'K^*B)a + (2\beta x'C'K^*B + 2\beta m_\theta'K^*B + \beta m_\theta'L^{*\prime}B)a\} .$$

(iii) Since $m_q(i) = \int m_\theta(i)q(d\theta)$, $m_q(i)$ is the Bayes estimate of $m_\theta(i)$ when q is the prior distribution. Hence the optimal strategy we found in th. 5.9 can also be formulated as: at each moment use the Bayes estimate for $m_\theta$ in the formula for the optimal action instead of $m_\theta$, which should be used if $\theta$ were known. In the linear system with known transition law, i.e. with $\theta = \{\theta\}$, it turns out that the optimal strategy is the same as if $\nu$ is concentrated at $m_\theta$. In that situation we are dealing with a deterministic system. This property, which we used in the proof of lemma 5.7, is called the *certainty equivalent principle* (cf. [Bertsekas (1976)]). We showed that in each state $(x,q)$ we may act as if $\nu$ is concentrated on $m_q$.

## 5.3 *A simple inventory control model*

In this section we consider an inventory control model which is closely related to the model described in example 5.2: the main difference is that $D(x) \neq A$ for all $x \in X$. Further we shall specify here the functions b, c and G of the example. The model we shall deal with is extensively studied by several authors: [Scarf (1959)], [Iglehart (1964)], [Rieder (1972)], [Zacks and Fennel (1973)] and [Waldmann (1976)]. Except for Zacks and Fennel all these authors prove structural properties of the optimal strategy under various conditions. Only Zacks and Fennel considered an easy-to-handle suboptimal strategy and they studied its behaviour using Monte Carlo methods. We also study a suboptimal strategy, namely a Bayesian equivalent rule, and we give bounds on the difference of its Bayesian total discounted return and the optimal value. Further we consider conditions under which this strategy is optimal. We start with a sketch of the model.

Model 5A: *inventory control model*

5.13 (i) $X := A := (-\infty,M]$, $M > 0$ is called the *capacity*.

(ii) $D(x) := [x,M]$, a is the *inventory* after ordering.

(iii) $P(\{a-y\} \mid x,a,y) = 1$, y represents the *demand*.

(iv) $Y := \{y \in \mathbb{R} \mid y \geq 0\}$.

(v) $c(x,a) := hx^{+} + px^{-} + k(a - x)$ where h is the *holding cost*, p the *penalty cost* for shortage and k the *production cost*; $h,p,k > 0$ and $k < \beta(k+p)$.

(It is easy to verify that, if $k \geq \beta(k+p)$ never ordering is optimal.)

We shall compare this model with model 5B.

Model 5B.

5.14 $D(x) := [0,M]$.

Further all specifications as in 5.13.

It is easy to verify that all assumptions of example 5.2 are fulfilled. The optimal action in (x,q) for model 5B is the minimizer in [0,M] of:

$$5.15 \qquad a \to [ka + \beta \int \{h(a-y)^{+} + p(a-y)^{-} - k(a-y)\}p(y,q)\nu(dy)\}] .$$

We shall determine the minimizer. The term between braces equals

$$5.16 \qquad \{k - \beta(p+k)\}a + m_q\beta(p+k) + \beta(h+p) \int_{[0,a]} (a-y)p(y,q)\nu(dy)$$

where

$$5.17 \qquad m_q := \int yp(y,q)\nu(dy) .$$

It is easy to verify that

$$f(a) := \int_{[0,a]} (a-y)p(y,q)\nu(dy) = \int_0^a du\{\int_{[0,u]} p(y,q)\nu(dy)\} .$$

Hence the function f is continuous and $f(\frac{1}{2}(a+b)) \leq \frac{1}{2}\{f(a) + f(b)\}$. Therefore f is convex. So 5.16 has a minimum in

$$5.18 \text{ (a)} \qquad \tilde{s}(q) := \inf\{a \in \mathbb{R} \mid \int_{[0,a]} p(y,q)\nu(dy) \geq \frac{p - \frac{1-\beta}{\beta}k}{p+h}\} .$$

According to 5.13 (v) there always is minimum of 5.16. Note that $q \to \tilde{s}(q)$, $q \in W$ is measurable.
Consequently

5.18 (b) $s(q) := \min\{M, \tilde{s}(q)\}$, $q \in W$

is a minimizer of 5.16 in the set $[0,M]$.
It is well known that for the inventory control model 5A with known parameter value $\theta$, the strategy:
*"order upto level* $s(\theta)$ *each period or wait if the inventory is larger than* $s(\theta)$"
is optimal (cf. [Iglehart (1964)]).
To define a Bayesian equivalent rule for model 5A, define the function F by:

$$F(x,\theta,a) := ka + \beta \int \{h(a-y)^+ + p(a-y)^- - k(a-y)\}p(y|\theta)\nu(dy) \ .$$

Hence, in state $(x,q) \in X \times W$ this Bayesian equivalent rule chooses a minimizer of $a \to \int q(d\theta)F(x,\theta,a)$ on the set $[x,M]$:

5.18 (c) the Bayesian equivalent rule chooses the action $a(x,q)$,
$a(x,q) = \max\{x, s(q)\}$ in state $(x,q) \in X \times W$.

We proceed with a definition:

5.19 Let v be the value function of model 5A, w the value function of model 5B and let $\hat{v}$ be the Bayesian discounted total costs under the Bayesian equivalent rule defined in 5.18 (c).

Note that the Bayesian equivalent rule 5.18 (c) defines a Bayesian stationary strategy for model 5A and also for model 5B. Note also that the Bayesian discounted total costs for both models is the same under this strategy, namely $\hat{v}(x,q)$, if $(x,q) \in X \times W$ is the starting state.
There is an optimal strategy for model 5A of the form *"choose the action* $\max\{x,t(q)\}$ *in state* $(x,q)$", where $t : W \to A$ is a measurable function such that:

5.20 $t(q)$ is a minimizer of

$$a \to ka + \beta \int v(a-y, T_y(q))p(y,q)\nu(dy)\}$$

on the set $[0,M]$.

This is proved in [Rieder (1972), th. 7.2 and th. 7.3] under the additional assumption that $\theta \to m_\theta$ is bounded. For practical purposes this result is only interesting if the value function v is known. Lemma 5.11 shows that

$t(q) \leq s(q)$ for $q \in W$, which is intuitively clear, since if it is not allowed to reduce the inventory, you will order more carefully.

Lemma 5.11

Let $t: W \to A$ be measurable such that $t(q)$ satisfies 5.20 for $q \in W$. Then $s(q) \geq t(q)$, for $q \in W$.

Proof.

Define

$$f(x,q) := v(x,q) - \{hx^+ + px^- - kx\}, \quad (x,q) \in X \times W .$$

From the optimality equation, $v = Uv$, for model 5A it follows that

$$(*) \qquad f(x,q) = \inf_{x \leq a \leq M} \{ka + \beta \int v(a-y, T_y(q)) p(y,q) \nu(dy)\} .$$

Hence $x \to f(x,q)$ is nondecreasing for all $q \in W$. By (*) we have:

$$(**) \qquad f(x,q) = \inf_{x \leq a \leq M} [ka + \beta \int \{h(a-y)^+ + p(a-y)^- - k(a-y)\} p(y,q) \nu(dy) + \beta \int f(a-y, T_y(q)) p(y,q) \nu(dy)] .$$

Remember that the function (cf. 5.15)

$$a \to ka + \beta \int \{h(a-y)^+ + p(a-y)^- - k(a-y)\} p(y,q) \nu(dy)$$

is convex and attains a minimum in $[0,M]$ (cf. 5.18 (a)).
The last term of (**) is a nondecreasing function of a, since $x \to f(x,q)$ is nondecreasing, for all $q \in W$. Hence a minimizer $t(q)$ of

$$a \to ka + \beta \int v(a-y, T_y(q)) p(y,q) \nu(dy)$$

in the set $a \in [0,M]$ must satisfy $t(q) \leq s(q)$, $q \in W$. □

In th. 5.12 we give bounds for the difference $\hat{v} - w$ (cf. 5.19). This difference is an upper bound for $\hat{v} - v$, the loss due to controlling the system with the Bayesian equivalent rule. The bounds are derived by comparing two strategies for model 5B. We compare the optimal strategy for this model, where $\tilde{A}_n = s(Q_n)$ for all $n \in \mathbb{N}$, with the strategy where $A_n = \max\{x, s(Q_n)\}$, the Bayesian equivalent rule defined in 5.18 (c). Note that the production costs at time n for these two strategies, differ if $s(Q_{n-1}) - Y_n - s(Q_n) > 0$.

Theorem 5.12

The functions v, w and $\hat{v}$ defined in 5.19, satisfy the following inequalities:

(i) $$w(x,q) \le v(x,q) \le \hat{v}(x,q), \quad (x,q) \in X \times W \ .$$

(ii) $$\hat{v}(x,q) - w(x,q) \le (\frac{\beta}{1-\beta} h + k)\{(x - s(q))^+ + \sum_{n=1}^{\infty} \beta^n E_q[\{s(Q_{n-1}) - Y_n - s(Q_n)\}^+]\} \ .$$

Proof.

(i) Note that for model 5B the lower bound for the action space is not essential, since $\tilde{s}(q) > 0$ (see 5.18 (a)). Hence it is obvious that $w(x,q) \le v(x,q)$. The right-hand side of the inequality is trivial.

(ii) Let $X_n$ denote the inventory at time n when the action at time n $A_n := \max\{s(Q_n), X_n\}$ is used, and $\tilde{X}_n$ the inventory if $\tilde{A}_n := s(Q_n)$ is used. Further let $X_0 = \tilde{X}_0 = x$. Since $X_{n+1} = A_n - Y_{n+1}$ and $\tilde{X}_{n+1} = \tilde{A}_n - Y_{n+1}$ it is easily verified that: $\tilde{X}_n \le X_n$, $n \in \mathbb{N}$. Let $S_n := s(Q_n)$, $n \in \mathbb{N}$, and consider the difference in immediate costs at time n:

(*) $$h(X_n^+ - \tilde{X}_n^+) + p(X_n^- - \tilde{X}_n^-) - k(X_n - \tilde{X}_n) + k(\max\{X_n, S_n\} - S_n) \le$$
$$\le h(X_n - \tilde{X}_n) + k(\max\{X_n - S_n, 0\} - X_n + \tilde{X}_n), \ n \in \mathbb{N} \ ,$$

since $X_n^- \le \tilde{X}_n^-$ for $n \in \mathbb{N}$. We consider the term with coefficient k first. We establish:

(**) $$k(\max\{X_n - S_n, 0\} - X_n + \tilde{X}_n) \le k(S_{n-1} - S_n - Y_n)^+, \ n \in \mathbb{N}^* \ .$$

To prove (**), let $X_n > S_n$. Then (**) holds, since $\tilde{X}_n = S_{n-1} - Y_n$. And if $X_n \le S_n$, we get $\max\{X_n - S_n, 0\} - X_n + \tilde{X}_n \le 0$, and so (**) holds. For n = 0 we have

$$k(\max\{X_0 - S_0, 0\} - X_0 + \tilde{X}_0) = k(x - s(q))^+ \ .$$

Hence if h = 0:

$$\hat{v}(x,q) - w(x,q) \le k\{(x - s(q))^+ + \sum_{n=1}^{\infty} \beta^n E_q[(S_{n-1} - S_n - Y_n)^+]\} \ .$$

For $h > 0$ we consider $h(X_n - \tilde{X}_n)$, $n \in \mathbb{N}$.

Note that $X_0 = \tilde{X}_0 = x$ and

$$X_1 - \tilde{X}_1 = \max\{x, s(q)\} - Y_1 - s(q) + Y_1 = \{x - s(q)\}^+ .$$

We shall prove

$$(***) \qquad X_n - \tilde{X}_n \le \{x - s(q)\}^+ + \sum_{k=1}^{n-1} \{S_{k-1} - Y_k - S_k\}^+, \quad n \in \mathbb{N}^*$$

(an empty sum vanishes).

We already verified (***) for $n = 1$. Assume it holds for $n$. Consider:

$$X_{n+1} - \tilde{X}_{n+1} = \max\{X_n, S_n\} - Y_{n+1} - S_n + Y_{n+1} = (X_n - S_n)^+ .$$

By the induction hypothesis:

$$X_n \le \tilde{X}_n + \{x - s(q)\}^+ + \sum_{k=1}^{n-1} \{S_{k-1} - Y_k - S_k\}^+ =$$

$$= \{x - s(q)\}^+ + \sum_{k=1}^{n-1} \{S_{k-1} - Y_k - S_k\}^+ + S_{n-1} - Y_n .$$

Hence

$$(X_n - S_n)^+ \le \{x - s(q)\}^+ + \sum_{k=1}^{n} \{S_{k-1} - Y_k - S_k\}^+ ,$$

which proves (***).

Now we add the upperbounds for the differences in holding costs:

$$h \sum_{n=1}^{\infty} \beta^n (X_n - \tilde{X}_n) \le h \frac{\beta}{1-\beta} \{x - s(q)\}^+ + h \sum_{n=1}^{\infty} \beta^n \sum_{k=1}^{n-1} \{S_{k-1} - Y_k - S_k\}^+ =$$

$$= \frac{h\beta}{1-\beta} \{x - s(q)\}^+ + h \sum_{k=1}^{\infty} \{S_{k-1} - Y_k - S_k\}^+ \sum_{n=k+1}^{\infty} \beta^n =$$

$$= \frac{h\beta}{1-\beta} \{x - s(q)\} + \frac{h\beta}{1-\beta} \sum_{k=1}^{\infty} \beta^k \{S_{k-1} - Y_k - S_k\}^+$$

which accounts for the term with h in the right-hand side of (ii). □

Corollary 5.13

If for all $q \in W$:

5.21 $$\int_{\{y \mid s(q)-y \le s(T_y(q))\}} p(y,q)\nu(dy) = 1$$

then, for $x \le s(q)$, we have $v(x,q) = w(x,q)$ and therefore the Bayesian equivalent rule (defined in 5.18 (c)) is optimal.

Example 5.3

We define (cf. 5.18 (a)):

$$s_{min} := \inf_{\theta \in \Theta} \tilde{s}(\theta), \quad s_{max} := \sup_{\theta \in \Theta} \tilde{s}(\theta) \text{ and } \delta := s_{max} - s_{min} .$$

Since

$$\int_{[0,s_{max}]} p(y \mid \theta)\nu(dy) \ge \frac{p - \frac{1-\beta}{\beta}k}{p+h} \ge \int_{[0,s_{min})} p(y \mid \theta)\nu(dy)$$

for all $\theta \in \Theta$, we have

$$s_{min} \le \tilde{s}(q) \le s_{max} \quad \text{for all } q \in W .$$

Note that $s(q) = \min\{\tilde{s}(q), M\}$ for $q \in W$. Further note that

5.22 $$\{\min(\tilde{s}(Q_{n-1}), M) - \min(\tilde{s}(Q_n), M) - Y_n\}^+ \le \{\tilde{s}(Q_{n-1}) - \tilde{s}(Q_n) - Y_n\}^+ .$$

Hence

$$\mathbf{E}_q[\{s(Q_{n-1}) - s(Q_n) - Y_n\}^+] \le \mathbf{E}_q[\int_{[0,\delta]} p(y,Q_{n-1})\nu(dy)] =$$

$$= \int_{[0,\delta]} p(y,q)\nu(dy) .$$

Therefore we have by th. 5.12

$$\hat{v}(x,q) - w(x,q) \le \{\frac{\beta}{1-\beta}h + k\}\{(x - s(q))^+ + \frac{\beta}{1-\beta}\int_{[0,\delta]} p(y,q)\nu(dy)\} .$$

Hence, if $x \le s(q)$ and $\int_{[0,\delta]} p(y \mid \theta)\nu(dy) = 0$ for all $\theta \in \Theta$, then the Bayesian equivalent rule is optimal.

Remarks.

(i) The statement of corollary 5.13 is not new. In [Veinott (1965), section 6] a similar condition is considered for a multi-product inventory model with dependent demand, to prove an analogous statement. In [Rieder (1972)] Veinott's result has been translated to the Bayesian inventory model. However, the inequality of th. 5.12 (ii) seems to be new.

(ii) For the problem with known parameter, i.e. when $q \in W$ is degenerate at $\theta \in \Theta$, all posterior distributions are degenerate at $\theta$ and therefore 5.21 holds. So we actually proved the optimality of the rule: "order up to level $s(\theta)$ at each stage" for this situation.

(iii) Condition for lemma 5.13 can be weakened by requiring 5.21 only for all possible posterior distributions of a given $q \in W$.

(iv) In [van Hee (1976)] a different proof of th. 5.12 is given.

We conclude this section with an extensive study of the behaviour of the Bayesian equivalent rule (5.18 (c)) for the inventory model with exponentially distributed demand, where we assume the parameter of the demand distribution to have a gamma prior distribution. We shall compute the bound given in th. 5.12 (ii).
We also consider an upper bound for the relative error if we use the Bayesian equivalent rule (5.18 (c)) instead of an optimal rule. This *relative error* is defined by:

5.23 $$\{\hat{v}(0,q) - v(0,q)\}/v(0,q) \ .$$

Remember that model 5B satisfies all assumptions of example 5.2, and note that we are minimizing now.
Hence (cf. example 5.2) we have

$$w(0,q) \geq \frac{1}{1-\beta} \int q(d\theta)\tilde{e}(\theta)$$

where

5.24 $$\tilde{e}(\theta) := \min_{a \in A} [ka + \beta \int \{h(a-y)^+ + p(a-y)^- - k(a-y)\}p(y|\theta)\nu(dy)].$$

Note that $v(0,q) \geq w(0,q)$. Therefore we have the following upper bound for 5.23 (cf. th. 5.12 (ii)):

5.25 $$\Delta(\beta,k,h,p,q) := \frac{(\frac{\beta}{1-\beta}h+k)\sum_{n=1}^{\infty}\beta^n \mathbb{E}_q[\{s(Q_{n-1}) - Y_n - s(Q_n)\}^+]}{\frac{1}{1-\beta}\int q(d\theta)\tilde{e}(\theta)} .$$

We first give, in lemma 5.14, conditions guaranteeing that $\lim_{n\to\infty}\tilde{s}(Q_n) = \tilde{s}(Z)$ $\mathbb{P}_q$-a.s. Under these conditions we have, by 5.22

5.26 $$\lim_{\beta\uparrow 1}\Delta(\beta,k,0,p,q) = 0$$

since

$$\lim_{\beta\uparrow 1}(1-\beta)\sum_{n=1}^{\infty}\beta^n \mathbb{E}_q[\{\tilde{s}(Q_{n-1}) - Y_n - \tilde{s}(Q_n)\}^+] =$$

$$= \lim_{n\to\infty}\mathbb{E}_q[\{\tilde{s}(Q_{n-1}) - Y_n - \tilde{s}(Q_n)\}^+] = 0 .$$

Hence the relative error (cf. 5.23) tends to zero in this case.

Lemma 5.14

Let for all $\theta \in \Theta$, the function $a \to \int_{[0,a]} p(y|\theta)\nu(dy)$ be continuous and (strictly) increasing in a neighbourhood $(\tilde{s}(\theta)-\delta, \tilde{s}(\theta)+\delta)$ for $\delta > 0$ (cf. 5.18 (a). Then

$$\lim_{n\to\infty}\tilde{s}(Q_n) = \tilde{s}(Z) \ , \ \mathbb{P}_q\text{-a.s.} \quad \text{for all } q \in W .$$

Proof.

Define for $q \in W$ and $a \in \mathbb{R}$, $F_q(a) := \int_{[0,a]} p(y,q)\nu(dy)$. Note that $F_q(a) = \int F_\theta(a)q(d\theta)$. Since $a \to F_\theta(a)$ is continuous for all $\theta \in \Theta$ the function $a \to F_q(a)$ is continuous for all $q \in W$.
According to th. 2.4 we have for each function $\theta \to F_\theta(a)$ a set $\Omega_a \in \mathcal{H}$ such that $\mathbb{P}_q[\Omega_a] = 1$ and

(*) $$\lim_{n\to\infty} F_{Q_n}(a) = F_Z(a) \text{ on } \Omega_a .$$

Let R be the set of rational numbers in $\mathbb{R}$. Define $\Omega^* := \bigcap_{a\in R}\Omega_a$. Note that $\mathbb{P}_q[\Omega^*] = 1$. Let $\ell := (p - \frac{1-\beta}{\beta}h)(p+h)^{-1}$. Then $F_Z(\tilde{s}(Z)) = \ell$.
Fix $\omega \in \Omega^*$.

Fix $a_1, a_2 \in R$ such that $a_1 < \tilde{s}(Z(\omega)) < a_2$ and $a_2 - a_1 < \delta$. Then

$$F_{Z(\omega)}(a_1) < \ell < F_{Z(\omega)}(a_2) \ .$$

Hence

$$\lim_{n\to\infty} F_{Q_n(\omega)}(a_1) \le \ell \le \lim_{n\to\infty} F_{Q_n(\omega)}(a_2) \ .$$

Therefore we have, for n sufficiently large

$$a_1 \le \tilde{s}(Q_n(\omega)) \le a_2 \ .$$

Hence

$$\lim_{n\to\infty} \tilde{s}(Q_n(\omega)) = \tilde{s}(Z(\omega)) \ . \qquad \square$$

Example 5.4 *Exponential demand, gamma prior distribution*

In this example we consider the inventory control model (model 5A), with an exponential demand distribution and a gamma distribution for the unknown parameter of the demand distribution.
Let $p(y|\theta) := \theta e^{-\theta y}$, $Y := \Theta := [0,\infty)$ and let $q = \Gamma(\lambda, N)$, where

$$\Gamma(\lambda,N)(B) := \int_B \frac{\theta^{N-1}}{(N-1)!} \lambda^N e^{-\lambda\theta} d\theta \ ,$$

for a Borel subset B of $\mathbb{R}$. It is easy to verify that

$$p(y,q) = \frac{N\lambda^N}{(\lambda+y)^{N+1}} \quad \text{and} \quad \int_0^a p(y,q)dy = 1 - \left(\frac{\lambda}{\lambda+a}\right)^N \ .$$

Let

5.27 $$c := (p+h)\{h + (1-\beta)\beta^{-1}k\}^{-1} \ .$$

Then the minimizer $a^*$ of 5.15 is $a^* = \lambda(c^{1/N} - 1)$. So we have here

5.28 $$\tilde{s}(q) = \lambda(c^{1/N} - 1) \quad \text{for } q = \Gamma(\lambda,N) \ .$$

Further we consider the posterior distribution $T_y(q)$. It is straightforward to verify that

5.29 $$T_y(q) = \Gamma(\lambda+y, N+1) \quad \text{if } q = \Gamma(\lambda,N) \ .$$

Therefore the posterior distribution after n observations is

5.30 $$Q_n = \Gamma(\lambda + \sum_{i=1}^n Y_i, N+n) \ .$$

Hence, using 5.28 we find:

$$5.31 \qquad \tilde{s}(Q_n) - \tilde{s}(Q_{n+1}) - Y_{n+1} = \{\lambda + \sum_{i=1}^{n} Y_i\}\{c^{\frac{1}{N+n}} - c^{\frac{1}{N+n+1}}\} - Y_{n+1} c^{\frac{1}{N+n+1}} .$$

For a fixed $\theta \in \Theta$ we compute, for positive constants a and b:

$$5.32 \qquad \mathbf{E}_\theta[\{a \sum_{i=1}^{n} Y_i + b - Y_{n+1}\}^+] = \frac{an}{\theta} + b - \frac{1}{\theta} + \frac{e^{-\theta b}}{\theta (a+1)^n}$$

(remember here that, given $\theta$, $Y_i$ is exponentially distributed and $\sum_{i=1}^{n} Y_i$ is $\Gamma(\theta,n)$-distributed).

Now we integrate both terms in 5.32 over $\theta$, with respect to the $\Gamma(\lambda,N)$-distribution:

$$5.33 \qquad \mathbf{E}_q[\{a \sum_{i=1}^{n} Y_i + b - Y_{n+1}\}^+] = an \frac{\lambda}{N-1} + b - \frac{\lambda}{N-1} + \frac{1}{(a+1)^n} \frac{1}{N-1} \frac{\lambda^N}{(\lambda+b)^{N-1}} .$$

Finally we substitute for a and b the appropriate values (cf. 5.31). Hence

$$5.34 \qquad \mathbf{E}_q[\{\tilde{s}(Q_n) - \tilde{s}(Q_{n+1}) - Y_{n+1}\}^+] =$$

$$= \frac{\lambda}{N-1}\{(n+N-1)c^{\frac{1}{N+n}} - (n+N)c^{\frac{1}{N+n+1}} + c^{\frac{2(n+N)-1}{(n+N)(n+N+1)}}\} .$$

According to 5.22 we have

$$5.35 \qquad \mathbf{E}_q[\{s(Q_n) - s(Q_{n+1}) - Y_{n+1}\}^+] \leq \mathbf{E}_q[\{\tilde{s}(Q_n) - \tilde{s}(Q_{n+1}) - Y_{n+1}\}^+]$$

with equality if $M = \infty$ (M is the capacity cf. 5.13).

Note that the minimum in 5.24 is nonincreasing if M tends to infinity. Hence $\Delta(\beta,k,h,p,q)$ is nondecreasing if M tends to infinity. Therefore we shall assume $M = \infty$.

It is easy to verify that:

$$5.36 \qquad \tilde{e}(\theta) = \{k(1-\beta) - \beta p\}\frac{\log c}{\theta} + \frac{\beta(p+k)}{\theta} + \frac{\beta(p+h)}{\theta}\{\frac{1}{c} + \log c - 1\} .$$

Integration with respect to $q = \Gamma(\lambda,N)$ yields:

$$5.37 \qquad \int \tilde{e}(\theta) q(d\theta) = \frac{\lambda}{N-1}\{k(1-\beta)\log c + \frac{\beta(p+h)}{c} + \beta h(\log c - 1) + \beta k\} .$$

Finally $\Delta(\beta,k,h,p,q)$ is determined by 5.34 and 5.36.

In the table below we display $\Delta(\beta,k,h,p,\Gamma(1,N))$ for various parameter values. We also display the upperbound of th. 5.12 (ii):

5.38 $$B(\beta,k,h,p,q) := \{\frac{\beta}{1-\beta}\, h + k\} \sum_{n=1}^{\infty} \beta^n E_q[\{s(Q_{n-1}) - Y_n - s(Q_n)\}^+], \qquad q = \Gamma(1,N)\ .$$

Remember that $M = \infty$ and $\lambda = 1$ in the table.

| $\beta$ | k | h | p | N | c | $B(\beta,k,h,p,\Gamma(1,N))$ | $\Delta(\beta,k,h,p,\Gamma(1,N))$ in % |
|---|---|---|---|---|---|---|---|
| 0.90 | 10 | 0 | 2 | 5 | 1.8 | 2.22 | 8 |
| 0.90 | 10 | 0 | 2 | 15 | 1.8 | 0.33 | 4 |
| 0.90 | 10 | 0 | 10 | 5 | 9 | 10.39 | 34 |
| 0.90 | 10 | 2 | 10 | 5 | 3.8 | 15.89 | 46 |
| 0.90 | 10 | 2 | 10 | 15 | 3.8 | 2.21 | 22 |
| 0.95 | 10 | 0 | 1 | 5 | 1.9 | 3.79 | 7 |
| 0.95 | 10 | 2 | 1 | 5 | 1.2 | 4.61 | 9 |
| 0.95 | 10 | 2 | 5 | 15 | 2.8 | 4.76 | 26 |
| 0.95 | 100 | 2 | 10 | 5 | 1.6 | 40.29 | 8 |
| 0.95 | 100 | 2 | 10 | 15 | 1.6 | 6.56 | 4 |
| 0.99 | 10 | 0 | 1 | 5 | 9.9 | 31.85 | 12 |
| 0.99 | 10 | 0 | 10 | 15 | 99.0 | 12.54 | 16 |
| 0.99 | 10 | 2 | 5 | 15 | 3.3 | 60.28 | 67 |
| 0.99 | 100 | 1 | 2 | 5 | 1.4 | 95.22 | 4 |
| 0.99 | 100 | 1 | 5 | 5 | 2.9 | 274.41 | 10 |
| 0.999 | 10 | 0 | 1 | 5 | 99.9 | 130.83 | 5 |
| 0.999 | 10 | 0 | 2 | 5 | 199.8 | 157.38 | 6 |
| 0.999 | 10 | 0 | 5 | 5 | 499.5 | 196.55 | 7 |

## 6. *Approximations*

In this chapter we give several approximations to the value function of the Bayesian control model (cf. 2.12). Special attention is paid to the situation where there is only one unknown parameter, i.e. where $\theta_i$ is a singleton for all indices $i \in I$ except one. In section 6.1 we consider upper and lower bounds on the value function and their use in successive approximations. These approximations are computable when X, Y, A and $\theta$ are finite. In chapter 7 we shall consider algorithms based on these approximations. In section 6.1 we also give a lower bound on the Bayesian discounted total return when a certain Bayesian equivalent rule is used and we also consider another easy-to-handle Bayesian Markov policy. Since in practice the set $\theta$ is seldom finite, we study the consequences of approximating of $\theta$ by a finite subset in section 6.2. Throughout this chapter we assume that r is bounded and I finite.

### 6.1 *Bounds on the value function and successive approximations*

The bounds we consider require the knowledge of the value function of the dynamic program with known parameter value $\theta$ for all $\theta \in \theta$ and of the expected discounted total return under several stationary strategies, also for all $\theta \in \theta$. First we introduce some notations:

6.1 (i) $\tilde{F}$ is the set of Bayesian Markov policies (cf. 3.10).

(ii) F is the set of Markov policies (cf. 3.9 and note that $F \subset \tilde{F}$).

We identify each Bayesian Markov policy with the Bayesian stationary strategy which is determined by it (hence we write $v(x,q,f)$, $f \in \tilde{F}$). An important role is played by a subset $\bar{F}$ of F satisfying:

$$6.2 \qquad \inf_{f\in\bar{F}} \sup_{x\in X} \{v(x,\theta) - v(x,\theta,f)\} = 0 \text{ for all } \theta \in \theta .$$

We shall assume that such a set $\bar{F}$ is given and that $v(x,\theta,f)$ is known for all $x \in X$, $\theta \in \theta$ and $f \in \bar{F}$. Note that, if there exists for all $\theta \in \theta$ a $f_\theta \in F$ that is optimal for all $x \in X$ for the dynamic program with known parameter value $\theta$, then the set $\{f_\theta, \theta \in \theta\}$ satisfies 6.2.
For each $f \in \tilde{F}$ we define the (non-linear) operator $L_f$ on the set of bounded measurable functions b on $X \times W$ as follows:

6.3 (i) $$(L_f b)(x,q) := \sum_{i\in I} 1_{K_i}(x,f(x,q)) \int \nu(dy)p_i(y,q)\{r(x,f(x,q),y) + \beta \int P(dx'|x,f(x,q),y)b(x',T_{i,y}(q))\} .$$

(ii) $L_f^n$ is the n-th iterate of $L_f$, $n \in \mathbb{N}^*$ and $L_f^0 b := b$ .

Note that $\sup_{f\in\tilde{F}} L_f b = Ub$ for each bounded measurable function b on $X \times W$ (cf. the remark following 3.10). We further note that, for $f \in \tilde{F}$

$$\lim_{n\to\infty} (L_f^n b)(x,q) = v(x,q,f) .$$

Although this is easily proved directly, it is an immediate consequence of th. 3.14 (iii) if we consider the model with $D(x,q) = \{f(x,q)\}$, $x,q \in X \times W$. For $f \in F$ and $\theta \in \Theta$ 6.3 (i) reduces to

$$(L_f b)(x,\theta) = \sum_{i\in I} 1_{K_i}(x,f(x)) \int \nu(dy)p_i(y|\theta_i)\{r(x,f(x),y) + \beta \int P(dx'|x,f(x),y)b(x',\theta)\}$$

which is the usual return operator for the discounted dynamic program with known transition law (cf. [Blackwell (1965)]).

On $X \times W$ we define two functions:

6.4 (i) $$w(x,q) := \int q(d\theta)v(x,\theta) .$$

(ii) $$\ell(x,q) := \sup_{f\in\bar{F}} \int q(d\theta)v(x,\theta,f) .$$

Note that $\ell$ depends on the choice of the subset $\bar{F}$ of F. Further we define for $n \in \mathbb{N}^*$, $\theta \in \Theta$ and $f \in F$:

6.5 $$\varphi_n(\theta,f) := \sup_x \{v(x,\theta) - (L_f^n v)(x,\theta)\}$$

$$\varphi_\infty(\theta,f) := \sup_x \{v(x,\theta) - v(x,\theta,f)\} .$$

Note that, if X is finite, $\lim_{n\to\infty} \varphi_n(\theta,f) = \varphi_\infty(\theta,f)$ since

$$\lim_{n\to\infty} (L_f^n v)(x,\theta) = v(x,\theta,f) .$$

Theorem 6.1

For $x \in X$, $q \in W$, $\pi \in \Pi$ and $n \in \bar{\mathbb{N}}^*$ we have:

(i) $\quad \ell(x,q) \leq v(x,q) \leq w(x,q)$.

(ii) $\quad w(x,q) - \ell(x,q) \leq \frac{1}{1-\beta^N} \inf_{f \in \bar{F}} \int q(d\theta)\varphi_N(\theta,f)$.

(iii) $\quad \mathbb{E}^{\pi}_{x,q}[\frac{1}{1-\beta^N} \inf_{f \in \bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f)]$

is nonincreasing in n and if

$$\mathbb{P}^{\pi}_{x,q}[\bigcap_{i \in I} \bigcap_{n \in \mathbb{N}^*} \{\tau(i,n) < \infty\}] = 1$$

then it tends to zero.

Proof.

In th. 3.16 we proved $v(x,q) \leq w(x,q)$. Further we have

$$\ell(x,q) = \sup_{f \in \bar{F}} \int q(d\theta) v(x,\theta,f) = \sup_{f \in \bar{F}} v(x,q,f) \leq v(x,q) .$$

We proceed with assertion (ii). Note that for $N \in \mathbb{N}^*$

$$(*) \qquad v(x,\theta) - v(x,\theta,f) = \sum_{k=0}^{\infty} \{(L_f^{kN} v)(x,\theta) - (L_f^{(k+1)N} v)(x,\theta)\},$$

since $\lim_{n \to \infty} (L_f^n v)(x,\theta) = v(x,\theta,f)$. For bounded measurable functions b and c on $X \times \theta$ we find in a familiar way (cf. [Denardo (1967), th. 1])

$$\sup_x \{(L_f b)(x,\theta) - (L_f c)(x,\theta)\} \leq \beta \sup_x \{b(x,\theta) - c(x,\theta)\}$$

and therefore

$$\sup_x \{(L_f^{kN} v)(x,\theta) - (L_f^{(k+1)N} v)(x,\theta)\} \leq \beta^{kN} \sup_x \{v(x,\theta) - (L_f^N v)(x,\theta)\}.$$

Consequently, using (*), we find:

$$(**) \qquad v(x,\theta) - v(x,\theta,f) \leq \frac{1}{1-\beta^N} \varphi_N(\theta,f) .$$

Note that, for $N = \infty$ (**) also holds.

By 6.4 we have

$$w(x,q) - \ell(x,q) \leq \inf_{f\in\bar{F}} \int q(d\theta)\{v(x,\theta) - v(x,\theta,f)\}$$

and so, using (**) we find the desired result.

We proceed with assertion (iii).

$$\mathbb{E}^{\pi}_{x,q}[\inf_{f\in\bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f) \mid \mathcal{F}_{n-1}] \leq \inf_{f\in\bar{F}} \mathbb{E}^{\pi}_{x,q}[\int Q_n(d\theta)\varphi_N(\theta,f) \mid \mathcal{F}_{n-1}] =$$

$$= \inf_{f\in\bar{F}} \int Q_{n-1}(d\theta)\varphi_N(\theta,f), \ \mathbb{P}^{\pi}_{x,q}\text{-a.s. (cf. th. 2.1)} .$$

Hence the sequence $\{\inf_{f\in\bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f) \mid n \in \mathbb{N}\}$ is a super martingale, which establishes the first part of (iii) and the existence of

$$\lim_{n\to\infty} \inf_{f\in\bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f) .$$

Assume that $\tau(i,n) < \infty$ for all $n \in \mathbb{N}^*$ and $i \in I$, $\mathbb{P}^{\pi}_{x,q}$-a.s.

From (**) it follows that $\varphi_N(\theta,f) \geq 0$ and from corollary 2.5 that

$$\lim_{n\to\infty} \int Q_n(d\theta)\varphi_N(\theta,f) = \varphi_N(Z,f), \ \mathbb{P}^{\pi}_{x,q}\text{-a.s.}$$

Hence we have $\mathbb{P}^{\pi}_{x,q}$-a.s.:

$$0 \leq \lim_{n\to\infty} \inf_{f\in\bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f) \leq \inf_{f\in\bar{F}} \lim_{n\to\infty} \int Q_n(d\theta)\varphi_N(\theta,f) = \inf_{f\in\bar{F}} \varphi_N(Z,f).$$

Note that:

$$v(x,\theta) \geq (L_f v)(x,\theta) \geq (L_f^N v)(x,\theta) \geq v(x,\theta,f), \ x \in X, \ \theta \in \Theta \text{ and } f \in \bar{F}.$$

Hence:

$$0 \leq \inf_{f\in\bar{F}} \varphi_N(\theta,f) \leq \inf_{f\in\bar{F}} \sup_{x} \{v(x,\theta) - v(x,\theta,f)\} = 0 \text{ (cf. 6.2)} .$$

Therefore we have

$$\lim_{n\to\infty} \inf_{f\in\bar{F}} \int Q_n(d\theta)\varphi_N(\theta,f) = 0, \ \mathbb{P}^{\pi}_{x,q}\text{-a.s.}$$

Since $(\theta,f) \to \varphi_N(\theta,f)$ is bounded, the dominated convergence theorem yields the desired result. □

Remark.

The bound given in th. 6.1 (ii) has significance only, if either $\varphi_N(\theta,f)$ itself, or an approximation of it, is known for $\theta \in \Theta$ and $f \in \bar{F}$.

If $v(x,\theta)$ is computed for all $x \in X$ and $\theta \in \Theta$, and if optimal Markov policies $f_\theta \in F$, $\theta \in \Theta$ are determined, then it requires more work to compute $v(x,\theta',f_\theta)$ than to compute $(L_{f_\theta} v)(x,\theta')$ for $x \in X$, $\theta,\theta' \in \Theta$. However,

$$\sup_x \{v(x,\theta) - v(x,\theta,f)\} \leq \frac{1}{1-\beta} \sup_x \{v(x,\theta) - (L_f v)(x,\theta)\}$$

(cf. (**) in the proof of th. 6.1). So for more work we get a better bound.

In th. 6.2 we consider successive approximations of the value function.

Theorem 6.2

For $x \in X$, $q \in W$

(i) $(U^n \ell)(x,q) \leq v(x,q) \leq (U^n w)(x,q)$ .

(ii) $(U^n w)(x,q)$ is nonincreasing and $(U^n \ell)(x,q)$ is nondecreasing in n.

Proof.

Part (i) is a direct consequence of th. 6.1 (i) since U is monotone. To prove part (ii) it suffices to show $Uw \leq w$ and $U\ell \geq \ell$. Using

$$p_i(y,q)T_{i,y}(q)(d\theta) = p_i(y|\theta_i)q(d\theta)$$

and 6.4 (i) we find

$$(Uw)(x,q) = \sup_{a \in D(x)} \int q(d\theta) \sum_{i \in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) +$$

$$+ \beta \int P(dx'|x,a,y)v(x',\theta)\} \leq \int q(d\theta)v(x,\theta) = w(x,q)$$

where the inequality follows from exchanging $\sup_{a \in D(x)}$ and $\int q(d\theta)$, and the optimality equation of the dynamic program with known parameter value, i.e. $v(x,\theta) = (Uv)(x,\theta)$. Using

$$v(x',T_{i,y}(q),f) = \int T_{i,y}(q)(d\theta)v(x',\theta,f)$$

we find

$$(U\ell)(x,q) = \sup_{a\in D(x)} \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y,q)\{r(x,a,y) +$$

$$+ \beta \int P(dx'|x,a,y) \sup_{f\in\bar{F}} v(x',T_{i,y}(q),f)\} \geq$$

$$\geq \sup_{f\in\bar{F}} \sup_{a\in D(x)} \int q(d\theta) \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) +$$

$$+ \beta \int P(dx'|x,a,y) v(x',\theta,f)\} \geq \sup_{f\in\bar{F}} v(x,q,f) = \ell(x,q) \ . \qquad \square$$

In th. 6.3 we consider for each $\varepsilon > 0$ a Bayesian stationary strategy, which is easy to handle, and which is (nearly) as good as all stationary strategies in $\bar{F}$. Moreover the strategy processes new information concerning the unknown parameter in the following sence. If, under the strategy f, $\tau(i,n) < \infty$, $\mathbb{P}^f_{x,q}$-a.s. for all $i \in I$ and $n \in \mathbb{N}^*$, then we have

$$\limsup_{n\to\infty} \mathbb{E}^f_{x,q}[v(X_n,Q_n) - v(X_n,Q_n,f)] \leq \frac{\varepsilon}{1-\beta} \ .$$

Theorem 6.3

Fix $\varepsilon > 0$ and let f be a Bayesian Markov policy such that for $(x,q) \in X \times W$

$$(L_f\ell)(x,q) \geq (U\ell)(x,q) - \varepsilon \ .$$

Then

$$v(x,q,f) \geq \ell(x,q) - \frac{\varepsilon}{1-\beta}$$

and if $\mathbb{P}^f_{x,q}[\bigcap_{i\in I} \bigcap_{n\in\mathbb{N}^*} \{\tau(i,n) < \infty\}] = 1$, then

$$\limsup_{n\to\infty} \mathbb{E}^f_{x,q}[\{v(X_n,Q_n) - v(X_n,Q_n,f)\}] \leq \frac{\varepsilon}{1-\beta} \ .$$

Proof.

Let 1 be the function on $X \times W$ which is identically equal to one. By the proof of th. 6.2 (ii):

$$L_f\ell \geq U\ell - \varepsilon 1 \geq \ell - \varepsilon 1 \ .$$

Assume: $L_f^n\ell \geq \ell - \varepsilon(\frac{1-\beta^n}{1-\beta})1$. Then:

$$L_f^{n+1}\ell \geq L_f\ell - \varepsilon\beta(\frac{1-\beta^n}{1-\beta})1 \geq \ell - \varepsilon\{1 + \beta(\frac{1-\beta^n}{1-\beta})\}1 = \ell - \varepsilon(\frac{1-\beta^{n+1}}{1-\beta})1 .$$

Hence, if we let n tend to infinity, we get:

$$v(x,q,f) = \lim_{n\to\infty} (L_f^n \ell)(x,q) \geq \ell(x,q) - \frac{\varepsilon}{1-\beta} .$$

This proves the first statement.

The second statement is a consequence of th. 6.1 (iii), since

$$v(X_n,Q_n) - v(X_n,Q_n,f) \leq w(X_n,Q_n) - \ell(X_n,Q_n) +$$

$$+ \frac{\varepsilon}{1-\beta} \leq \inf_{\bar{f}\in\bar{F}} \int Q_n(d\theta)\varphi_\infty(\theta,\bar{f}) + \frac{\varepsilon}{1-\beta} . \qquad \square$$

For the Bayesian equivalent rule considered in section 4.1 (cf. 4.3a) we give in th. 6.4 a lower bound on its Bayesian discounted total return. Hence we consider in th. 6.4 a Bayesian Markov policy such that

$$\int q(d\theta) \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) +$$

$$+ \beta \int P(dx'|x,a,y) v(x',\theta)\}$$

is maximized within an $\varepsilon$-bound.

The strategy is "adaptive" in the same sence as the strategy in th. 6.3.

Theorem 6.4

Let $\varepsilon > 0$ and let f be a Bayesian Markov policy such that, for $(x,q) \in X \times W$:

$$(L_f w)(x,q) \geq (Uw)(x,q) - \varepsilon .$$

Then f is a Bayesian equivalent rule as considered above, and

$$v(x,q,f) \geq w(x,q) - \frac{1}{1-\beta}\{\inf_{\bar{f}\in\bar{F}} \int q(d\theta)\varphi_1(\theta,\bar{f}) + \varepsilon\} .$$

If $\mathbb{P}^f_{x,q}[\bigcap_{i\in I} \bigcap_{n\in\mathbb{N}^*} \{\tau(i,n) < \infty\}] = 1$, then

$$\limsup_{n\to\infty} \mathbb{E}^f_{x,q}[\{v(X_n,Q_n) - v(X_n,Q_n,f)\}] \leq \frac{\varepsilon}{1-\beta} .$$

Proof.

To verify that f is a Bayesian equivalent rule as considered above note that

$$(L_f w)(x,q) = \int q(d\theta) \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)\{r(x,a,y) +$$

$$+ \beta \int P(dx' \mid x,a,y) v(x',\theta) \} \quad .$$

We have $\mathbb{P}^f_{x,q}$-a.s.:

$$(L_f w)(X_n,Q_n) = \tilde{r}(X_n,Q_n,A_n) + \beta \mathbf{E}^f_{x,q}[\, w(X_{n+1},Q_{n+1}) \mid X_n,Q_n]$$

(cf. 3.1(e) for the definition of $\tilde{r}$). Hence:

$$\mathbf{E}^f_{x,q}[\sum_{n=0}^{\infty} \beta^n \tilde{r}(X_n,Q_n,A_n)] = \mathbf{E}^f_{x,q}[\sum_{n=0}^{\infty} \beta^n (L_f w)(X_n,Q_n) - \sum_{n=1}^{\infty} \beta^n w(X_n,Q_n)] .$$

And therefore, by the definition of f we have:

$$(*) \qquad v(x,q,f) = w(x,q) + \mathbf{E}^f_{x,q}[\sum_{n=0}^{\infty} \beta^n \{ (L_f w)(X_n,Q_n) - w(X_n,Q_n)\}] \geq$$

$$\geq w(x,q) + \mathbf{E}^f_{x,q}[\sum_{n=0}^{\infty} \beta^n \{ (Uw)(X_n,Q_n) - w(X_n,Q_n)\}] - \frac{\varepsilon}{1-\beta} .$$

Since

$$(Uw)(x,q) = \sup_{f \in \mathfrak{F}} \int (L_f v)(x,\theta) q(d\theta), \quad (x,q) \in X \times W ,$$

we have:

$$(Uw)(x,q) - w(x,q) \geq \sup_{\bar{f} \in \bar{F}} \int q(d\theta) \{ (L_{\bar{f}} v)(x,\theta) - v(x,\theta)\} \geq$$

$$\geq -\inf_{\bar{f} \in \bar{F}} \int q(d\theta) \varphi_1(\theta,\bar{f}) .$$

And therefore:

$$(**) \qquad \mathbf{E}^f_{x,q}[(Uw)(X_n,Q_n) - w(X_n,Q_n)] \geq -\mathbf{E}^f_{x,q}[\inf_{\bar{f} \in \bar{F}} \int Q_n(d\theta) \varphi_1(\theta,\bar{f})] \geq$$

$$\geq -\inf_{\bar{f} \in \bar{F}} \mathbf{E}^f_{x,q}[\int Q_n(d\theta) \varphi_1(\theta,\bar{f})] = -\inf_{\bar{f} \in \bar{F}} \int q(d\theta) \varphi_1(\theta,\bar{f}) .$$

Combination of (*) and (**) yields the first statement.

To prove the second statement assume that $\tau(i,n) < \infty$, $\mathbb{P}^f_{x,q}$-a.s. Note that, by the first statement:

$$v(X_n,Q_n) - v(X_n,Q_n,f) \leq w(X_n,Q_n) - v(X_n,Q_n,f) \leq$$

$$\leq \frac{1}{1-\beta} \inf_{\bar{f} \in \bar{F}} \int Q_n(d\theta) \varphi_1(\theta,\bar{f}) + \frac{\varepsilon}{1-\beta} .$$

Hence the desired result follows from th. 6.1 (iii). □

Remark.

Note that, by th. 6.1 (ii),

$$w(x,q) - \frac{1}{1-\beta} \inf_{\bar{f} \in \bar{F}} \int q(d\theta)\varphi_1(\theta,\bar{f}) \leq \ell(x,q) \ .$$

Hence the lowerbound on $v(x,q,f)$ in th. 6.4 is not better than the lowerbound we found for the strategy in th. 6.3.

In practise, when we are dealing with finite sets X, Y and A we often have to approximate the value function v. Th. 6.2 gives us the opportunity to do this as accurately as we like. However it is impossible to compute, for example, $(U^N w)(x,q)$ for all $(x,q) \in X \times W$, since W is not finite. Nevertheless it is possible to compute $(U^N w)(x,q)$ for a fixed $q \in W$, since the number of possible posterior distributions after N transitions is finite. Hence we have to determine a *horizon* $N \in \mathbb{N}$ such that

$$|(U^N w)(x,q) - (U^N \ell)(x,q)| \leq \varepsilon$$

where $\varepsilon > 0$ is the maximal allowed error in the approximation for $v(x,q)$. Then we compute $\frac{1}{2}\{(U^N w)(x,q) + (U^N \ell)(x,q)\}$, which is an acceptable approximation for $v(x,q)$ (cf. th. 6.2). To determine N we have to compute first $(U^n w)(x,q) - (U^n \ell)(x,q)$ for $n = n_0, n_0+1, \ldots, N$, where $n_0$ is a lowerbound on the horizon. In general the horizon determination in this way is very time consuming compared to the backward induction to compute $(U^N\{\frac{1}{2}(w+\ell)\})(x,q)$, another acceptable approximation of $v(x,q)$. To see this, we note that in general the sets $W_n(q)$ and $W_m(q)$ of possible posterior distributions of q after n and m transitions, respectively, are disjoint if $m \neq n$ (cf. the remarks at the end of this section).

Hence, to compute $(U^n b)(x,q)$ for some bounded measurable function b on $X \times W$, we first have to compute $(Ub)(x,\tilde{q})$ for $x \in X$ and all $\tilde{q} \in W_{n-1}(q)$ and afterwards $(Ub)(x,\tilde{q})$ for $x \in X$ and $\tilde{q} \in W_{n-2}(q)$ etc. So we have computed, together with $(U^n b)(x,q)$, the set of values

$$\bigcup_{m=1}^{n-1} \{(U^m b)(x,\tilde{q}) \mid x \in X,\ \tilde{q} \in W_{n-m}(q)\} \ .$$

However, to compute $(U^{n+1} b)(x,q)$, we need the values:

$$\bigcup_{m=1}^{n} \{ (U^m b)(x,\tilde{q}) \mid x \in X,\ \tilde{q} \in W_{n+1-m}(q) \}$$

and since $W_{n+1-m}(q) \cap W_{n-m}(q) = \emptyset$ in general, we cannot use the already computed values for the computation of $(U^n b)(x,q)$, to determine $(U^{n+1} b)(x,q)$. (We return to this matter in the next chapter.)

It will be clear that it would be nice to have a simpler method to determine a suitable horizon. Indeed such a procedure exists when we are dealing with the simple parameter structure that we introduce below.

*Assumption on the parameter structure*

6.6 Let $I := \{1,2,\ldots,t\}$ and let $\theta_i$ be a singleton for $i = 2,\ldots,t$. Further let $\{L_1,L_2\}$ be a measurable partition of X and let $K_1 := L_1 \times A$.

The models 4 and 5, considered in chapter 5, satisfy 6.6 in a trivial way: there we have $L_2 = \emptyset$. In chapter 7 we consider other models satisfying 6.6 (cf. examples 7.4 and 7.5). In the rest of this section we assume that 6.6 holds.

Note that for states $x \in L_2$ the transition law is completely known and for $x \in L_1$ it is incompletely known but the chosen action does not influence the kind of information we get after the transition. It is easy to verify that $q = \bigotimes_{i \in I} q_i$ for all $q \in W$, in this situation, since $q_i(\{\theta_i\}) = 1$ for all $i \geq 2$.

Consider the stopping time $\sigma$

6.7 $$\sigma := \inf\{n > 0 \mid X_n \in L_1\} .$$

We shall use the optimal reward operator $U_\sigma$ (cf. section 3.2). Let $x \in L_2$ and let b be a bounded measurable function on $X \times W$. Recall:

$$(U_\sigma b)(x,q) = \sup_{\pi \in \Pi_0} \mathbf{E}^{\pi}_{x,q} [ \sum_{n=0}^{\sigma-1} \beta^n r(X_n,A_n,Y_{n+1}) + \beta^\sigma b(X_\sigma,Q_\sigma) ] .$$

Next we discuss a nice property of this operator.

If $X_0 \in L_2$ then we have $(X_n,A_n) \in \bigcup_{2 \leq i \leq t} K_i$ for $n < \sigma$. Hence the expectation of the first term does not depend on $q \in W$ (however, it does depend on the known parameter values $\theta_2,\ldots,\theta_t$). Further we note that

6.8 $\tau(1,1) = \sigma + 1$ if $X_0 \in L_2$

(cf. 2.17 for the definition of $\tau$). Hence, if $X_0 \in L_2$ then $Q_\sigma = Q_0 = q$, and therefore we may write:

6.9 $$(U_\sigma b)(x,q) = \sup_{\pi \in \Pi_0} E_x^\pi [\sum_{n=0}^{\sigma-1} \beta^n r(X_n, A_n, Y_{n+1}) + \beta^\sigma b(X_\sigma, q)], \quad x \in L_2$$

since the expectation does not depend on q.
The computation of $(U_\sigma b)(x,q)$ for $x \in L_2$ is an ordinary dynamic programming problem which is feasible if X, Y and A are finite (this will be clarified in chapter 7).

In th. 6.5 we assume that the function b on $X \times W$ is an approximation of v, such that

$$|v(x,q) - b(x,q)| \leq \varepsilon(q), \quad x \in X, \ q \in W .$$

First we introduce some notations:

6.10 (i) For $q \in W$ and $y_1, \ldots, y_n \in Y$ we define the probability $\chi_q(y_1, \ldots, y_n)$ on $\Theta$ by

$$\chi_q(y_1, \ldots, y_n)(B) := \int_B \prod_{j=1}^n p_1(y_j|\theta_1) q(d\theta) \{ \int_\Theta \prod_{j=1}^n p_1(y_j|\theta_1) q(d\theta) \}^{-1}$$

if the denominator is positive;

$:= q(B)$ otherwise $(B \in \mathcal{T})$ .

(ii) $$E(q,\varepsilon,n) := \int q(d\theta) \{ \int \ldots \int \nu(dy_1) \ldots \nu(dy_n) \cdot \prod_{j=1}^n p_1(y_j|\theta_1) \varepsilon(\chi_q(y_1, \ldots, y_n)) \}$$

where $\varepsilon$ is a real-valued, bounded measurable function on W.

It is easy to verify that, if $L_2 = \emptyset$ then $E(q,\varepsilon,n) = E_q[\varepsilon(Q_n)]$ since $\tau(1,n) = n$ for all $n \in \mathbb{N}^*$, in this situation (note that here the expectation is independent of the starting state and the strategy).

6.11 (i) $$\varepsilon_0(q) := \tfrac{1}{2} \inf_{f \in \bar{F}} \int q(d\theta) \sup_{x \in L_1} \{v(x,\theta) - v(x,\theta,f)\};$$

$$b_0(x,q) := \tfrac{1}{2}\{w(x,q) + \ell(x,q)\} .$$

$$\text{(ii)}\quad \varepsilon_k(q) := \tfrac{1}{2}\,\frac{1}{1-\beta^k}\,\inf_{f\in\bar{F}}\int q(d\theta)\varphi_k(\theta,f);$$

$$b_k(x,q) := w(x,q) - \varepsilon_k(q),\ k \in \bar{\mathbb{N}}^* .$$

Theorem 6.5

(i) Let b be a bounded measurable function on $X \times W$ and let $\varepsilon$ be a bounded nonnegative measurable function on W such that for $x \in L_1$ and $q \in W$:

6.12 $$|v(x,q) - b(x,q)| \le \varepsilon(q) .$$

Then, for $x \in X$, $q \in W$, $n \in \mathbb{N}^*$:

6.13 $$|v(x,q) - (U_\sigma^n b)(x,q)| \le \beta^n E(q,\varepsilon,n) \quad \text{if } x \in L_1$$
$$\le \beta^{n-1} E(q,\varepsilon,n-1) \quad \text{if } x \in L_2 .$$

(ii) In particular the functions $b_k$ and $\varepsilon_k$ (cf. 6.11) for $k \in \bar{\mathbb{N}}$ satisfy 6.12 and $E(q,\varepsilon_k,n)$ is nonincreasing in n with $\lim_{n\to\infty} E(q,\varepsilon_k,n) = 0$, for $q \in W$.

Proof.

Part (i). Define the operator $\tilde{U}_\sigma$ on the bounded measurable functions on $X \times W$ by:

$$(\tilde{U}_\sigma f)(x,q) := \sup_{\pi\in\Pi_0} \mathbf{E}^\pi_{x,q}[\beta^\sigma f(X_\sigma, Q_\sigma)] .$$

Note that this is an optimal reward operator of the kind we studied in section 3.2, for the model with r identically zero. We define $\varepsilon^*(x,q) := \varepsilon(q)$, $x \in X$, $q \in W$. Using corollary 3.13 and th. 3.14 (ii), we find:

$$v = U_\sigma^n v = U_{\sigma_n} v \le U_{\sigma_n}(b+\varepsilon^*) \le (U_{\sigma_n} b) + (\tilde{U}_{\sigma_n} \varepsilon^*),$$

where $\sigma_1 := \sigma$ and $\sigma_n := \sigma \circ \sigma_{n-1}$ (cf. 3.14 and 3.23). And similarly

$$v = U_{\sigma_n} v \ge U_{\sigma_n}(b-\varepsilon^*) \ge (U_{\sigma_n} b) - (\tilde{U}_{\sigma_n} \varepsilon^*) .$$

Hence

$$|v(x,q) - (U_\sigma^n b)(x,q)| \le (\tilde{U}_{\sigma_n} \varepsilon^*)(x,q),\ x \in X,\ q \in W .$$

Next we consider $(\tilde{U}_{\sigma_n} \varepsilon^*)(x,q)$ in more detail.

Note that $\tau(1,1) = \sigma + 1$ if $X_0 \in L_2$ and $\tau(1,2) = \sigma + 1$ if $X_0 \in L_1$ (remember that $\tau(1,1) = 1$ if $X_0 \in L_1$). By induction it is easy to verify that

$$\sigma_n = \inf\{k > \sigma_{n-1} \mid X_k \in L_1\},\ n = 2,3,\ldots .$$

We show, using induction, that for $n \in \mathbb{N}^*$:

$$\tau(1,n) = \sigma_n + 1 \text{ if } X_0 \in L_2 \quad \text{and} \quad \tau(1,n+1) = \sigma_n + 1 \text{ if } X_0 \in L_1 .$$

For $n = 1$ the statement is true, so assume it holds for $n$ ($n \geq 2$). If $X_0 \in L_2$ we have

$$\tau(1,n+1) = \inf\{k > \tau(1,n) \mid X_{k-1} \in L_1\} =$$

$$= \inf\{k > \sigma_n + 1 \mid X_{k-1} \in L_1\} = 1 + \inf\{k > \sigma_n \mid X_k \in L_1\} = \sigma_{n+1} + 1.$$

Similarly if $X_0 \in L_1$. Remember that $\tau(1,n) \geq n$ for $n \in \mathbb{N}^*$. Hence

$$(\tilde{U}_{\sigma_n} \varepsilon^*)(x,q) = \sup_{\pi \in \Pi_0} \mathbb{E}^{\pi}_{x,q}[\beta^{\sigma_n} \varepsilon(Q_{\sigma_n})] \leq \sup_{\pi \in \Pi_0} \beta^k \mathbb{E}^{\pi}_{x,q}[\varepsilon(Q_{\tau(1,k)-1})]$$

with $k = n$ if $x \in L_2$ and $k = n+1$ if $x \in L_1$.
Note that $Q_n = \bigotimes_{i \in I} Q_{i,n}$ and $Q_{i,n}(\{\theta_i\}) = 1$ for $i \geq 2$. According to 2.26 we have for $B \in \mathcal{T}$ $\mathbb{P}^{\pi}_{x,q}$-a.s. (cf. the proof of th. 2.1):

$$Q_n(B) = \int_B \prod_{\{\ell > 0 \mid \tau(1,\ell) \leq n\}} p_1(Y_{\tau(1,\ell)} \mid \theta_1) q(d\theta)$$

$$\cdot \Big\{ \int_{\theta} \prod_{\{\ell > 0 \mid \tau(1,\ell) \leq n\}} p_1(Y_{\tau(1,\ell)} \mid \theta_1) q(d\theta) \Big\}^{-1} .$$

Hence for $k = 2,3,\ldots$ we have $\mathbb{P}^{\pi}_{x,q}$-a.s.:

$$Q_{\tau(1,k)-1}(B) = \int_B \prod_{j=1}^{k-1} p_1(Y_{\tau(1,j)} \mid \theta_1) q(d\theta)$$

$$\cdot \Big\{ \int_{\theta} \prod_{j=1}^{k-1} p_1(Y_{\tau(1,j)} \mid \theta_1) q(d\theta) \Big\}^{-1} .$$

Therefore

$$Q_{\tau(1,k)-1} = \chi_q(Y_{\tau(1,1)}, \ldots, Y_{\tau(1,k-1)}),\ \mathbb{P}^{\pi}_{x,q}\text{-a.s.}$$

By lemma 2.2 we find:

$$\mathbf{E}^{\pi}_{x,q}[\varepsilon(Q_{\tau(1,k)-1})] \leq E(q,\varepsilon,k-1), \quad k = 2,3,\ldots .$$

This proves part (i). We proceed with part (ii).
It is easy to verify that 6.12 holds for the functions $b_k$ and $\varepsilon_k$, $k \in \bar{\mathbf{N}}$. Hence 6.13 holds by part (i). As already noted we have

$$E(q,\varepsilon,n) = \mathbf{E}_q[\varepsilon(Q_n)]$$

for the model with $L_2 = \emptyset$ (note that the expectation is independent of the starting state x and the strategy $\pi$, here). For this model the assumption of th. 6.3 (iii) is true, which implies that $E(q,\varepsilon,n)$ converges monotonically to zero, as n tends to infinity, in case $k \in \bar{\mathbf{N}}^*$. For $k = 0$ the proof is analogous. □

In chapter 7 we discuss algorithms in which the horizon determination is based on th. 6.5 (ii). It turns out that computation of $E(q,\varepsilon,n)$ is rather easy compared with the backward induction.

Corollary 6.6

If $L_2 = \emptyset$ (and 6.6 holds) we have

$$|v(x,q) - (U^n b_k)(x,q)| \leq \beta^n \mathbf{E}_q[\varepsilon_k(Q_n)], \quad n \in \mathbf{N}^*, \quad k \in \bar{\mathbf{N}}$$

($b_k$ and $\varepsilon_k$ are defined in 6.11).
This statement is already proved in the proof th. 6.5 (ii).

For the functions $\varepsilon_k$, $k \in \bar{\mathbf{N}}$ defined in 6.11, 6.10 (ii) reduces to:

6.14
$$E(q,\varepsilon_0,n) = \tfrac{1}{2} \int \ldots \int \nu(dy_1)\ldots\nu(dy_n)$$
$$\cdot \inf_{f\in\bar{F}} \int q(d\theta) \prod_{j=1}^{n} p_1(y_j|\theta_1) \sup_{x\in L_1} \{v(x,\theta_1) - v(x,\theta_1,f)\}$$

and for $k \in \bar{\mathbf{N}}^*$:

$$E(q,\varepsilon_k,n) = \tfrac{1}{2} \frac{1}{1-\beta^k} \int \ldots \int \nu(dy_1)\ldots\nu(dy_n)$$
$$\cdot \inf_{f\in\bar{F}} \int q(d\theta) \prod_{j=1}^{n} p_1(y_j|\theta_1)\varphi_k(\theta_1,f) .$$

To verify this, note that

$$\varepsilon_k(\chi_q(y_1,\ldots,y_n)) = \tfrac{1}{2}\,\frac{1}{1-\beta^k}\,\inf_{f\in\bar{F}}\frac{\int q(d\theta)\prod_{j=1}^{n} p_1(y_j|\theta_1)\varphi_k(\theta,f)}{\int q(d\theta)\prod_{j=1}^{n} p_1(y_j|\theta)}\,,\quad k\in\bar{\mathbb{N}}^*\,,$$

if the denominator is positive.

Further note that, for all $q \in W$:

$$\varepsilon_0(q) \le \varepsilon_\infty(q) \le \varepsilon_{k+1}(q) \le \varepsilon_k(q), \quad \text{for } k \in \mathbb{N}^*\,,$$

and note that $\varepsilon_0(q) = \varepsilon_\infty(q)$ if $L_2 = \emptyset$.

We conclude this section with some remarks. The first four remarks complement the results we derived in this section. The last three remarks concern other approaches, not treated here.

Remarks.

(i) In most situations the sets of possible posterior distributions at successive stages are disjoint. However the following example shows that this is not always the case.

Let 6.6 hold and let $L_2 = \emptyset$. Further let $\theta_1 := \{t, 1-t\}$, $0 < t < \tfrac{1}{2}$ and let $Y := \{0,1\}$ and $p_1(y|\theta_1) := \theta_1^y(1-\theta_1)^{1-y}$, $y \in Y$, $\theta_1 \in \theta_1$. It is easy to verify that if $q(\{t\}) := \pi$, then the posterior distribution after n transitions is:

$$Q_n(\{t\}) = \{1 + (\frac{t}{1-t})^{(n-2\sum_{i=1}^{n} Y_i)}\cdot\frac{(1-\pi)}{\pi}\}^{-1}\,.$$

Hence if $n > m$ and $\sum_{i=m+1}^{n} Y_i = \frac{n-m}{2}$ then $Q_n = Q_m$.

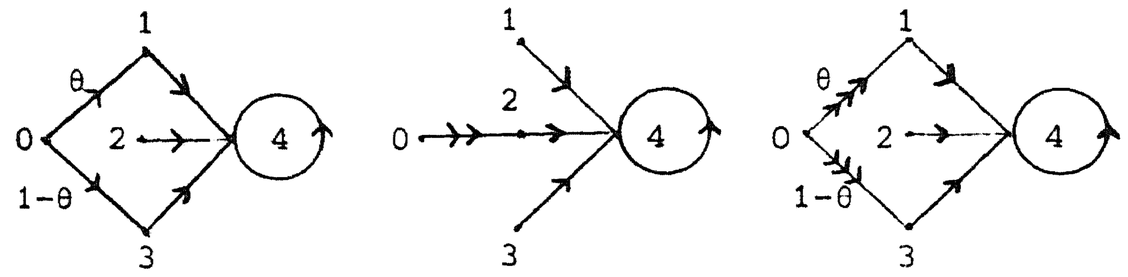
(ii) We have already suggested a choice for the set $\bar{F}$ (see 6.2). Now we consider:

$$\bar{F} := \{f \in F \mid \text{for some } \theta \in \theta\colon v(x,\theta) = v(x,\theta,f) \text{ for all } x \in X\}\,.$$

At first sight one may expect that the best Markov policy for the Bayes criterion can be found in $\bar{F}$. However for an example we show that

$$\sup_{f\in\bar{F}} v(x,q,f) < \sup_{f\in F} v(x,q,f)\,.$$

Counterexample




$X := \{0,1,2,3,4\}$, $D(0) := A := \{1,2,3\}$, $D(x) := \{1\}$, $x \neq 0$.
$\theta := \{\frac{1}{10}, \frac{9}{10}\}$. Consider the transition probability $P(x'|x,a)$ from $X \times A$ to $X$:

$$P(1|0,1) = 1 - P(3|0,1) = \theta, \quad P(3|0,3) = 1 - P(1|0,3) = \theta$$

$$P(2|0,2) = P(4|1,1) = P(4|2,1) = P(4|3,1) = P(4|4,1) := 1 .$$

Only in the states 1, 2 and 3 a reward is obtained: 110, 70 and 10 respectively. It is easy to fit this example into our framework. In case of known parameter values actions 1 or 3 are optimal in state 0 but action 2 is never optimal. We identify the three possible Markov policies with the actions chosen in state 0. Hence $\bar{F} = \{1,3\}$. Let $q \in W$ be such that $q(\{\theta\}) = \frac{1}{2}$ for $\theta \in \Theta$ and let $\beta = \frac{1}{2}$. Then

$$v(0,q,2) = 35 \quad \text{and} \quad v(0,q,1) = v(0,q,3) = 30 .$$

(iii) If there exists a $f^* \in F$ such that $v(x,\theta,f^*) = v(x,\theta)$ for all $x \in X$, $\theta \in \Theta$ then $v(x,q,f^*) = \int q(d\theta)v(x,\theta) = w(x,q)$, for $x \in X$ and $q \in W$ and therefore $f^*$ is optimal.

(iv) If $b$ is a bounded measurable function on $X \times W$ such that $b(x,q) = \int b(x,\theta)q(d\theta)$ for all $x \in X$, $q \in W$ then

$$(U^n b)(x,q) \leq \int q(d\theta)(U^n b)(x,\theta), \quad n \in \mathbb{N}^* .$$

To prove this note that using arguments of the proof of th. 6.2 we find: $(Ub)(x,q) \leq \int q(d\theta)(Ub)(x,\theta)$. Hence, by putting $b'(x,q) := \int q(d\theta)(Ub)(x,\theta)$ and repeating the argument, we have

$$(U^2 b)(x,q) \leq (Ub')(x,q) \leq \int q(d\theta)(U^2 b)(x,\theta) ,$$

since $(Ub')(x,\theta) = (U^2 b)(x,\theta)$. The statement follows by induction.

(v) In [Martin (1967)] the usual method of successive approximations is described for Bayesian control models with finite state and action spaces. Martin suggested the use of "scrap functions" b on $X \times W$ that are constant on W (in fact Martin specifies a function $b^*$ on X and he sets $b(x,q) := b^*(x)$ for $q \in W$). Then he approximates $v(x,q)$ by $(U^n b)(x,q)$. The difficulty of this method is the choice of the horizon such that $|v(x,q) - (U^n b)(x,q)|$ is sufficiently small. He gives the following bound for this difference (cf. [Martin (1967), th. 3.4.3]):

$$(*) \qquad \beta^n \max\{\bar{b} - \frac{m}{1-\beta}, \frac{M}{1-\beta} - \underline{b}\}$$

where

$$M := \sup_{x,a,y} r(x,a,y), \quad m := \inf_{x,a,y} r(x,a,y), \quad \bar{b} := \sup_x b^*(x)$$

and

$$\underline{b} := \inf_x b^*(x) .$$

To verify this, note:

$$(U^n 0)(x,q) + \beta^n \frac{m}{1-\beta} \le (U^n v)(x,q) \le (U^n 0)(x,q) + \beta^n \frac{M}{1-\beta}$$

and

$$(U^n 0)(x,q) + \beta^n \underline{b} \le (U^n b)(x,q) \le (U^n 0)(x,q) + \beta^n \bar{b}.$$

Since $v = U^n v$ we have

$$\beta^n\{\frac{m}{1-\beta} - \bar{b}\} \le v(x,q) - (U^n b)(x,q) \le \beta^n\{\frac{M}{1-\beta} - \underline{b}\} .$$

It is obvious that this bound (*) is minimized by setting $b_m := b_M := \frac{1}{2} \frac{M+m}{1-\beta}$. Then the bound becomes $\frac{1}{2} \frac{\beta^n}{1-\beta}(M-m)$ which is poor, in general. In our approach a better scrap function is suggested for the special parameter structure given in 6.6, and the convergence of the posterior distributions is used to get a smaller horizon (see th. 6.5) (see also chapter 7 for some examples).

(vi) The use of upper and lower bounds is also suggested in [Satia and Lave (1973)]. The authors consider bounds of the form:

$$ub(x,q) := \sup_{f \in F} \sup_{\theta \in \theta_q} v(x,\theta,f)(1-\varepsilon) + \frac{M}{1-\beta}\varepsilon$$

$$\ell b(x,q) := \sup_{f \in F} \inf_{\theta \in \theta_q} v(x,\theta,f)(1-\varepsilon) + \frac{m}{1-\beta}\varepsilon$$

where $\theta_q \subset \theta$ such that $\mathbb{P}_q[Z \in \theta_q] \geq 1 - \varepsilon$ for some fixed $\varepsilon > 0$ and m and M are defined in the foregoing remark.

They compute their bounds with very time-consuming algorithms for Markov games. It is clear that $ub(x,q) \geq w(x,q)$ and $\ell b(x,q) \leq \ell(x,q)$, if $\bar{F}$ is defined as in remark (ii).

(vii) In [Waldmann (1976)] the space W is approximated by a finite subset of W, i.e. a finite (measurable) partition of W is constructed, and in each set of this partition a representative is chosen. Then the transition law is modified such that the process only visits these representative points. Waldmann suggests to solve the modified dynamic program with state space $X \times \tilde{W}$, where $\tilde{W}$ is the set of representative points. The value function of this dynamic program is an approximation for the value function of the original model. The idea of approximating a dynamic program with an uncountable state space by one with a finite state space is also found in [Whitt (1976)]. Whitt also provides bounds on the approximation.

(viii) In [Van Hee (1977)] a generalization of the well-known MacQueen extrapolation is considered (see [MacQueen (1966)]) for the situation where 6.6 holds and $L_2 = \emptyset$.

## 6.2 *Discretizations*

Although most of the material presented in section 6.2 is valid if X, Y and A are noncountable, the results have practical relevance only if these sets are finite. However, we do not assume that $\theta$ is finite, but rather we study the problems caused by $\theta$ being infinite.

First we consider the determination of the upper and lower bounds given in th. 6.1. We recall that, if X, Y and A are finite, the computation of $(U_\sigma^{n_{\frac{1}{2}}}(w + \ell))(x,q)$ for fixed $q \in W$ and $n \in \mathbb{N}^*$ is rather simple if $w(x',q')$ and $\ell(x',q')$ are known, for $x' \in X$, $q' \in W$. To approximate these upper and lower bound we approximate $\int v(x,\theta)q(d\theta)$ and $\int v(x,\theta,f)q(d\theta)$ using straightforward numerical integration methods.

Afterwards we shall consider another approach, namely the "finitization" of the parameter set in advance. This means that we only consider prior distributions that are concentrated in finitely many points. It is easy to verify that in that situation, all posterior distributions are also concentrated on this finite subset of $\theta$.

For both cases we give bounds on the errors caused by the discretizations.

We start with a result on perturbations of the function $\theta \to v(x,\theta)$. In the proof of th. 6.7 we use the same technique as used in [Whitt (1976), th. 6.3].

Theorem 6.7

Let $\theta,\tilde{\theta} \in \Theta$. Then:

$$\sup_{x\in X} \{v(x,\theta) - v(x,\tilde{\theta})\} \leq \frac{1}{1-\beta} \operatorname{span}(r) \frac{\frac{1}{2}\Delta(\theta,\tilde{\theta})}{1-\beta+\frac{1}{2}\beta\Delta(\theta,\tilde{\theta})} ,$$

where

$$6.15 \qquad \Delta(\theta,\tilde{\theta}) := \max_{i\in I} \int \nu(dy) \left| p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i) \right|$$

and

$$\operatorname{span}(r) := \sup_{x,a,y} r(x,a,y) - \inf_{x,a,y} r(x,a,y) .$$

Proof.

First assume $\inf_{x,a,y} r(s,a,y) = 0$. For each $\varepsilon > 0$ there is an action $a \in D(x)$ such that

$$v(x,\theta) \leq \varepsilon + \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) p_i(y|\theta_i)$$

$$\cdot \{r(x,a,y) + \beta \int P(dx'|x,a,y) v(x',\theta)\} .$$

Hence:

$$v(x,\theta) - v(x,\tilde{\theta}) \leq \varepsilon + \sum_{i\in I} 1_{K_i}(x,a) \int \nu(dy) [\{p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)\} r(x,a,y) +$$

$$+ \beta \int P(dx'|x,a,y) \{v(x',\theta) p_i(y|\theta_i) - v(x',\tilde{\theta}) p_i(y|\tilde{\theta}_i)\}] \leq$$

$$\leq \varepsilon + \sum_{i\in I} 1_{K_i}(x,a) [\int \nu(dy) \{p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)\}^+ \sup_{x,a,y} r(x,a,y) +$$

$$+ \beta \int \nu(dy) P(dx'|x,a,y) \{v(x',\theta) - v(x',\tilde{\theta})\} \min\{p_i(y|\theta_i), p_i(y|\tilde{\theta}_i)\}$$

$$+ \beta \int \nu(dy) P(dx'|x,a,y) \{p_i(y|\theta_i) - \min(p_i(y|\theta_i), p_i(y|\tilde{\theta}_i))\} v(x',\theta)] .$$

Remember that $\int \nu(dy) p_i(y|\theta_i) = 1$. Note that, for $i \in I$

$$1 - \int \nu(dy) \min\{p_i(y|\theta_i), p_i(y|\tilde{\theta}_i)\} = \int \nu(dy) \{p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)\}^+ =$$

$$= \tfrac{1}{2} \int \nu(dy) |p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)| .$$

Let $\Delta := \Delta(\theta,\tilde{\theta})$ and $M := \sup_{x,a,y} r(x,a,y)$. Then

$$v(x,\theta) - v(x,\tilde{\theta}) \le \varepsilon + \tfrac{1}{2}\Delta M + \beta \sum_{i \in I} 1_{K_i}(x,a)$$

$$\cdot \Big[\sup_{x'} |v(x',\theta) - v(x',\tilde{\theta})| \{1 - \tfrac{1}{2} \int \nu(dy) |p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)|\} +$$

$$+ \frac{M}{1-\beta} \tfrac{1}{2} \int \nu(dy) |p_i(y|\theta_i) - p_i(y|\tilde{\theta}_i)|\Big] .$$

Note that

$$\sup_{x'} |v(x',\theta) - v(x',\tilde{\theta})| \le \frac{1}{1-\beta} M .$$

Hence

$$v(x,\theta) - v(x,\tilde{\theta}) \le \varepsilon + \tfrac{1}{2}\Delta M + \beta \sup_{x'} |v(x',\theta) - v(x',\tilde{\theta})| +$$

$$+ \beta \tfrac{1}{2} \Delta \{\frac{M}{1-\beta} - \sup_{x'} |v(x',\theta) - v(x',\tilde{\theta})|\} .$$

And therefore, by rearranging terms and omitting $\varepsilon$, we find:

$$(*) \qquad \sup_x \{v(x,\theta) - v(x,\tilde{\theta})\} \le \frac{\frac{1}{2}\Delta}{1-\beta} M + \beta(1 - \tfrac{1}{2}\Delta) \sup_x |v(x,\theta) - v(x,\tilde{\theta})| .$$

If $m := \inf_{x,a,y} r(x,a,y) \neq 0$ then we first subtract m from r and afterwards we add m again. This causes M to be replaced by span(r). Now we exchange $\theta$ and $\tilde{\theta}$. Then we get

$$\sup_x |v(x,\theta) - v(x,\tilde{\theta})| \le \tfrac{1}{2} \frac{\Delta}{1-\beta} \operatorname{span}(r) + \beta(1 - \tfrac{1}{2}\Delta) \sup_x |v(x,\theta) - v(x,\tilde{\theta})|$$

which proves the theorem. □

Remarks.

(i) If $\{\theta \to p_i(y|\theta_i);\ y \in Y\}$ is equicontinuous for all $i \in I$ then the function $\{\theta \to v(x,\theta);\ x \in X\}$ is equicontinuous. This is an immediate consequence of th. 6.7.

(ii) If $f \in F$ then

$$\sup_x \left| v(x,\theta,f) - v(x,\tilde{\theta},f) \right| \leq \frac{\operatorname{span}(r)}{1-\beta} \frac{\frac{1}{2}\Delta(\theta,\tilde{\theta})}{1-\beta+\frac{1}{2}\beta\Delta(\theta,\tilde{\theta})} .$$

The proof is exactly the same if we assume $D(x) = \{f(x)\}$, $x \in X$.

Assume 6.6 and identify $\theta_1$ and $\theta$. We shall split up the parameter space into a measurable partition $\{B_1,\ldots,B_n\}$ and we assume that in each set $B_j$ a *point of discretization* $b_j$ is fixed. Further we suppose that, for $j = 1,..,n$ and $x \in X$, $v(x,b_j)$ is known and also that for $j = 1,\ldots,n$ a Markov policy $f_j \in F$ is known such that $v(x,b_j) = v(x,b_j,f_j)$ for all $x \in X$.
It is easy to verify that, if $f_k \in \bar{F}$ for $k = 1,\ldots,n$, then

$$6.16 \quad \text{(i)} \quad \ell(x,q) \geq \max_{1\leq k\leq n} \sum_{j=1}^{n} v(x,b_j,f_k)q(B_j) - \frac{\operatorname{span}(r)}{1-\beta} \frac{\frac{1}{2}\tilde{\Delta}}{1-\beta+\frac{1}{2}\beta\tilde{\Delta}}$$

$$\text{(ii)} \quad w(x,q) \leq \sum_{j=1}^{n} v(x,b_j)q(B_j) + \frac{\operatorname{span}(r)}{1-\beta} \frac{\frac{1}{2}\tilde{\Delta}}{1-\beta+\frac{1}{2}\tilde{\Delta}}$$

where $\tilde{\Delta} := \max_{1\leq j\leq n} \sup_{\theta\in B_j} \Delta(\theta,b_j)$ .

Hence we derived an upper and a lower bound for $v(x,q)$ involving only the points of discretization. Statements similar to 6.16 are possible with the other bounds considered in th. 6.1. Note that the difference in the bounds of 6.16 is positive , even if q is degenerate. If we assume more structure we may derive better bounds. Let $\theta$ be an interval on the real line: $\theta := [b_0,b_n]$ and let $b_0 < b_1 < \ldots < b_n$ be the points of discretization. Further assume that $\theta \to v(x,\theta,f_j)$ is *nondecreasing* for $x \in X$, $j = 1,\ldots,n$ (an example of this situation is considered in example 6.4). Then it is straightforward to verify that, for $(x,q) \in X \times W$:

$$6.17 \qquad \max_{1\leq k\leq n} \sum_{j=1}^{n} v(x,b_{j-1},f_k)q([b_{j-1},b_j)) \leq v(x,q) \leq \sum_{j=1}^{n} v(x,b_j)q([b_{j-1},b_j)).$$

Even if q is degenerate, the upper and lower bound in 6.17 differ at least:

$$\min_{1\leq j\leq n} \{v(x,b_j) - v(x,b_{j-1})\} .$$

Now we shall consider the discretization in advance. We only treat the situation where I is a singleton. We omit the dependence on $i \in I$ in the notations.

Theorem 6.8

Let I be a singleton, let $b_1,\ldots,b_n \in \Theta$ be the points of discretization and let $\{B_1,\ldots,B_n\}$ be a measurable partition of $\Theta$ such that $b_j \in B_j$, $j = 1,\ldots,n$. Let $q \in W$ and define $\varphi \in W$ such that $\varphi(\{b_j\}) = q(B_j)$, $j = 1,\ldots,n$. Then:

$$\sup_x |v(x,q) - v(x,\varphi)| \le \frac{\operatorname{span}(r)}{1-\beta} \sum_{j=1}^{n} \int_{B_j} q(d\theta) \frac{\frac{1}{2}\Delta(\theta,b_j)}{1-\beta+\frac{1}{2}\beta\Delta(\theta,b_j)} \le$$

$$\le \frac{\operatorname{span}(r)}{1-\beta} \frac{\frac{1}{2}\Delta}{1-\beta+\frac{1}{2}\beta\Delta}$$

where

6.18 $$\Delta := \sum_{j=1}^{n} \int_{B_j} q(d\theta)\Delta(\theta,b_j)$$

($\Delta(\theta,b_j)$ has been defined in 6.15; note that I is a singleton).

*Proof.*

Fix $\varepsilon > 0$. There is a $\pi \in \Pi_0$ such that for a fixed $x \in X$

$$v(x,q) - v(x,\varphi) \le \varepsilon + v(x,q,\pi) - v(x,\varphi,\pi) .$$

Hence

$$v(x,q) - v(x,\varphi) \le \varepsilon + \sum_{k=0}^{\infty} \beta^k \sum_{j=1}^{n} \{ \int_{B_i} q(d\theta) \mathbf{E}^{\pi}_{x,\theta}[r(X_k,A_k,Y_{k+1})] -$$

$$-\mathbf{E}^{\pi}_{x,b_j}[r(X_k,A_k,Y_{k+1})]q(B_i)\} .$$

Let

$$f(y_1,\ldots,y_{k+1}) := \int \pi_0(da_0|x) \int P(dx_1|x,a_0,y_1)\ldots$$

$$\cdot \int P(dx_k|x_{k-1},a_{k-1},y_k) \int \pi_k(da_k|x,a_0,y_1,..,y_k,x_k) r(x_k,a_k,y_{k+1}) .$$

It is easy to verify that $f(y_1,\ldots,y_{k+1})$ is a version of

$$\mathbf{E}^{\pi}_{x,\theta}[r(X_k,A_k,Y_{k+1}) \mid Y_1 = y_1,\ldots,Y_{k+1} = y_{k+1}] \quad \text{for all } \theta \in \Theta .$$

Note that $m \le f \le M$ where $M := \sup_{x,a,y} r(x,a,y)$ and $m := \inf_{x,a,y} r(x,a,y)$.
Then we have

$$v(x,q) - v(x,\varphi) \le \varepsilon + \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1})$$

$$\cdot \sum_{i=1}^{n} \int_{B_i} q(d\theta)\{\prod_{j=1}^{k+1} p(y_j \mid \theta) - \prod_{j=1}^{k+1} p(y_j \mid b_i)\} f(y_1,\dots,y_{k+1}) \le$$

$$\le \varepsilon + \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1}) \sum_{i=1}^{n} \int_{B_i} \{\prod_{j=1}^{k+1} p(y_j|\theta) - \prod_{j=1}^{k+1} p(y_j \mid b_i)\}^{+} M -$$

$$- \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1}) \sum_{i=1}^{n} \int_{B_i} q(d\theta)\{\prod_{j=1}^{k+1} p(y_j|\theta) -$$

$$- \prod_{j=1}^{k+1} p(y_j \mid b_i)\}^{-} m = \varepsilon + \tfrac{1}{2} \operatorname{span}(r) \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1})$$

$$\cdot \sum_{i=1}^{n} \int_{B_i} q(d\theta) \left| \prod_{j=1}^{k+1} p(y_j|\theta) - \prod_{j=1}^{k+1} p(y_j|b_i) \right| .$$

(Here we use

$$\int \dots \int \nu(dy_1)\dots\nu(dy_{k+1}) \prod_{j=1}^{k+1} p(y_j|\theta) = 1, \text{ for all } \theta \in \Theta) .$$

Likewise there is a $\pi^* \in \Pi_0$ such that $v(x,\varphi) - v(x,q) \le \varepsilon + v(x,\varphi,\pi^*) - v(x,q,\pi^*)$. Therefore we have:

$$(*) \qquad \sup_x |v(x,q) - v(x,\varphi)| \le \tfrac{1}{2} \operatorname{span}(r) \sum_{k=0}^{\infty} \beta^k \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1})$$

$$\cdot \sum_{i=1}^{n} \int_{B_i} q(d\theta) \left| \prod_{j=1}^{k+1} p(y_j|\theta) - \prod_{j=1}^{k+1} p(y_j|b_i) \right| .$$

Let

$$\alpha_k := \tfrac{1}{2} \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1}) \sum_{i=1}^{n} \int_{B_i} q(d\theta) \left| \prod_{j=1}^{k+1} p(y_j|\theta) - \prod_{j=1}^{k+1} p(y_j|b_i) \right| .$$

Further let $c_1,\dots,c_{k+1},d_1,..,d_{k+1}$ be nonnegative numbers. The following inequality is immediate

$$(**) \qquad \min\{\prod_{j=1}^{k+1} c_j, \prod_{j=1}^{k+1} d_j\} \ge \prod_{j=1}^{k+1} \min\{c_j,d_j\} .$$

It is easy to verify that

$$\alpha_k = 1 - \int \dots \int \nu(dy_1)\dots\nu(dy_{k+1}) \sum_{i=1}^{n} \int_{B_i} q(d\theta)\min\{\prod_{j=1}^{k+1} p(y_j|\theta),\ \prod_{j=1}^{k+1} p(y_j|b_i)\}.$$

Using (**) we find, after changing the order of integration

$$\alpha_k \le 1 - \sum_{i=1}^{n} \int_{B_i} q(d\theta)\Big[\int \nu(dy)\min\{p(y|\theta),\ p(y|b_i)\}\Big]^{k+1} .$$

Hence (*) becomes:

$$\sup_x \left|v(x,q) - v(x,\varphi)\right| \le \operatorname{span}(r)\Big\{\frac{1}{1-\beta} - \sum_{i=1}^{n} \int_{B_i} q(d\theta)\frac{F(\theta,b_i)}{1-\beta F(\theta,b_i)}\Big\} ,$$

where $F(\theta,b_i) := \int \nu(dy)\min\{p(y|\theta),p(y|b_i)\}$. Note that $\sum_{i=1}^{n} \int_{B_i} q(d\theta) = 1$ and $F(\theta,b_i) = 1 - \frac{1}{2}\Delta(\theta,b_i)$.
Hence we get the first inequality:

$$\sup_x \left|v(x,q) - v(x,\varphi)\right| \le \operatorname{span}(r) \sum_{i=1}^{n} \int_{B_i} q(d\theta)\frac{\frac{1}{2}\Delta(\theta,b_i)}{(1-\beta)\{1-\beta+\frac{1}{2}\beta\Delta(\theta,b_i)\}} .$$

Since the function $s \to \frac{\frac{1}{2}s}{1-\beta+\frac{1}{2}\beta s}$ is concave on $[0,1]$ we find using Jensen's inequality:

$$\sup_x \left|v(x,q) - v(x,\varphi)\right| \le \frac{\operatorname{span}(r)}{1-\beta}\,\frac{\frac{1}{2}\Delta}{1-\beta+\frac{1}{2}\beta\Delta} ,$$

which proves the second inequality. □

Remarks.

(i) We can also use the proof of th. 6.8 to compare the original model with a slightly different Bayesian control model. Let $\tilde{\theta} := \{1,2,\dots,n\}$ be the parameter space of the modified model and let $\tilde{p}(\cdot|\theta)$ be a probability density with respect to the measure $\nu$, for $\theta \in \tilde{\theta}$. Further let $\varphi(\{j\}) := q(B_j)$, $j = 1,\dots,n$ be the prior distribution on $\tilde{\theta}$ and $\tilde{v}(x,\varphi)$ be the value of the modified model. All other specifications of the modified model are as in the original model. Then the statement of th. 6.8 remains valid with $v(x,\varphi)$ replaced by $\tilde{v}(x,\varphi)$ and with $\Delta(\theta,b_j)$ replaced by $\int \nu(dy)\left|p(y|\theta) - \tilde{p}(y|j)\right|$.

(ii) Note that, in case I is a singleton, th. 6.7 is a consequence of th. 6.8. To verify this let $q \in W$ be degenerate at $\theta$ and let the partition of 6.16 consist of $\theta$ only with discretization point $\theta$. Then apply the first inequality of th. 6.8.

Corollary 6.9

Let I be a singleton and let $\tilde{v}(x,q)$ be the value of the model with known transition law, given by $\nu(dy)P(dx'|x,a,y)p(y,q)$. Then we have

$$\sup_x |v(x,q) - \tilde{v}(x,q)| \leq \frac{\operatorname{span}(r)}{1-\beta} \frac{\frac{1}{2}\Delta^*}{1-\beta+\frac{1}{2}\beta\Delta^*}$$

where

$$\Delta^* := \int q(d\theta) \int \nu(dy) |p(y|\theta) - p(y,q)| .$$

Note that this is an example of the situation considered in remark (i) above, if we set $\tilde{\theta} := \{1\}$ and $\tilde{p}(\cdot|1) := p(\cdot,q)$.

If there is a $b \in \theta$ such that $p(\cdot|b) = p(\cdot,q)$ for some $q \in W$, then corollary 6.9 is a special case of th. 6.8.

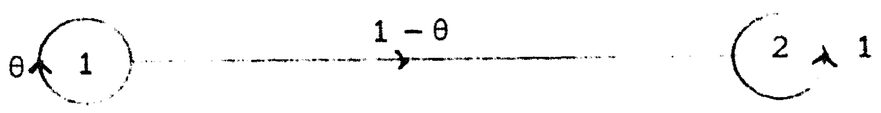
In practice one often considers the value $\tilde{v}(x,q)$, defined in corollary 6.9 as an approximation to $v(x,q)$. This is justified by the following interpretation. In the Bayesian approach the prior distribution q is determined, using data from the past. Then the Bayes estimation of the density is computed: $p(y,q) = \int q(d\theta)p(y|\theta)$, for $y \in Y$, and finally this density is considered to be the true one.

We conclude this section with some examples and remarks.

Example 6.1

The first bound in th. 6.8 is tight.
Consider the model with only one action in each state, and with $X := \{1,2\}$, $Y := \{0,1\}$, $A := \{1\}$, $\theta := \{0,1\}$

$$P(1|1,1,1) := 1,\ P(2|1,1,0) := 1,\ P(2|2,1,0) := P(2|2,1,1) := 1$$
$$p(1|\theta) := \theta,\ r(1,1,1) := 1,\ r(1,1,0) := r(2,1,0) := r(2,1,1) := 0.$$



Let the only point of discretization be 0, hence $B_1 = \theta$. Let $q \in W$ be defined by $q(\{0\}) = q(\{1\}) = \frac{1}{2}$. It is easy to verify that $v(1,q) = \frac{1}{2}\frac{1}{1-\beta}$. The prior $\varphi$ defined in th. 6.8 has all mass in the point 0. Hence $v(1,\varphi) = 0$ and $v(1,q) - v(1,\varphi) = \frac{1}{2}\frac{1}{1-\beta}$. Now we consider the first bound of th. 6.8. We have

$$\int \nu(dy)\,|p(y|\theta) - p(y|0)| = |\theta - 0| + |1 - \theta - 1| = 2\theta,\ \theta \in \Theta .$$

Hence the bound becomes $\frac{1}{2}\frac{1}{1-\beta}$.

Example 6.2

The bound of th. 6.7 is tight. Consider the example above and let $\theta := 0$ and $\tilde{\theta} := 1$. Then $|v(1,\theta) - v(1,\tilde{\theta})| = \frac{1}{1-\beta}$ and

$$\int \nu(dy)\,|p(y|\theta) - p(y|\tilde{\theta})| = 2|\theta - \tilde{\theta}| = 2 .$$

Hence the bound is $\frac{1}{1-\beta}$ also.

Example 6.3

The bounds of th. 6.8 behave badly if $\beta$ tends to 1. Consider the model of example 6.1 and modify it as follows: $\Theta := [0,1]$, $r(1,1,1) := r(1,1,0) := 1$. Let the only point of discretization be $\frac{1}{2}$ and let $q \in W$ be homogeneous. Hence $v(1,\theta) = \frac{1}{1-\beta\theta}$ and so $v(1,q) = -\frac{1}{\beta}\log(1-\beta)$ and $v(1,\varphi) = \frac{1}{1-\frac{1}{2}\beta}$. The first bound of th. 6.8 becomes:

$$\frac{1}{1-\beta}\int_0^1 \frac{|\theta - \frac{1}{2}|}{1-\beta+\beta|\theta-\frac{1}{2}|}\,d\theta = \frac{1}{1-\beta}\{\frac{1}{\beta} - \frac{2(1-\beta)}{\beta^2}\log(\frac{1-\frac{\beta}{2}}{1-\beta})\} = O(\frac{1}{1-\beta}) .$$

Note that $|v(1,q) - v(1,\varphi)| = O(\frac{1}{1-\beta})$.

In the next example we consider a situation where $\theta \to v(x,\theta,f_i)$ is monotone (cf. 6.17).

Example 6.4

Consider the inventory model: model 5B where the demand is exponentially distributed:

$$p(y|\theta) = \theta e^{-\theta y},\ \theta \in [a,b],\ 0 < a < b < \infty .$$

For each $\theta \in [a,b]$ the optimal strategy is characterized by a number $s_\theta$, $s_\theta \in [a,b]$, such that, at time zero the inventory level is brought to $s_\theta$ and afterwards at each stage the demand is supplied. For a fixed $s \in [a,b]$ we determine the total expected costs corresponding to the strategy as described above with s instead of $s_\theta$.

$$v(x,\theta,s) = hx^+ + px^- + k(s-x) + \sum_{n=1}^{\infty} \beta^n \mathbb{E}_\theta [h(s-Y_n)^+ + p(s-Y_n)^- + kY_n] \ .$$

It is easy to verify that

$$\mathbb{E}_\theta [h(s-Y_n)^+ + p(s-Y_n)^- + kY_n] = (h+p)(s + \frac{1}{\theta} e^{-\theta s} - \frac{1}{\theta}) - ps + (p+k)\frac{1}{\theta} \ .$$

If $k > h$, then this function is decreasing in $\theta$.
Hence for each point of discretization and each $x \in X$, $\theta \to v(x,\theta,f_i)$ is decreasing. (Remember that we considered costs instead of rewards in this example.)

Remarks.

(i) If $\theta \subset \mathbb{R}$ and for some $f \in F$, $\theta \to v(x,\theta,f)$ is *convex* then

$$\int q(d\theta) v(x,\theta,f) \geq v(x, \int q(d\theta)\theta, f) \ ,$$

by Jensen's inequality, and if $\theta \to v(x,\theta)$ is *concave* then $w(x,q) \leq v(x, \int q(d\theta)\theta)$. These properties are sometimes useful in approximating upper and lower bounds.

(ii) In [Whitt (1976)] discretizations of the state and action spaces are considered for discounted dynamic programs. If we apply Whitt's approach here we have to discretize the set of posterior distributions W, i.e. we have to fix a finite measurable partition $B_1,\ldots,B_n$ of W and in each set $B_i$ a representant $b_i$. Then the original model is compared with the model with a perturbed transition law, which causes the process to visit the points $b_i$, $i \in \{1,\ldots,n\}$ only. However, if $(\tilde{X}_n,\tilde{Q}_n)$ is the state at time n of this new process, then $\tilde{Q}_n$ is not the posterior distribution of Z, in general.
So th. 6.5 is not valid anymore.

# 7. COMPUTATIONAL ASPECTS AND EXAMPLES

In this chapter we consider algorithms for the computation of the value function (cf. 2.12) in two special cases of the Bayesian control model. In section 7.1 we consider the model where the index set I is a singleton. Here we also consider the rate of convergence of the algorithm. In section 7.2 we consider Bayesian control models where assumption 6.6 holds and where in addition the set $L_1$ is a singleton (cf. 6.6). Finally, in section 7.3, we study some examples of the models considered in sections 7.1 and 7.2, and we illustrate the quality of the algorithms by numerical data. The algorithms are based on the approximations given in th. 6.5. Throughout this chapter we assume that X, Y, A and $\Theta$ are finite sets. (For notational convenience we write $q(\theta)$ instead of $q(\{\theta\})$.)

## 7.1 *Algorithm for models where I is a singleton*

In this section we assume that the index set I is a singleton.
We consider an algorithm, based on th. 6.5, to approximate $v(x,q)$ for all $x \in X$ and one fixed prior distribution $q \in W$. The accuracy of the approximation has to be given in advance. In section 6.1 we already considered the set of all possible posterior distributions after n transitions

$$7.1 \qquad W_n(q) := \{\chi_q(y_1,\ldots,y_n) \mid y_1,\ldots,y_n \in Y\},\ q \in W \text{ (cf. 6.10 (i))} .$$

Since I is a singleton we omit the subscript $i \in I$ in this section. We first give the algorithm and afterwards we discuss each of its steps. Let $\Delta > 0$ be given, and let $\tilde{v}(x,q)$ be the approximation to $v(x,q)$. If $\max_x |v(x,q) - \tilde{v}(x,q)| \leq \Delta$ for a fixed $q \in W$ then we say that the accuracy of the approximation is (at least) $\Delta$.
In the algorithms the symbol ":=" denotes an assignment instead of a definition.

*Algorithm* 1

part 1: *parameter influence*

(a) For all $\theta \in \Theta$ and $x \in X$ determine $v(x,\theta)$ and an optimizing $f_\theta \in F$ (i.e. $v(x,\theta) = v(x,\theta,f_\theta)$ for $\theta \in \Theta$). Let $\bar{F} := \{f_\theta,\ \theta \in \Theta\}$ (cf. 6.2).

(b) For all $\theta \in \Theta$ and $f \in \bar{F}$ determine $\varphi_\infty(\theta,f) = \max_x \{v(x,\theta) - v(x,\theta,f)\}$.

part 2: *horizon determination*

(c) Set $n := n_0$ (a lower estimate of the horizon, e.g. $n_0 := 0$).

(d) Compute (cf. 6.14):

$$\delta := E(q,\varepsilon_\infty,n) = \tfrac{1}{2} \sum_{y_1,\ldots,y_n} \min_{f\in\bar{F}} \sum_{\theta} q(\theta) \prod_{j=1}^{n} p(y_j|\theta)\varphi_\infty(\theta,f) \ .$$

(e) If $\beta^n\delta \leq \Delta$ then go to (f), otherwise set $n := n+1$ and go to (d).

part 3: *backward induction*

(f) For all $q' \in W_n(q)$ set $v_n(x,q') := \tfrac{1}{2}\{w(x,q') + \ell(x,q')\}$, $x \in X$.

(g) Set $k := n-1$.

(h) For all $q' \in W_k(q)$ compute $p(y,q')$ and $T_y(q')$ for all $y \in Y$, and then, for $x \in X$:

$$v_k(x,q') := \max_{a\in D(x)} \sum_{y} p(y,q')\{r(x,a,y) + \beta \sum_{x'} P(\{x'\}|x,a,y)v_{k+1}(x',T_y(q'))\}.$$

(i) If $k > 0$ then set $k := k-1$ and go to (h).
Otherwise: stop.

At the end of the algorithm the values $\tilde{v}(x,q) := v_0(x,q)$, $x \in X$ have been computed and it follows from th. 6.5 that the accuracy is at least $\Delta$.
We proceed with a discussion of each of the steps of the algorithm.

*Remarks on algorithm 1*

(i) The computations of step (a) can be carried out by a standard method, such as the *policy iteration algorithm* or the *method of successive approximations* with the *MacQueen extrapolation* (cf. [Ross (1970), section 6.8], [MacQueen (1966)]). More sophisticated methods can be found in [Van Nunen (1976), section 7.3] and in [Hastings and Van Nunen (1977)]. Note that $\bar{F}$ in step (a) satisfies 6.2.
It often occurs that, if the differences between the parameter values are small then also the differences in the value function are small (cf. th. 6.7). Hence if $f_\theta \in F$ is optimal, if $\theta \in \Theta$ is the true parameter and if $\tilde{\theta} \in \Theta$ is near to $\theta$, then it is wise to start the policy iteration for the parameter value $\tilde{\theta}$ with the policy $f_\theta$. And likewise we recommend to start the successive approximations for $\tilde{\theta}$ with the scrap function $v(\cdot,\theta)$.

(ii) In step (b) we have to determine $v(x,\theta,f)$ for all $x \in X$, $\theta \in \Theta$ and $f \in \bar{F}$. If $N := \#\Theta$ then it requires the solution of at most $N(N-1)$ systems of linear equations, in fact we only have to solve them all if we found for each $\theta \in \Theta$ a different optimal policy. Instead of the function $\varphi_\infty(\theta,f)$ we may also use the function

$\varphi_1(\theta,f) = \max_x \{v(x,\theta) - (L_f v)(x,\theta)\}$, if we replace $\varepsilon_\infty(\cdot)$ by $\varepsilon_1(\cdot)$ in step (d).

It is easy to verify that the computation of $(L_f v)(x,\theta)$ for $x \in X$, $\theta \in \Theta$ and $f \in \bar{F}$ requires less effort than the computation of $v(x,\theta,f)$ for $x \in X$, $\theta \in \Theta$ and $f \in \bar{F}$. However we have to do more work in part 2 of the algorithm in this case (cf. th. 6.1).

(iii) Let $Y = \{0,1,\ldots,m\}$. We may compute $E(q,\varepsilon_\infty,n)$ in the following way:

$$7.2 \qquad E(q,\varepsilon_\infty,n) = \tfrac{1}{2} \sum_{k_0,\ldots,k_m} \frac{n!}{\prod_{j=0}^{m} k_j!} \min_{f\in\bar{F}} \sum_{\theta} q(\theta) \prod_{j=0}^{m} p(j|\theta)^{k_j} \varphi_\infty(\theta,f)$$

with summation over all $k_0,k_1,\ldots,k_m \in \mathbb{N}$ such that $\sum_{j=0}^{m} k_j = n$.

Note that we have to sum over $\binom{m+n}{n}$ terms here, so the amount of work to compute $\delta$ in this way is very large if n is large. Therefore we suggest another approach in case $\{p(\cdot|\theta),\ \theta \in \Theta\}$ is an *exponential family* of the following form

$$7.3 \qquad p(y|\theta) = a(\theta)b(y)\exp\{c(\theta)y\},\ \theta \in \Theta,\ y \in Y = \{0,1,\ldots,m\}$$

where a and b are nonnegative functions such that

$$\sum_{Y} a(\theta)b(y)\exp\{c(\theta)y\} = 1, \quad \text{for all } \theta \in \Theta .$$

In this case the posterior distribution of $q \in W$ after n observations $y_1,\ldots,y_n$ becomes:

$$7.4 \qquad \frac{a(\theta)^n \exp\{c(\theta)\sum_{j=1}^{n} y_j\} q(\theta)}{\sum_{\theta'} a(\theta')^n \exp\{c(\theta')\sum_{j=1}^{n} y_j\} q(\theta')}$$

(provided that the denominator does not vanish).

Hence the number of different posterior distributions is:

$$\#W_n(q) = \#\{\sum_{j=1}^{n} y_j \mid y_j \in Y,\ j = 1,\ldots,n\} = nm + 1 .$$

(This number is small compared to $\binom{m+n}{n}$, the number of terms in 7.2.) Note that there is a one-to-one correspondence between $\Sigma_{j=1}^{n} Y_j$ and $Q_n$, due to relation 7.4. Instead of computing $E(q,\varepsilon_\infty,n)$ as proposed in step (d) we approximate this quantity in the following way. Note that here

$$E(q,\varepsilon_\infty,n) = \mathbb{E}_q[\varepsilon_\infty(Q_n)] = \sum_{q' \in W_n(q)} \mathbb{P}_q[Q_n = q']\varepsilon_\infty(q') \ .$$

It is relatively easy to compute $\varepsilon_\infty(q')$ for each $q' \in W_n(q)$. Instead of computing $\mathbb{P}_q[Q_n = q']$ directly, we approximate this probability by the normal probability in the following way. Let $q' \in W_n(q)$ correspond to all sequences $y_1,\ldots,y_n \in Y$ with $\Sigma_{j=1}^{n} y_j = s$, $0 \le s \le nm$. Further let

$$7.5 \qquad \mu_\theta := \sum_{j=0}^{m} jp(j|\theta) \quad \text{and} \quad \sigma_\theta^2 := \sum_{j=0}^{m} (j - \mu_\theta)^2 p(j|\theta) \ .$$

Then we have

$$\mathbb{P}_q[Q_n = q'] = \sum_{\theta} q(\theta)\mathbb{P}_\theta[\sum_{j=1}^{n} Y_j = s]$$

and therefore

$$7.6 \qquad \mathbb{P}_q[Q_n = q'] \approx \sum_{\theta} q(\theta)\{\Phi(\frac{s + \frac{1}{2} - n\mu_\theta}{\sigma_\theta\sqrt{n}}) - \Phi(\frac{s - \frac{1}{2} - n\mu_\theta}{\sigma_\theta\sqrt{n}})\}$$

(where $\Phi$ is the standard normal distribution function).
Note that $\mu_\theta$ and $\sigma_\theta$ can be computed in advance, also in part 1.
Since $\# W_n(q')$ is relatively small it is easy to approximate $E(q,\varepsilon_\infty,n)$ in this way.

(iv) Besides the convergence due to the discount factor $\beta$, we also use in step (e) the convergence of the posterior distributions. In fact we might replace the stop criterion by "$\delta \le \Delta$" without loosing convergence of the algorithm (cf. th. 6.5).
Instead of an absolute stop criterion we might use a relative criterion. For instance we could use the inequality

$$\beta^n\delta\{1 + \max_x|\ell(x,q)|\}^{-1} \le \Delta$$

instead of $\beta^n\delta \le \Delta$ , in step (e).

(v) The backward induction in part 3 requires the following storage capacity for numbers

$\{\#W_n(q) + \#W_{n-1}(q)\} \# X = \{(2n - 1)m + 2\} \# X$

if $Y = \{0,1,\ldots,m\}$, $\{p(\cdot|\theta), \theta \in \Theta\}$ is an exponential family of the form 7.3, and n is the horizon determined in part 2.

We note that the work in part 1 has to be carried out only once, while we have to perform part 2 and part 3 for each prior distribution $q \in W$ for which we want to approximate $v(x,q)$ , $x \in X$.

We continue with the discussion of a simple modification of the algorithm to determine in part 3 upper and lower bounds on $v(x',q')$, $q' \in \bigcup_{k=1,\ldots,n-1} W_k(q)$ . These bounds shall allow us to exclude some sub-optimal actions, during the backward induction procedure. To derive these bounds we proceed as follows. Let n be the horizon determined in part 2 and remember that

$$v_n(x',q') = \tfrac{1}{2}\{w(x',q') + \ell(x',q')\} \quad \text{for} \quad q' \in W_n(q),\ x' \in X .$$

According to corollary 6.6 we have

$$7.7 \qquad |v(x',q') - (U^{n-k}v_n)(x',q')| \leq \beta^{n-k}\mathbb{E}_{q'}[\varepsilon_\infty(Q_{n-k})]$$

for $q' \in W_k(q)$, $k \in \{0,1,\ldots,n-1\}$ and $x' \in X$ .

Further note that, according to the Markov property of $\{Q_k, k \in \mathbb{N}\}$ (cf. th. 5.1):

$$7.8 \qquad \mathbb{E}_{q'}[\varepsilon_\infty(Q_{n-k})] = \sum_y \mathbb{E}_{T_y(q')}[\varepsilon_\infty(Q_{n-k-1})]p(y,q')$$

for $q' \in W_k(q)$ , $k \in \{0,1,\ldots,n-1\}$ and $x' \in X$ .

Hence the values $\mathbb{E}_{q'}[\varepsilon_\infty(Q_{n-k})]$ for $q' \in W_k(q)$ are upper and lower bounds on $v(x',q')$, for $x' \in X$.

These values are easy to compute by 7.9.

7.9 (i) Let $\gamma_n(q') := \varepsilon_\infty(q')$ for $q' \in W_n(q)$ .

(ii) For $k = n-1, n-2,\ldots,1$ compute

$$\gamma_k(q') := \sum_y p(y,q')\gamma_{k+1}(T_y(q')), \quad \text{for } q' \in W_k(q) .$$

Note that the computations of 7.9(ii) can be incorporated in step (h) of the algorithm.

If we use the normal approximation, as suggested in remark (iii), then we loose our exact accuracy. However if we incorporate the computations of 7.9, then we have exact bounds after the execution of step 3.

It is obvious that an action $a \in D(x)$ is *sub-optimal* in state $(x,q')$ $q' \in W_k(q')$, $x \in X$ and $k \in \{0,1,\ldots,n-1\}$ if

7.10
$$\sum_y p(y,q')[r(x,a,y) + \beta \sum_{x'} P(\{x'\}|x,a,y)\{v_{k+1}(x',T_y(q')) + \gamma_{k+1}(T_y(q'))] \leq$$
$$\leq (U(v_{k+1} - \gamma_{k+1}))(x,q') \ .$$

We conclude this section with a qualitative statement concerning the rate of convergence of the algorithm.

We start with some preparations. Remember that X, Y, A and $\Theta$ are finite sets.

7.11 A *maximum likelihood estimator* $M_n$ of the parameter based on the observations $Y_1,Y_2,\ldots,Y_n$ is a $\Theta$-valued function of $Y_1,\ldots,Y_n$ such that

$$\prod_{j=1}^{n} p(Y_j|M_n) \geq \prod_{j=1}^{n} p(Y_j|\theta) \quad \text{on } \Omega, \quad \text{for all } \theta \in \Theta \ .$$

Lemma 7.1

There are numbers k and a, $k,a > 0$ such that for all $\theta \in \Theta$

$$\mathbb{P}_\theta[M_n \neq \theta] \leq k \exp\{-an\} \ , \quad n \in \mathbb{N}^* .$$

Proof.

Define on $\Omega$:

$$Z_j(\theta,\varphi) := \log\{\frac{p(Y_j|\theta)}{p(Y_j|\varphi)}\} \ , \quad j \in \mathbb{N}^* , \quad \theta,\varphi \in \Theta .$$

(let $\log 0 = -\infty$, $\log \infty = \infty$ and let $0 \cdot \infty = 0$).

Note that

(*)
$$\{M_n \neq \theta\} \subset \{\max_{\varphi \neq \theta} \prod_{j=1}^{n} p(Y_j|\varphi) \geq \prod_{j=1}^{n} p(Y_j|\theta)\} =$$

$$= \{\min_{\varphi\neq\theta} \prod_{j=1}^{n} \frac{p(Y_j|\theta)}{p(Y_j|\varphi)} \leq 1\} = \bigcup_{\varphi\neq\theta} \{\sum_{j=1}^{n} Z_j(\theta,\varphi) \leq 0\} .$$

By a Chebyshev-type inequality we have for all $t \leq 0$ and $\theta \in \Theta$:

$$(**) \qquad \mathbb{P}_\theta[\sum_{j=1}^{n} Z_j(\theta,\varphi) \leq 0] \leq \mathbb{E}_\theta[\exp\{t \sum_{j=1}^{n} Z_j(\theta,\varphi)\}] = \mathbb{E}_\theta[\exp\{tZ_j(\theta,\varphi)\}]^n .$$

Note that $f_{\theta,\varphi}(t) := \mathbb{E}_\theta[\exp\{tZ_j(\theta,\varphi)\}] = \sum_y \{\frac{p(y|\varphi)}{p(y|\theta)}\}^{-t} p(y|\theta)$ is finite for $t \leq 0$, $\theta,\varphi \in \Theta$ and independent of $j \in \mathbb{N}^*$.

Further note that $f_{\theta,\varphi}(0) = f_{\theta,\varphi}(-1) = 1$, for $\theta,\varphi \in \Theta$.

Since the function $t \to x^{-t}$ $(x > 0)$ is strictly convex, except when $x = 1$, we conclude that for each pair $\theta,\varphi \in \Theta$ there is a number $t$, $-1 < t < 0$ such that $f_{\theta,\varphi}(t) < 1$, except when $\frac{p(y|\theta)}{p(y|\varphi)} = 1$ for all $y$ with $p(y|\theta) > 0$. However, if $p(y|\theta) = p(y|\varphi)$ for all $y \in Y$ then, by the separation assumption (cf. 2.1), we have $\theta = \varphi$.

Hence there is for each pair $\theta,\varphi \in \Theta$, $\varphi \neq \theta$ a number $t_{\theta,\varphi}$ such that $-1 \leq t_{\theta,\varphi} \leq 0$ and $f_{\theta,\varphi}(t_{\theta,\varphi}) < 1$.

Hence by (*) and (**) we have:

$$\mathbb{P}_\theta[M_n \neq \theta] \leq \sum_{\varphi\neq\theta} \mathbb{P}_\theta[\sum_{j=1}^{n} Z_j(\theta,\varphi) \leq 0] \leq \sum_{\varphi\neq\theta} \{f_{\theta,\varphi}(t_{\theta,\varphi})\}^n .$$

Finally let $m := \max_{\theta\neq\varphi} f_{\theta,\varphi}(t_{\theta,\varphi})$, $a := -\log m$ and $k := \#\Theta - 1$.

Then we have

$$\mathbb{P}_\theta[M_n \neq \theta] \leq k \exp\{-an\} , \quad \text{for all } \theta \in \Theta . \qquad \square$$

The statement of lemma 7.1 is contained in th. 5.3.1 given in [Zacks (1971)], with a proof that is incorrect but easy to repair. Since our situation is less general our proof is easier. However, the idea has been borrowed from Zacks.

In th. 7.4 we use the maximum likelihood estimator to choose a Markov policy $f_{M_n} \in \bar{F}$ such that on $\Omega$:

$$\sum_\theta \max_x \{v(x,\theta) - v(x,\theta,f_{M_n}\}Q_n(\theta) \geq \min_{f\in\bar{F}} \sum_\theta \varphi_\infty(\theta,f)Q_n(\theta) .$$

This bound is used implicitly to show that one has *exponential convergence* in part 2 of the algorithm.

Theorem 7.2

Let $f_\theta \in F$ be an optimal policy for $\theta \in \Theta$ and let $\bar{F} := \{f_\theta, \theta \in \Theta\}$. There are positive numbers k and a such that

$$\mathbb{E}_q[\min_{f\in\bar{F}} \sum_\theta \max_x \{v(x,\theta) - v(x,\theta,f)\}Q_n(\theta)] \le k \exp\{-an\} .$$

Proof.

For $f \in \bar{F}$ define $A_f := \{\theta \in \Theta \mid v(x,\theta) = v(x,\theta,f) \quad \text{for all } x \in X\}$ and $B_f := \Theta \backslash A_f$ .

Further let $\Delta := \max_{f\in\bar{F}} \max_\theta \max_x \{v(x,\theta) - v(x,\theta,f)\}$ .

Note that

$$\mathbb{E}_q[\min_{f\in\bar{F}} \sum_\theta \max_x \{v(x,\theta) - v(x,\theta,f)\}Q_n(\theta)] \le \Delta \cdot \mathbb{E}_q[\min_{f\in\bar{F}} Q_n(B_f)] =$$

$$= \Delta\, \mathbb{E}_q[\min_{f\in\bar{F}} \mathbb{P}_q[Z \in B_f \mid Y_1,\ldots,Y_n]] .$$

Note further that:

$$\min_{f\in\bar{F}} \mathbb{P}_q[Z \in B_f \mid Y_1,\ldots,Y_n] \le \sum_\theta 1_{\{M_n=\theta\}} \mathbb{P}_q[Z \in B_{f_\theta} \mid Y_1,\ldots,Y_n]$$

and, since $M_n$ is a function of $Y_1,\ldots,Y_n$ (cf. 7.11), we obtain:

$$\mathbb{E}_q[\min_{f\in\bar{F}} \sum_\theta \max_x \{v(x,\theta) - v(x,\theta,f)\}Q_n(\theta)] \le \Delta \cdot \mathbb{P}_q[Z \in B_{f_{M_n}}] =$$

$$= \Delta \sum_\theta q(\theta)\, \mathbb{P}_\theta[\theta \not\in A_{f_{M_n}}] .$$

Since $\theta \in A_{f_\theta}$ for all $\theta \in \Theta$, we conclude that $\theta \not\in A_{f_\varphi}$ implies $\theta \ne \varphi$. Hence $\{\theta \not\in A_{f_{M_n}}\} \subset^\theta \{\theta \ne M_n\}$ .

Finally the desired result follows from lemma 7.1. □

## 7.2 *Algorithm for models with known transition law except for one state*

In this section we assume that assumption 6.6 holds and that the set $L_1$ is a singleton (cf. 6.6). Hence the transition law is known except for one state.

Throughout this section $X = \{0,1,\ldots,N\}$ and $L_1 = \{0\}$. Hence $L_2 = \{1,2,\ldots,N\}$. The algorithm is also based on th. 6.5 and it consists of three parts again. Each part is a modification of the corresponding part of algorithm 1.
We start with a discussion of these parts and afterwards we describe the algorithm.
We start with part 1.
If we want to compute $\mathbb{E}^{f}_{x,q}[\Sigma_{n=0}^{\sigma-1} \beta^n r(X_n,A_n,Y_{n+1}) + \beta^{\sigma} b(X_\sigma, Q_\sigma)]$ (cf. 6.7) for some $f \in F$ then we only have to specify the actions in the states of $L_2$. We shall use this property and therefore we introduce some useful notations:

7.12 (i) $F^* := \{f : L_2 \to A \mid f(x) \in D(x)\}$ .

(ii) $c_f(x) := \mathbb{E}^{f}_{x}[\Sigma_{n=0}^{\sigma-1} \beta^n r(X_n,A_n,Y_{n+1})]$ , $f \in F^*$, $x \in L_2$ .

(iii) $d_f(x) := \mathbb{E}^{f}_{x}[\beta^{\sigma}]$ , $f \in F^*$, $x \in L_2$ .

(iv) $g(x,e) := \max_{f \in F^*} \{c_f(x) + d_f(x)e\}$ , $e \in \mathbb{R}$ , $x \in L_2$ .

As already noted in section 6.1 (cf. 6.9), the determination of $(U_\sigma b)(x,q)$ $x \in L_2$ is an ordinary dynamic programming problem. To see this, extend the state space to $\{-1,0,1,\ldots,N\}$ and let $P(\{-1\}\mid 0,a,y) := P(\{-1\}\mid -1,a,y) := 1$ for all $a \in A$ and $y \in Y$. Further we define for this model $r(0,a,y) := b(0,q)$ and $r(-1,a,y) := 0$ for $a \in A$ and $y \in Y$, where $q \in W$ is fixed. It is easy to verify that the value function of this model in $x \in L_2$ equals $(U_\sigma b)(x,q)$ of the original model. Therefore we have

7.13 $$(U_\sigma b)(x,q) = \max_{f \in F^*} \{c_f(x) + d_f(x)b(0,q)\} = g(x,b(0,q)) \ , \quad x \in L_2 \ .$$

Let two numbers $\underline{e}$ and $\bar{e}$ be fixed such that: $\underline{e} \le v(0,q) \le \bar{e}$ for all $q \in W$. Note that, if $m \le r \le M$, $m,M \in \mathbb{R}$ then $\underline{e} := m(1-\beta)^{-1}$ and $\bar{e} := M(1-\beta)^{-1}$ will do.
It is easy to compute $g(x,e)$ for all $\underline{e} \le e \le \bar{e}$ and $x \in L_2$. This is due to the following properties.

**Lemma 7.3**

For each $x \in X$ the function $e \to g(x,e)$ is nondecreasing, convex and piecewise linear.

The proof of this lemma is trivial.

It is a consequence of lemma 7.3 that for each $f \in F^*$ the set $\{e \mid \underline{e} \leq e \leq \bar{e},\ g(x,e) = c_f(x) + d_f(x)e\}$ is a closed interval.

For $x \in L_2$ and $a \in D(x)$ we define

7.14 (i) $$\tilde{r}(x,a) := \sum_{i \in I} 1_{K_i}(x,a) \sum_{y} p_i(y|\theta_i) r(x,a,y) .$$

(ii) $$\tilde{P}(x'|x,a) := \sum_{i \in I} 1_{K_i}(x,a) \sum_{y} p_i(y|\theta_i) P(\{x'\}|x,a,y) .$$

(note that these definitions are consistent with definitions 3.1 (d) and (e), since $\theta_i$ is a singleton for $i \neq 1$, $i \in I$).

Lemma 7.4

Let $f \in F^*$ be optimal for some $e_1$, $\underline{e} \leq e_1 \leq \bar{e}$, i.e. $g(x,e_1) = c_f(x) + d_f(x)e_1$, for $x \in L_2$.
Then f is optimal for all $e \in [e_1, e_2]$ and non-optimal for $e > e_2$ where

7.15 $$e_2 := \max_{x \in L_2} \left[\min \left\{ \frac{c_f(x) - \tilde{r}(x,a) - \beta \sum_{x' \in L_2} \tilde{P}(x'|x,a) c_f(x')}{-d_f(x) + \beta \sum_{x' \in L_2} \tilde{P}(x'|x,a) d_f(x') + \beta \tilde{P}(0|x,a)} \right\}\right] ,$$

where the minimum has to be taken over all $a \in D(x)$ for which the denominator is positive (the minimum over the empty set is infinite).

Proof.

Note that, for $f \in F^*$ and $x \in L_2$:

$$c_f(x) + d_f(x)e = \tilde{r}(x,f(x)) + \beta \sum_{x' \in L_2} \tilde{P}(x'|x,f(x)) \{c_f(x') + d_f(x')\ e\} + \beta \tilde{P}(0|x,f(x))e.$$

Moreover for $a = f(x)$ the denominator in 7.15 vanishes.
By Howard's policy improvement routine (cf. [Ross (1970), corollary 6.4]) the policy $f \in F^*$ is optimal if and only if for all $a \in D(x)$ and $x \in L_2$:

$$c_f(x) + d_f(x)e \geq \tilde{r}(x,a) + \beta \sum_{x' \in L_2} \tilde{P}(x'|x,a)\{c_f(x') + d_f(x')e\} + \beta\tilde{P}(0|x,a)e \ .$$

Hence $f \in F^*$ is optimal if and only if for all $a \in D(x)$ and $x \in L_2$:

$$(*)\ c_f(x) - \tilde{r}(x,a) - \beta \sum_{x' \in L_2} \tilde{P}(x'|x,a)c_f(x') \geq e\{- d_f(x) + \beta \sum_{x' \in L_2} \tilde{P}(x'|x,a)d_f(x') + \beta\tilde{P}(0|x,a)\}.$$

So, if the denominator in 7.15 is less than or equal to zero, then (*) holds for all $e > e_1$, if it holds for $e_1$. On the other hand, if the denominator in 7.15 is greater than zero, then (*) holds for all $e > e_1$ such that e is less than or equal to the $e_2$.
This proves the lemma. □

According to lemma 7.4 we now have the following procedure to determine $g(x,\cdot)$ for $x \in L_2$. Compute for $e_1 := \underline{e}$ an optimal $f \in F^*$. Then determine $e_2$ by 7.15. If $e_1 < e_2 < \bar{e}$ then compute for $e_2$ a new optimal policy $f \in F^*$ and repeat this procedure. So $e_1, e_2, e_3, \ldots$ are computed. If during this process $e_n = e_{n-1}$ then we have to determine another optimal policy for the value $e_{n-1}$ such that $e_n > e_{n-1}$. Note that there always is such a policy (cf. lemma 7.3) and that we have to examine only finitely many Markov policies since X and A are finite.
Let us denote the set of optimal Markov policies determined in this way by $\bar{F}^*$. Using the values $c_f(x)$ and $d_f(x)$ for $x \in L_2$ and $f \in \bar{F}^*$ it is easy to determine $v(0,\theta)$ and optimal policies $f_\theta \in F$ for $\theta \in \Theta$. So, $\bar{F} = \{f_\theta, \theta \in \Theta\}$ and $v(0,\theta,f)$ for $\theta \in \Theta$ and $f \in \bar{F}$ are easy to compute.
This concludes the discussion of part 1.

In part 2 of the algorithm a suitable horizon is determined.
Here we have to compute for $q' \in W_n(q)$ the value $\varepsilon_0(q')$ and afterwards $E(q,\varepsilon_0,n)$ (cf. 6.11 and th. 6.5).
However, $\varepsilon_0(q')$ has a simple form in this case:

$$\varepsilon_0(q') = \tfrac{1}{2} \min_{f \in \bar{F}} \sum_{\theta \in \Theta} q(\theta)\{v(0,\theta) - v(0,\theta,f)\} \ .$$

Therefore we have, in a way similar to 6.14:

$$7.16\ E(q,\varepsilon_0,n) = \tfrac{1}{2}\{\sum_{\theta \in \Theta} q(\theta)v(0,\theta) - \sum_{y_1,\ldots,y_n} \max_{f \in \bar{F}} \sum_{\theta \in \Theta} q(\theta) \prod_{j=1}^{n} p_1(y_j|\theta_1)v(0,\theta,f)\}.$$

If $\{p_1(\cdot|\theta_1),\ \theta_1 \in \theta_1\}$ is an exponential family, then we might use the normal approximation to approximate $E(q,\varepsilon_0,n)$ (cf. remark (iii) section 7.1).

In the final part of the algorithm the backward induction is carried out. Here we use

$$7.17 \qquad (U_\sigma b)(0,q) = \max_{a\in D(0)} \sum_y p_1(y,q)[r(0,a,y) + \beta P(\{0\}|0,a,y)b(0,T_{1,y}(q)) +$$

$$+ \beta \sum_{x\in L_2} P(\{x\}|0,a,y) \max_{f\in \bar{F}^*} \{c_f(x) + d_f(x)b(0,T_{1,y}(q))\}] \quad .$$

Finally we summarize the steps of the algorithm.

*Algorithm* 2

part 1 : *parameter influence*.

(a) Determine for all $e \in [\underline{e},\bar{e}]$ an optimal $f \in F^*$ (cf. 7.12) and simultaneously $c_f(x)$ and $d_f(x)$ for $x \in L_2$. So $\bar{F}^*$ and $g(x,e)$ are determined for $x \in L_2$ and $\underline{e} \le e \le \bar{e}$.

(b) For all $\theta \in \theta$ compute $v(0,\theta)$ and an optimal $f_\theta \in F$ (using the results of step (a)). Hence $\bar{F}$ is determined.

(c) For all $\theta \in \theta$ and $f \in \bar{F}$ determine $v(0,\theta,f)$

part 2 : *horizon determination*.

(d) Compute $w(0,q) = \Sigma_{\theta\in\theta}\, q(\theta)\, v(0,\theta)$.

(e) Set $n := n_0$ ($n_0$ is a lower estimate of the horizon).

(f) Compute $\delta$:

$$\delta := \sum_{y_1,\ldots,y_n} \min_{f\in F} \sum_{\theta\in\theta} q(\theta) \cdot \prod_{j=1}^{n} p_1(y_j|\theta_1) v(0,\theta,f)\} \quad \text{(cf. 7.16)}.$$

(g) If $\frac{1}{2}\beta^n\{w(0,q) - \delta\} \le \Delta$ then go to step (h), otherwise set $n := n + 1$ and go to step (f) ($\Delta$ is the desired accuracy).

part 3 : *backward induction*.

(h) For $q' \in W_n(q)$ set : $v_n(0,q') := \frac{1}{2}\{w(0,q') + \ell(0,q')\}$ .

(i) Set $k := n - 1$.

(j) For all $q' \in W_k(q)$ compute $p(y,q')$ and $T_{1,y}(q')$ for all $y \in Y$, and then compute

$$v_k(0,q') = \max_{a\in D(0)} \sum_y p_1(y,q)[r(0,a,y) + \beta P(\{0\}|0,a,y)v_{k+1}(0,T_{1,y}(q')) +$$

$$+\beta \sum_{x\in L_2} P(\{x\}|0,a,y) \max_{f\in \bar{F}^*}\{c_f(x) + d_f(x)v_{k+1}(0,T_{1,y}(q'))\}] .$$

(k) If $k > 0$ then set $k := k - 1$ and go to (j), otherwise go to (ℓ).

(ℓ) For $x \in L_2$ compute

$$v_0(x,q) := \max_{f\in \bar{F}^*}\{c_f(x) + d_f(x)v_0(0,q)\} .$$

Note that $\tilde{v}(x,q) := v_0(x,q)$ is an approximation of $v(x,q)$ of the desired accuracy.

Finally we note that a modification to determine upper and lower bounds in part 3 of algorithm 2 can be incorporated in a way similar to the modification of algorithm 1. It is straightforward to modify th. 7.2 to show that we have exponential convergence in part 2 of algorithm 2.

## 7.3 *Numerical examples*

In this section we present 5 examples. The first three examples illustrate algorithm 1 (cf. section 7.1), and the last two examples illustrate algorithm 2 (cf. section 7.2).

### Example 7.1 *Inventory control*

We consider a well-known inventory control model. The model we studied in section 5.3 (model 5A) is a special case.
The cost function is

$$c(x,a) := hx^+ + px^- + k(a - x) + K\{1 - \delta(a,x)\} , \qquad K \geq 0 .$$

Here K represents the *order cost* or the cost for starting the production. Note that the cost K is incured only if the inventory is brought to a higher level. If $K = 0$ then we are dealing with model 5A again. In [Rieder (1972)] it is proved that there is an optimal Bayesian Markov policy f of the following form:

7.18 $$f(x,q) = S(q) \quad \text{if} \quad x < s(q)$$
$$= x \quad \text{if} \quad x \geq s(q) , \qquad (x,q) \in X \times W$$

where s and S are measurable functions from W to the interval [0,M].

(Note that s = S if K = 0). We specify the numerical data of the model.

In this example the demand is binomially distributed.

$$p(y|\theta) = \binom{5}{y}\theta^{y}(1-\theta)^{5-y} , \qquad y \in Y = \{0,1,\ldots,5\} ,$$

and $\theta \in \Theta = \{0.1,0.2,\ldots,0.9\}$.

The prior distribution q is uniform on $\Theta$, i.e. $q(\{\theta\}) = \frac{1}{9}$ for $\theta \in \Theta$.

The various costs are:

$$h = 0.1, \qquad p = 5, \qquad k = 3 .$$

We consider two cases: K = 0 and K = 1. Finally the discount factor is $\beta = 0.9$.

In the tables below, we display the optimal strategies $s(\theta)$ and $(s(\theta),S(\theta))$ for the models with known parameter values and the value function, at starting-inventory level zero for the various parameter values. Further we display, for several horizons, the value $E(q,\varepsilon_\infty,n)$ and finally we display the optimal actions for the first 3 stages and the value function at inventory level zero for the prior distributions $q' \in \cup_{n=0}^{2} W_n(q)$.

table 1 (K = 0)

| $\theta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $s(\theta)$ | 1 | 2 | 3 | 4 | 4 | 4 | 5 | 5 | 5 |
| $v(0,\theta)$ | 21 | 37 | 52 | 68 | 82 | 97 | 111 | 124 | 137 |

(The optimal strategy, if $\theta$ is known, is "order up to level $s(\theta)$ iff $x < s(\theta)$")

table 2 (K = 0)

| horizon n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $E(q,\varepsilon_\infty,n)$ | 3.6 | 1.7 | 1.1 | 0.8 | 0.6 | 0.5 | 0.4 | 0.3 |

table 3 (K = 1)

| θ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| s(θ) | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| S(θ) | 2 | 3 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| v(0,θ) | 24 | 42 | 59 | 76 | 91 | 107 | 121 | 134 | 147 |

(The optimal strategy here is: "order up to S(θ) iff x < s(θ)").

table 4 (K = 1)

| horizon n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| $E(q,\varepsilon_\infty,n)$ | 3.2 | 1.5 | 1.0 | 0.8 | 0.6 | 0.5 | 0.4 | 0.3 |

In table 5 we represent the posterior distributions $q' \in W_n(q)$ by $\Sigma_{i=1}^{n} y_i$, $y_i \in Y$ (cf. remark (iii) section 7.1).

table 5

| $\Sigma_{i=1}^{n} y_i$ | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| n | | K | | | | | | | | | | | |
| 0 | s(q) | 0 | 5 | | | | | | | | | | |
| 0 | v(0,q) | 0 | 82 | | | | | | | | | | |
| 0 | s(q) | 1 | 4 | | | | | | | | | | |
| 0 | S(q) | 1 | 5 | | | | | | | | | | |
| 0 | v(0,q) | 1 | 90 | | | | | | | | | | |
| 1 | s(q') | 0 | 2 | 3 | 4 | 5 | 5 | 5 | | | | | |
| 1 | v(q') | 0 | 36 | 52 | 72 | 93 | 111 | 125 | | | | | |
| 1 | s(q') | 1 | 2 | 2 | 3 | 4 | 4 | 5 | | | | | |
| 1 | S(q') | 1 | 3 | 3 | 4 | 5 | 5 | 5 | | | | | |
| 1 | v(0,q') | 1 | 41 | 58 | 80 | 101 | 120 | 134 | | | | | |
| 2 | s(q') | 0 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 5 | 5 | 5 |
| 2 | v(q') | 0 | 27 | 34 | 45 | 58 | 70 | 83 | 94 | 106 | 117 | 126 | 132 |
| 2 | s(q') | 1 | 1 | 2 | 2 | 2 | 3 | 3 | 4 | 4 | 4 | 5 | 5 |
| 2 | S(q') | 1 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 5 | 5 | 5 | 5 |
| 2 | v(0,q') | 1 | 31 | 39 | 51 | 65 | 78 | 91 | 104 | 116 | 127 | 136 | 142 |

(The optimal strategies are, if K = 0: "order up to s(q') iff x < s(q')" and if K = 1: "order up to S(q') iff x < s(q')").

Note that $\max_x\{\ell(x,q) - w(x,q)\} = E(q,\varepsilon_\infty,0)$.
It is remarkable that, although the variations in $\theta \to v(0,\theta)$ are large, the upper and lower bounds on $v(0,q)$ are very close.

Example 7.2 *Replacement under additive damage*

We consider the following replacement problem. In each time interval $(n-1,n]$, $n \in \mathbb{N}^*$, there is a random shock $Y_n$, which is observed at time n. The random variables $Y_1, Y_2, Y_3, \ldots$ are i.i.d., $Y_n \in Y = [0,\infty)$ and they act on the state of the system in the following way:

$$7.19 \qquad X_{n+1} = \min\{\delta(A_n,0)X_n + Y_{n+1}, x^*\} \quad , \; n \in \mathbb{N},$$

where $X_n$ is the state of the system at time n, $X_n \in X = [0,x^*]$ and $A_n$ is the action at time n. The action space is $A = \{0,1\}$. Action 0 means "do not replace" and action 1 means "replace" the machine. $D(x) := A$ for $x \in [0,x^*)$, $D(x^*) := \{1\}$. Replacement takes place instanteneous.
If the system is in state $x^*$ then replacement is more expensive than in the other states. The costs are:

$$c(x,a) = m(x)\delta(a,0) + R\delta(a,1) \quad , \quad \text{if} \quad 0 \le x < x^* \quad , \quad a \in A$$
$$= R^* \quad , \quad \text{if} \quad x = x^* \quad , \quad a \in A$$

where $R^* > R$. Here $m(x)$ are the maintenance costs for one period, if the state of the system is x. We assume that $x \to m(x)$, $x \in X$ is real-valued, measurable and nondecreasing, and $0 \le m(0)$, $m(x^*) < R$. It is further assumed that the distribution of $Y_n$ is incompletely known with density $p(\cdot|\theta)$ with respect to a $\sigma$-finite measure $\nu$ on Y, with $\theta \in \Theta$ where $\Theta$ is a complete separable metric space called the parameter space.
It is easy to transform this model into a Bayesian control model with index set I a singleton (cf. example 2.3).
Before we consider numerical data we first establish a property concerning the form of an optimal strategy for this Bayesian control model. In lemma 7.5 we show that the optimal strategy is characterized by so-called *control limits* in the following way. There is an optimal Bayesian Markov policy f such that

7.20 $\quad f(x,q) = 0 \quad \text{if} \quad x \leq s(q)$

$\qquad\qquad = 1 \quad \text{if} \quad x > s(q) \ , \qquad (x,q) \in X \times W \ ,$

where s is a measurable function from W to X.

(The values s(q) are called control limits).

The proof proceeds in a familiar way (cf.[Ross (1970), th. 6.9]).

Lemma 7.5

There is an optimal strategy of the form, given in 7.20.

Proof.

We first show that the value function v is nondecreasing in the first coordinate.

Consider the sequence of successive approximations $v_0, v_1, v_2, \ldots$ of v where $v_0 := 0$, $v_n := Uv_{n-1}$. By th. 3.14 we have $\lim_{n\to\infty} v_n(x,q) = v(x,q)$, for $(x,q) \in X \times W$.

Note that, for $0 \leq x < x^*$, $q \in W$ and $n \in \mathbb{N}^*$:

$$(*) \qquad v_n(x,q) = \min\{m(x) + \beta \int \nu(dy)p(y,q)v_{n-1}(\min(x+y,x^*),T_y(q)) \ , \quad R + \beta \int \nu(dy)p(y,q)v_{n-1}(\min(y,x^*),T_y(q))\} \ .$$

It is easily verified that for $q \in W$ and $n \in \mathbb{N}^*$:

$$(**) \qquad v_n(x^*,q) = R^* + \beta \int \nu(dy)p(y,q)v_{n-1}(\min(y,x^*),T_y(q)) \ .$$

Hence $x \to v_1(x,q)$ is nondecreasing, since $x \to m(x)$ is nondecreasing for all $q \in W$. Suppose that $x \to v_{n-1}(x,q)$ is nondecreasing for all $q \in W$. Then, by (*), $x \to v_n(x,q)$ is nondecreasing for all $q \in W$. Hence, by induction, $x \to v_n(x,q)$ is nondecreasing for all $q \in W$ and $n \in \mathbb{N}$ and therefore $x \to v(x,q)$ is nondecreasing for all $q \in W$.

It is straightforward to verify that $(x,q) \to v_1(x,q)$ is measurable and therefore, by (*) and an induction argument, $(x,q) \to v(x,q)$ is measurable (cf. lemma 1.6 (iii)). Define, for $x \in X$ and $q \in W$:

$$d(x,q) := m(x) + \beta \int \nu(dy)p(y,q)v(\min(x+y,x^*),T_y(q))$$

and

$$b(q) := R + \beta \int \nu(dy)p(y,q)v(\min(y,x^*),T_y(q)) \ .$$

It is easy to verify that $q \to d(x,q)$ and $q \to b(q)$ are measurable for $x \in X$.

Note that v satisfies the functional equation:

(***) $\qquad v(x,q) = \min\{d(x,q),b(q)\} \ , \qquad 0 \le x < x^* \ , \quad q \in W.$

It is straightforward to prove that a strategy which chooses in each state $(x,q)$, $0 \le x < x^*$ and $q \in W$, a minimizing action in (***) is optimal.

Let

$$s(q) := \sup\{x \mid 0 \le x < x^* \ , \ d(x,q) \le b(q)\} \ .$$

If $q \to s(q)$ is measurable, then the policy f defined in 7.20 is optimal.

To verify the measurability of $q \to s(q)$ note that for $0 \le a < x^*$:

$$\{q \in W \mid s(q) > a\} = \{q \in W \mid d(x,q) \le b(q) \quad \text{for some} \quad x > a\} =$$

$$= \{q \in W \mid d(x,q) \le b(q) \quad \text{for some rational number } x > a\}.$$

Since $q \to d(x,q)$ and $q \to b(q)$ are measurable for all $x \in X$ we conclude that $q \to s(q)$, $q \in W$ is measurable. □

In the numerical example the following data are used: $x^* = 25$, $m(x) = 0$ for all $x \in X$, $R = 75$, $R^* = 125$ and $\beta = 0.95$. Further $p(y \mid \theta) = \binom{9}{y}\theta^9(1-\theta)^{9-y}$, for $y \in Y = \{0,1,\ldots,9\}$ and $\theta \in \Theta = \{0.1,0.2,\ldots,0.9\}$.

Hence, if we start the system in an integer x, $0 \le x \le 25$ then the state $X_n$ is always an integer (cf. 7.19).

The prior distribution q on $\theta$ is the uniform distribution, i.e. $q(\{\theta\}) = \frac{1}{9}$, for $\theta \in \Theta$.

In table 6 we display the optimal strategies for the models with known parameter values and the value function for a new machine, i.e. in $x = 0$, for the various parameter values.

table 6

| $\theta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $s(\theta)$ | 5 | 7 | 9 | 10 | 11 | 12 | 13 | 13 | 14 |
| $v(0,\theta)$ | 461 | 546 | 600 | 639 | 670 | 693 | 714 | 731 | 745 |

(the optimal strategy, when $\theta$ is the true parameter, is "replace iff $x > s(\theta)$").

In table 7 we show for several horizons n: $E(q,\varepsilon_\infty,n)$.

table 7

| horizon n | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $E(q,\varepsilon_\infty,n)$ | 6.7 | 2.9 | 1.9 | 1.5 | 1.2 | 1.0 | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 |

Finally we show in table 8 for all $q' \in \cup_{n=0}^{2} W_n(q)$ the optimal control limit $s(q')$ and the optimal value $v(0,q')$. As in table 5 we represent $q' \in W_n(q)$ by $\Sigma_{i=1}^{n} y_i$ , $y_i \in Y$ .

table 8

| $\sum_{i=1}^{n} y_i$ | n = 0 | | n = 1 | | n = 2 | | $\sum_{i=1}^{n} y_i$ | n = 2 | |
|---|---|---|---|---|---|---|---|---|---|
| | s(q') | v(0,q') | s(q') | v(0,q') | s(q') | v(0,q') | | s(q') | v(0,q') |
| 0 | 10 | 645 | 6 | 495 | 5 | 472 | 10 | 11 | 679 |
| 1 | | | 7 | 529 | 6 | 483 | 11 | 12 | 691 |
| 2 | | | 8 | 574 | 6 | 506 | 12 | 13 | 702 |
| 3 | | | 9 | 615 | 7 | 535 | 13 | 13 | 712 |
| 4 | | | 10 | 649 | 8 | 566 | 14 | 13 | 721 |
| 5 | | | 11 | 675 | 9 | 593 | 15 | 13 | 730 |
| 6 | | | 12 | 698 | 9 | 615 | 16 | 14 | 736 |
| 7 | | | 13 | 716 | 10 | 634 | 17 | 14 | 740 |
| 8 | | | 13 | 730 | 11 | 651 | 18 | 14 | 743 |
| 9 | | | 14 | 738 | 11 | 666 | | | |

**Example** 7.3 *Heads or tails*

We consider a simple game with only one player, who may choose heads (action 1) or tails (action 2) of a coin with unknown probabilities: the probability of heads is $\theta$, $0 \leq \theta \leq 1$. The system has two states and only in state 1 the player has a choice. In state 2, the system stays there with probability $\theta$ or it goes to state 1 with probability $1 - \theta$. If "heads" has been chosen then the system stays in state 1 with probability $\theta$ and if "tails" has been chosen then it remains in state 1 with probability $1 - \theta$. Otherwise the system moves to state 2.

Only in state 1 an immediate reward 100 is obtained, independent of the action chosen. The discount factor is $\beta = 0.9$, and the prior distribution q on $\Theta = [0,1]$ is:

$$q(\{\tfrac{i}{10}\}) = \frac{1}{9} \quad \text{for } i = 1,2,\ldots,9 \ .$$

It is easy to transform this problem into a model considered in section 7.1. At first glance one might think that the optimal strategy is: "if the system is in $(1,\tilde{q})$ then choose heads if $\int \theta \tilde{q}(d\theta) \geq \frac{1}{2}$ and choose tails otherwise" ($\tilde{q} \in W$).

However, if we consider a prior distribution which is concentrated on the set $\{0,1\}$, then it is straightforward to verify that the action "heads" is optimal in state $(1,\tilde{q})$ if and only if $\tilde{q}(\{1\}) \geq \frac{1-\beta}{2-\beta}$ .

In table 9 the optimal actions and the value function are displayed for the models with known parameter values. In table 10 we present, for four horizons and all possible posterior distributions, the value function in state 1, the upper and lower bound in state 1 and the optimal action. Note that there is a one-to-one correspondence between the posterior distributions and the number of "heads" for each horizon.

table 9

| $\theta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| action (in 1) | 2 | 2 | 2 | 2 | 1 or 2 | 1 | 1 | 1 | 1 |
| $v(1,\theta)$ | 910 | 820 | 730 | 640 | 550 | 561 | 578 | 609 | 678 |

table 10

| horizon n | number of heads | $\ell(1,\cdot)$ | $v(1,\cdot)$ | $w(1,\cdot)$ | action |
|---|---|---|---|---|---|
| 0 | 0 | 568 | 639 | 675 | 1 (heads) |
| 1 | 0 | 589 | 611 | 631 | 1 |
| | 1 | 670 | 687 | 718 | 2 (tails) |
| 2 | 0 | 604 | 611 | 622 | 1 |
| | 1 | 562 | 616 | 649 | 1 |
| | 2 | 739 | 744 | 758 | 2 |
| 3 | 0 | 616 | 618 | 623 | 1 |
| | 1 | 575 | 595 | 617 | 1 |
| | 2 | 636 | 652 | 682 | 2 |
| | 3 | 781 | 783 | 788 | 2 |

For the horizon n = 14 we have $\mathbb{E}_q[\varepsilon_\infty(Q_n)] = 8.5$. Note that the range between the upper and lower bounds are relatively large compared with the examples 7.1 and 7.2.

**Example 7.4** *Taxi driver's problem*

We consider a model that fits into the setting of section 7.2. At the cab-rank a taxi driver is offered a run of size $Y_n \in Y \subset \mathbb{N}$, at each stage n. If he accepts this run he will be away for $Y_n$ stages and if he refuses the run he remains in the cab-rank. The random variables $Y_1, Y_2, Y_3, \ldots$ are i.i.d. and observable if and only if the taxi driver is in the cab-rank. The distribution of $Y_n$ is incompletely known. Only at the cab-rank the taxi driver chooses an action $a \in A = Y$. Action a means: "accept all runs larger than or equal to a and refuse the runs smaller than a". Only if he accepts a run of size $y \in Y$ he obtains a reward $r(y)$, where $y \to r(y)$ is nondecreasing. To transform this model into a Bayesian control model define the state space X by

$$X := \{0\} \cup \{(n,k) \mid k = 1,\ldots,n-1, \quad n \in \mathbb{N}^*\}$$

and the transition law by:

$$P(\{n, k+1)\} \mid (n,k),a,y) = 1 \quad \text{for} \quad 1 \le k < n-1, \quad n,k \in \mathbb{N}^*,$$

$$P(\{0\} \mid (n,n-1),a,y) = 1 \quad \text{for } n \in \mathbb{N}^*,$$

$$P(\{(n,1)\} \mid 0,a,n) = 1 \quad \text{for } n \in \mathbb{N}^*.$$

Note that, if the system is in state $(n,k)$, $1 \le k \le n-1$, the taxi has been away for k time units on a trip of n time units in total. Further $p(\cdot \mid \theta)$ is the probability density of $Y_n$ with respect to the counting measure on $\mathbb{N}$, for $\theta \in \Theta$, where $\Theta$ is the parameter space.
We consider the operator $U_\sigma$ (cf. 6.7) and we obtain the optimality equation for the taxi driver's problem:

$$v(0,q) = \sup_{a \in A}\Big[\sum_{y \ge a} p(y,q)\{r(y) + \beta^y v(0,T_y(q))\} + \beta \sum_{y<a} p(y,q)v(0,T_y(q))\Big].$$

In the numerical example we used the following data:

$$Y = \{1,2,\ldots,10\}, \quad r(y) = e^y, \quad y \in Y, \quad \beta = 0.9,$$

$$p(y \mid \theta) = b(\theta) \cdot \exp\{-(y-\theta)^2\}, \quad y \in Y, \quad \theta \in \Theta = \{1,2,\ldots,10\}$$

where $b(\theta) = (\Sigma_{y \in Y} \exp\{-(y-\theta)^2\})^{-1}$.

The prior distribution q is uniform, i.e. $q(\theta) = 0.1, \quad \theta \in \Theta$.
Note that $\{p(\cdot|\theta), \theta \in \Theta\}$ is an exponential family and there is a one-to-one correspondence between $W_n(q)$ and $\{\Sigma_{i=1}^{n} y_i | y_i \in Y\}$ (cf. 7.1). In table 11 we display optimal strategies and the value function v in state 0 for the models with known parameter values.
In table 12 we display for several horizons n the value $E(q,\varepsilon_0,n)$ (cf. 7.16). This quantity is obtained using normal approximation (cf. remark (iii), section 7.1) for $n > 5$.

table 11

| $\theta$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| action | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 10 |
| $v(0,\theta)$ | 6 | 16 | 43 | 118 | 321 | 874 | 2376 | 6448 | 15997 | 21207 |

table 12

| horizon n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E(q,\varepsilon_0,n)$ | 327 | 253 | 153 | 125 | 80 | 64 | 44 | 36 | 25 | 20 | 13 | 11 | 8 | 6 | 5 |

For all $q' \in W_1(q) \cup W_2(q)$ the value function in $(0,q')$, $v(0,q')$ and the optimal actions are shown. (Note that each posterior distribution is represented by $\Sigma_{i=1}^{n} y_i$, $y_i \in Y$.)

table 13

| n = 0 | | v(q) = 4508 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| n = 1 | $y_1$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| | $v(0,q')$ | 8 | 17 | 49 | 134 | 366 | 995 | 2697 | 6992 | 14960 | 19859 |
| | action | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 10 | 10 |
| n = 2 | $y_1 + y_2$ | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | |
| | $v(0,q')$ | 6 | 9 | 15 | 28 | 44 | 76 | 120 | 209 | 326 | |
| | action | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | |
| | $y_1 + y_2$ | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
| | $v(0,q')$ | 568 | 887 | 1544 | 2412 | 4187 | 6493 | 10665 | 15704 | 19100 | 20798 |
| | action | 7 | 7 | 8 | 8 | 9 | 9 | 10 | 10 | 10 | 10 |

**Example** 7.5 *Compound replacement*

We consider another model that fits into the setting of section 7.2. Consider a replacement problem with two types of machines. If the controller decides to replace his machine he is not sure of what type the new machine will be. The probabilistic behaviour of both types of machines is known. However the probability of obtaining a machine of type 1 is $\theta \in (0,1)$, and a machine of type 2 is $1 - \theta$. The parameter $\theta$ is unknown. We first describe the machines. The life time of a machine of type i (i = 1 or 2) is geometric with parameter $p_i$, $p_1 = 0.9$ and $p_2 = 0.8$. If a machine is "alive" at stage n the controller may replace it (action 1) by a new one (of type 1 with probability $\theta$ and of type 2 with probability $1- \theta$). The costs of such a replacement are $C_i$ for machine i: $C_1 = 10$, $C_2 = 21$. If the controller decides to keep the machine i at stage n he has to pay maintenance costs $m_i(n)$ with probability $p_i$ and he has to pay costs for an emergency replacement $R_i$ with probability $(1 - p_i)$. Here $m_1(n) = 3(\frac{n}{10} + 1)^2 - 3$ for $n \in \mathbb{N}^*$,
$m_2(n) = 4(\frac{n}{10} + 1)^2$, $n \in \mathbb{N}^*$, $R_1 = 10$ and $R_2 = 25$.
The discount factor is $\beta = 0.9$.
We transform this model into a Bayesian control model in the following way. Let $Y = \{0,1\}$ and $p(y|\theta) = \theta^y(1 - \theta)^{1-y}$, $y \in Y$, $\theta \in (0,1)$. Let $X = \{0\} \cup \{(i,n) | i = \{1,2\}, n \in \mathbb{N}^*\}$ and $A = \{0,1\}$. State (i,n) means that we have a machine of type i of age n. State 0 means that we are replacing the machine. The actions have the same meaning as in example 7.2. The transition law is given by

$$P(\{(i,n+1)\}|(i,n),0,y) = p_i \ .$$

$$P(\{0\}|(i,n),0,y) = 1 - p_i \ .$$

$$P(\{0\}|(i,n),1,y) = 1 \ , \quad i \in \{1,2\}, \quad n \in \mathbb{N}^* \text{ and all } y \in Y \ .$$

$$P(\{(y,1)\}|0,0,y) = P(\{(y,1)\}|0,1,y) = 1 \ , \quad y \in Y \ .$$

The cost function is:

$$c((i,n),0) = p_i m_i(n) + (1 - p_i)R_i \ .$$

$$c((i,n),1) = C_i \ , \quad i \in \{1,2\}, \quad n \in \mathbb{N}^* \ .$$

$$c(0,0) = c(0,1) = 0 \ .$$

We consider the operator $U_\sigma$ (cf. 6.7). It is easy to show, in a way similar to the proof of lemma 7.5 that the "interesting" strategies in the set $F^*$ (cf. 7.12(i)) are characterized by two control limits $n_1, n_2 \in \mathbb{N}^*$ such that the controller chooses action 0 if and only if the system is in state (i,n)

with $n < n_i$, $i \in \{1,2\}$. This means that in 7.13 the maximizing $f \in F^*$ is found in this class, for all admissible scrapfunctions b.
Using this property, we find the optimality equation (cf. 7.17):

7.21 $$v(0,q) = \beta p(1,q) \min_{n \in \mathbb{N}^*} \{A_1(n) + B_1(n) v(0,T_1(q))\} +$$

$$+ \beta p(0,q) \min_{n \in \mathbb{N}^*} \{A_2(n) + B_2(n) v(0,T_0(q))\}$$

where

$$A_i(n) = \sum_{\ell=1}^{n-1} (\beta p_i)^{\ell-1} \{p_i m_i(\ell) + (1 - p_i) R_i\} + (\beta p_i)^{n-1} C_i$$

$$B_i(n) = \beta \sum_{\ell=1}^{n-1} (\beta p_i)^{\ell-1} (1 - p_i) + (\beta p_i)^{n-1} \quad , \quad i \in \{1,2\} \quad , \quad n \in \mathbb{N}^*$$

(here $A_i(n)$ equals $c_f(\{i,1\})$ and $B_i(n)$ equals $d_f(\{i,1\})$ where $c_f$ and $d_f$ are defined in 7.12 and $f \in F^*$ is the strategy that replaces only in states $(i,k)$ with $k \geq n$).

In the numerical example we have modified the model in such a way that the state space X becomes finite: in states $(i,10)$, $i \in \{1,2\}$ we allow only the action 1, i.e. we always replace the system in these states.

The prior distribution q is given by: $q(\{\frac{j}{10}\}) = \frac{1}{9}$, $j = 1,2,\ldots,9$.
Hence all posterior distributions are concentrated on the set $\{\frac{1}{10}, \ldots, \frac{9}{10}\}$.
In table 14 we display the values $v(0,\theta)$, $\theta \in \Theta$ and the optimal strategies for these parameter values. The strategies are characterized by pairs of numbers $(n_1,n_2)$ indicating the control limits for both machines.
After that, in table 15 we display the values $E(q,\varepsilon_0,n)$ (cf. 7.16) for several horizons n.

table 14

| $\theta$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---|---|---|---|---|---|---|---|---|---|
| $v(0,\theta)$ | 87 | 79 | 71 | 64 | 57 | 51 | 45 | 39 | 34 |
| $n_1,n_2$ | 10,6 | 10,5 | 10,5 | 9,4 | 9,4 | 8,3 | 7,2 | 7,2 | 6,2 |

table 15

| horizon | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $E(q,\varepsilon_0,n)$ x 0.01 | 56 | 46 | 40 | 35 | 30 | 27 | 25 | 23 | 20 | 19 | 18 | 17 |

Note that each posterior distribution $\tilde{q}$ of q is completely characterized by the number of times the replaced machine is of type 1, i.e. $\Sigma_{i=1}^{n} y_i$ determines the posterior distribution. In table 16 we display, for the first 7 stages, the value function in $(0,\tilde{q})$, $v(0,\tilde{q})$ for $\tilde{q} \in \cup_{n=0}^{6} W_n(q)$ and the optimal control limits $n_1$ and $n_2$ for the two types of machines (for example if $n = 2$ and $\Sigma_{i=1}^{2} y_i = 1$, then $n_1 = 7$, $n_2 = 3$ and $v(0,\tilde{q}) = 48$).

table 16

| n \ $\Sigma_{i=1}^{n} y_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| 0 | 49<br>7,3 | | | | | | |
| 1 | 56<br>7,3 | 42<br>6,2 | | | | | |
| 2 | 60<br>8,3 | 48<br>7,3 | 38<br>6,2 | | | | |
| 3 | 63<br>8,4 | 53<br>7,3 | 43<br>6,2 | 35<br>6,2 | | | |
| 4 | 65<br>8,4 | 57<br>8,3 | 48<br>7,3 | 40<br>6,2 | 34<br>6,2 | | |
| 5 | 66<br>8,4 | 60<br>8,3 | 52<br>7,3 | 45<br>7,2 | 38<br>6,2 | 33<br>6,2 | |
| 6 | 67<br>9,4 | 62<br>8,4 | 55<br>7,3 | 48<br>7,3 | 42<br>6,2 | 36<br>6,2 | 32<br>6,2 |

# APPENDIX A. RESULTS FROM ANALYSIS

## 1. Analytic sets, semi-analytic functions and related subjects

We summarize some pertinent facts about analytic sets and semi-analytic functions. For analytic sets we refer to [Parthasarathy (1967)]. Similar summaries are found in [Blackwell,Freedman and Orkin (1974)] and [Shreve (1977)].

Let N be the Cartesian product of countably many copies of $\mathbb{N}^*$ where $\mathbb{N}^*$ is endowed with the discrete topology and N with the product topology. Let X be a complete separable metric space.

A subset $A \subset X$ is called *analytic* if there is a continuous function from N to X with $f(N) = A$, moreover $\emptyset$ is analytic. The following properties hold. The proofs are found in [Parthasarathy (1967) chapter I section 3]).

A1 Each Borel subset of a Borel space is analytic.

A2 Countable unions, intersections and Cartesian products of analytic sets are analytic.

A3 If A and B are analytic subsets of Borel spaces $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ respectively and if f is a Borel measurable function from X to Y then f(A) and $f^{-1}(B)$ are analytic.

As a consequence of A3 we have:

A4 If A is an analytic subset of $X \times Y$ then $\mathrm{proj}_X(A)$ is analytic.

Let $(X,\mathcal{X})$ be a Borel space. For each $p \in P(\mathcal{X})$ we have the $\sigma$-field

$\mathcal{X}_p := \{B \cup A \mid B \in \mathcal{X}$ , and there is a $C \in \mathcal{X}$ such that $A \subset C$ and $p(C) = 0\}$ .

$\mathcal{X}_p$ is called the *completion* of $\mathcal{X}$ with respect to p. The *universal* $\sigma$-field $\mathcal{U}_X$ is defined by $\mathcal{U}_X := \bigcap_{p \in P(\mathcal{X})} \mathcal{X}_p$. $A \in \mathcal{U}_X$ is called *universally measurable*.

A5 Every analytic subset of a Borel space is universally measurable.

For a proof see [Christensen (1974) th. 1.5 or th. 1.7].

A6 For each probability $p \in P(\mathcal{X})$ where $(X,\mathcal{X})$ is a Borel space, there is a unique extension $p^*$ on $\mathcal{U}_X$. And for each real-valued function f on X which is $\mathcal{U}_X$-measurable there is for each $p \in P(\mathcal{X})$ an $\mathcal{X}$-measurable function $\tilde{f}$ such that $f = \tilde{f}$ p-a.s. (the proof is straightforward).

A7 *Kuratowski theorem* (see [Parthasarathy (1967) chapter I corollary 3.3]). If $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ are Borel spaces and $f : X \to Y$ is Borel measurable and one-to-one, then f(X) is a Borel subset of Y and $f^{-1}$ is Borel measurable.

Let $(X,\mathcal{X})$ be a Borel-space. A real-valued function f on X is called *lower semi-analytic* (l.s.a.) if $\{x|f(x) < c\}$ is analytic, for $c \in \mathbb{R}$, and f is called *upper semi-analytic* (u.s.a.) if -f is l.s.a.

The following properties hold.

A8 If f and g are l.s.a. (u.s.a.) then f + g is l.s.a. (u.s.a.) If $f_k$, $k \in \mathbb{N}$ are l.s.a. then $\inf_k f_k$ is l.s.a. and if $f_k$, $k \in \mathbb{N}$ are u.s.a. then $\sup_k f_k$ is u.s.a.

Proof.

Note that: $\{x|f(x) + g(x) < c\} = \bigcup_{y\in Q} \{x|f(x) < y, g(x) < c - y\}$ where Q is the set of rational numbers. Further $\{x|\inf_k f_k(x) < c\} = \bigcup_k \{x|f_k(x) < c\}$. So by A2 the statement has been proved for l.s.a. functions. For u.s.a. functions the proof follows from the definition. □

A9 Let $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ be Borel space, and let $g : X \to Y$ be a Borel function and f a l.s.a. (u.s.a.) function on Y. Then $f \circ g$ is l.s.a. (u.s.a.).

Proof.

Let f be l.s.a.
Then $\{y|f(y) < c\}$ is analytic, for $c \in \mathbb{R}$. Hence $\{x|f(g(x)) < c\} = \{x|g(x) \in \{y|f(y) < c\}\}$ is analytic, by A3.
Similarly if f is u.s.a. □

A10 Let $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ be Borel spaces and $f : X \times Y \to \mathbb{R}$ be bounded from above and measurable /l.s.a./u.s.a. Further let P be a transition probability from X to Y. Then the function $x \to \int f(x,y)P(dy|x)$ is measurable/ l.s.a./u.s.a.

Proof.

If f is l.s.a./u.s.a. the proof can be found in [Shreve (1977) th. 2.4]. Note that a function is measurable if and only if it is both l.s.a. and u.s.a., which proves the statement if f is measurable. □

A11 If f is l.s.a. or u.s.a. then f is universally measurable.
(the proof is trivial).

## 2. *Semi-continuous functions and measurable selections*

Let X be a metric space. A real-valued function f on X is called *upper semi-continuous* (u.s.c.) at $x_0 \in X$ if $\limsup_{n\to\infty} f(x_n) \le f(x_0)$ for any sequence $\{x_n | x_n \in X, n \in \mathbb{N}\}$ such that $\lim_{n\to\infty} x_n = x_0$ and f is called *lower semi-continuous* (l.s.c.) if -f is u.s.c.

A12 Let f be u.s.c. on the metric space X. Then there is a nonincreasing sequence of bounded continuous functions $\{f_k, k \in \mathbb{N}\}$ on X such that $\lim_{k\to\infty} f_k = f$.
For a proof see [Hausdorff (1957) section 4.2].

A13 If f and g are u.s.c. (l.s.c.) then f + g is u.s.c. (l.s.c.). If $\{f_k, k \in \mathbb{N}\}$ is a nonincreasing sequence of nonpositive u.s.c. functions then $f := \lim_{k\to\infty} f_k$ is u.s.c.
For a proof see [Hinderer (1970) page 32].

A14 Let g be continuous and nonnegative, and let f be u.s.c. (l.s.c.) and bounded. Then f·g is u.s.c. (l.s.c.).
For a proof cf. [Hinderer (1970) lemma 5.5(ii)].

A15 Let X and Y be metric spaces, g a continuous function from X to Y and f is u.s.c. (l.s.c.) on Y. Then $g \circ f$ is also u.s.c. (l.s.c.).

Proof.

Let $x_n \in X$ $(n \in \mathbb{N})$ with $\lim_{n\to\infty} x_n = x_0$. Then, since $\lim_{n\to\infty} g(x_n) = g(x_0)$:

$$\limsup_{n\to\infty} f(g(x_n)) \le f(g(x_0)).$$ □

A16 Let A be an index set and let $f_k$, $k \in A$ be l.s.c. Then $\sup_{k\in A} f_k$ is l.s.c.
If $f_k$, $k \in A$ is u.s.c. then $\inf_{k\in A} f_k$ is u.s.c.

Proof.

Let $x_n \in X$, $n \in \mathbb{N}$ and $\lim_{n\to\infty} x_n = x_0$. Then

$$\liminf_{n\to\infty} \sup_{k\in A} f_k(x_n) \ge \liminf_{n\to\infty} f_k(x_n) \ge f_k(x_0) \text{ for all } k \in A.$$ □

We continue with a result of Schäl on measurable selections. We first introduce some notations.

Let $(X,\mathcal{X})$ and $(A,\mathcal{A})$ be Borel spaces.

$L(X \times A) := \{f : X \times A \to \mathbb{R} \mid f$ is bounded Borel measurable and $a \to f(x,a)$ is continuous$\}$.

$\hat{L}(X \times A) := \{f : X \times A \to \mathbb{R} \mid f$ is Borel measurable, bounded from above and $f$ is the limit of some nonincreasing sequence of functions $f_n \in L(X \times A)\}$.

A17 Let $A$ be compact and $f \in \hat{L}(X \times A)$. Then there is a measurable mapping $g : X \to A$ such that

$$f(x,g(x)) = \max_{a \in A} f(x,a) \ .$$

For a proof see [Schäl (1975) th. 12.1].

## APPENDIX B. REMARKS ON THE MINIMAX CRITERION

Consider the Bayesian control model (cf. 2.1). Instead of rating the strategies $\pi \in \Pi$ by their Bayesian discounted total returns (cf 2.12) we might say $\pi^*$ is at least as good as $\pi$ in state x, if

$$\inf_{\theta\in\Theta} v(x,\theta,\pi^*) \geq \inf_{\theta\in\Theta} v(x,\theta,\pi) \ .$$

Let $\varepsilon \geq 0$. A strategy $\pi^* \in \Pi$ is called $\varepsilon$-*minimax* in state $x \in X$, if

$$\inf_{\theta\in\Theta} v(x,\theta,\pi^*) \geq \sup_{\pi\in\Pi} \inf_{\theta\in\Theta} v(x,\theta,\pi) - \varepsilon \ .$$

A 0-minimax strategy is simply called *minimax*.

A term "maximin" would be preferable, however in statistical decision theory the term "minimax" is current since one is interested in minimizing the expected loss instead of maximizing the expected return (cf. [Wald (1947)]).

We shall discuss a nice property of the Bayes criterion, which the minimax criterion does not have. Let $\pi = (\pi_0,\pi_1,\ldots) \in \Pi$ be an optimal strategy for the Bayes criterion, if the process is started in state $x \in X$ and if $q \in W$ is the prior distribution. Let the history at stage 1 be $(x,a,y,x') \in H_1$ and define similar to 3.22 the "tail-strategy" $\pi^* = (\pi_0^*,\pi_1^*,\ldots)$ by

$$\pi_k^*(\cdot|x_0,a_0,y_1,x_1,\ldots,y_k,x_k) := \pi_{k+1}(\cdot|x,a,y,x_0,a_0,y_1,x_1,\ldots,y_k,x_k)$$

for $k \in \mathbb{N}$. Then it is easy to verify that the strategy $\pi^*$ is optimal for the Bayes criterion if the process is started in x' with respect to the prior distribution $\Sigma_{i\in I} 1_{K_i}(x,a)T_{i,y}(q)$.

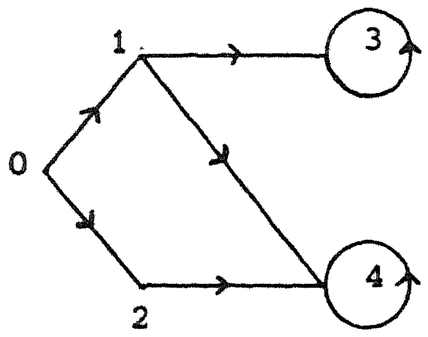
Hence the decision maker, who chooses a strategy that is optimal for the Bayes criterion, uses at each stage a strategy that is optimal for the Bayes criterion from that stage on, with respect to an "updated" prior distribution. In fact this property is the well-known "principle of optimality", for the equivalent dynamic program (model 2). In [Groenewegen (1978)] this principle is studied extensively . The property discussed above, enables us to compute the value function and the optimal actions by backward induction.

However, we show by an example that the minimax criterion does not have this property. A decision maker, who prefers the minimax criterion might be considered as a pessimist. However in the example he seems to forget his pessimism after one transition.

Another unpleasant property of the minimax criterion is that we may not

restrict our attention to the nonrandomized strategies. In the example it turns out that none of the nonrandomized strategies is optimal.

Example.




We start with an informal description. Only in states 0 and 1 there are two actions. In all other states the transitions are deterministic. If the decision maker chooses action 1 in state 0 then the next state will be 1 with probability $\theta$ and 2 with probability $1 - \theta$. If he chooses action 2, then the system moves to state 1 with probability $1 - \theta$ and to state 2 with probability $\theta$. In state 1 the two actions have the same effect with respect to the states 3 and 4. The parameter $\theta$ is unknown. In state 2 the reward is large compared with the rewards in the other states. This causes the "least favourable" parameter value for each strategy to be completely determined by the action chosen in state 0, if he starts there. We continue with a formal description of the example in terms of model 1.

$X = \{0,1,2,3,4\}$, $Y = \{0,1\}$, $A = \{1,2\}$, $D(0) = D(1) = A$, $D(3) = D(2) = D(4) = \{1\}$, $\theta = \{0.1,0.9\}$ and I is a singleton. The transition probabilities are determined by a function $F : X \times A \times Y \to X$ in the following way $P(\{F(x,a,y)\}|x,a,y) = 1$ (cf. example 2.3 chapter 1):

$F(0,1,y) = 1\delta(1,y) + 2\delta(0,y)$, $F(0,2,y) = 2\delta(1,y) + 1\delta(0,y)$

$F(1,1,y) = 3\delta(1,y) + 4\delta(0,y)$, $F(1,2,y) = 4\delta(1,y) + 3\delta(0,y)$ and $F(2,1,y) = 4$, $F(3,1,y) = 3$, $F(4,1,y) = 4$ for all $y \in Y$.

Further $p(y|\theta) = \theta^{y}(1 - \theta)^{1-y}$ for $y \in Y$ and $\theta \in \theta$.

The reward function r is given by

$r(0,1,y) = 8$, $r(0,2,y) = 1$, $r(1,1,y) = 25$, $r(1,2,y) = 20$

$r(2,1,y) = 200$, $r(3,1,y) = 2$, $r(4,1,y) = 14$ for all $y \in Y$, and the discount factor $\beta = \frac{1}{2}$. We omit y in the notation for r. First we consider a decision maker who starts in state 0. Any strategy for him can be characterized by three numbers a, b, and c. Here a is the probability of choosing action 1 in state 0, b the probability of choosing 1 in state 1 if in state 0 action 1 is chosen, and c is the probability of choosing action 1 in state 1 if in state 0 action 2 is chosen.

Let $v(0,\theta,(a,b,c))$ be the expected discounted total return in state 0 for the strategy given by a, b and c, if $\theta$ is the true parameter value. It is

straightforward to verify that

| $\theta$ | $v(0,\theta,(a,b,c))$ |
|---|---|
| 0.1 | $a(0.73b - 6.57c + 83.32) + 6.57c + 22.14$ |
| 0.9 | $a(-2.07b + 0.23c - 65.48) - 0.23c + 98.94$ |

Note that, for fixed b and c, the maximum over $a \in [0,1]$ of $\min_{\theta\in\Theta} v(0,\theta,(a,b,c))$ is attained for

$$a = \frac{76.8 - 6.8c}{148.8 - 6.8c + 2.8b}$$

and

$$f(b,c) := \max_a \min_\theta v(0,\theta,(a,b,c)) = \frac{76.8 - 6.8c}{148.8 - 6.8c + 2.8b}(0.73b - 6.57c + 83.32) + 6.57c + 22.14 .$$

Further note that $f(b,c)$ attains its maximum over $(b,c) \in [0,1]^2$ in a boundary point. It turns out that the optimal pair $(b,c)$ is $(0,1)$, and $f(0,1) = 65.54 \ldots$ .

Hence the optimal strategy is: $a = 0.49\ldots$ , $b = 0$ and $c = 1$.

It is easy to verify that all nonrandomized strategies are less good than this strategy.

Next we consider the minimax strategy for the situation that a second decision maker starts in state 1. Suppose this second decision maker has the same information concerning the unknown parameter as the first decision maker, i.e. he performs a Bernoulli trial with parameter $\theta$. Hence he works with the conditional distribution, given this observation. However, since this experiment is independent of the process, it does not change the transition law for the second decision maker.

(Note that if the parameter set would be $\{0,1\}$ then the observation of the experiment would reduce the parameter set to a singleton.)

The strategies for the second decision maker are characterized by the probability d of choosing action 1. Note that

| $\theta$ | $v(1,\theta,(d))$ |
|---|---|
| 0.1 | $37.8d + 23.2(1 - d)$ |
| 0.9 | $28.2d + 32.8(1 - d)$ |

Hence the optimal strategy is: d = ½.

So, if the first decision maker reaches state 1, he does not randomize but he chooses action 2 if he has chosen action 1 in state 0, otherwise he chooses action 1, and the second decision maker randomizes between the two actions with probability ½. The first decision maker acts in state 1 as if he knows the true parameter in state 1.

## REFERENCES


Aoki, M., Optimization of stochastic systems.
New York etc., Academic Press (1967).

Bauer, H., Wahrscheinlichkeitstheorie und Grundzüge der Masstheorie.
Berlin, Walter de Gruyter, (1968).

Behara, M. and Nath, P., Additive and nonadditive entropies of finite measurable partitions. In: Probability and information theory II.
Lecture Notes in Mathematics 296, Springer Verlag (1973); 102-138.

Bellman, R., Dynamic programming.
Princeton (N.J.), Princeton University Press (1957).

Bellman, R., Adaptive control processes: a guided tour.
Princeton (N.J.), Princeton University Press (1961).

Bertsekas, D.P., Dynamic programming and stochastic control.
New York etc., Academic Press (1976).

Blackwell, D., Discounted dynamic programming.
Ann. Math. Statist. 36 (1965), 226-235.

Blackwell, D., Freedman, D.A. and Orkin, M., The optimal reward operator in dynamic programming.
Ann. Probability 2 (1974), 926-941.

Christensen, J.P.R., Topology and Borel structure.
Amsterdam etc., North-Holland/American Elsevier (1974).

Denardo, E.V., Contraction mappings in the theory underlying dynamic programming.
SIAM Rev. 9 (1967), 165-177.

Derman, C., Denumerable state Markovian decision processes-average cost criterion.
Ann. Math. Statist. 37 (1966), 1545-1554.

Derman, C., Finite state Markovian decision processes.
New York etc., Academic Press (1970).

Derman, C. and Strauch, R.E., A note on memoryless rules for controlling sequential control processes.
Ann. Math. Statist. 37 (1966), 276-278.

Doob, J.L., Applications of the theory of martingales. In: Le calcul des probabilités et ses applications. Colloq. Internat. du CNRS 13 (1949) 23-27.

Dynkin, E.B., Controlled random sequences.
Theor. Probability Appl. 10 (1965), 1-14.

Fabius, J., Asymptotic behavior of Bayes' estimates.
Ann. Math. Statist. 35 (1964), 846-856.

Fox, B.L. and Rolph, J.E., Adaptive policies for Markov renewal programs.
Ann. Math. Statist. 1 (1973), 334-341.

Freedman, D.A., On the asymptotic behavior of Bayes estimates in the discrete case.
Ann. Math. Statist. 34 (1963), 1386-1403.

Freedman, D.A., On the asymptotic behavior of Bayes estimates in the discrete case II.
Ann. Math. Statist. 36 (1965), 451-456.

Furukawa, N., Fundamental theorems in a Bayes controlled process.
Bull. Math. Statist. 14 (1970), 103-110.

Groenewegen, L.P.J., Characterization of optimal strategies in dynamic games.
University of Technology Eindhoven, dissertation (1978).

Hastings, N.A.J. and van Nunen, J.A.E.E., The action elimination algorithm for Markov decision processes. Markov decision theory; ed. by H.C. Tijms and J. Wessels.
Amsterdam, Math. Centre Tracts 93 (1977), 161-178.

Hausdorff, F., Set theory.
New York, Chelsea (1957).

van Hee, K.M., Adaptive control of specially structured Markov chains. In: Dynamische Optimierung. Tagungsband des Sonderforschungsbereiches 72.
Bonner Mathematische Schriften 98 (1976), 99-116.

van Hee, K.M., Approximations in Bayesian controlled Markov chains. In: Markov decision theory; ed. by H.C. Tijms and J. Wessels.
Amsterdam, Math. Centre Tracts 93 (1977), 171-182.

van Hee, K.M., Hordijk, A. and van der Wal, J., Successive approximations for convergent dynamic programming. In Markov decision theory, ed. by H.C. Tijms and J. Wessels.
Amsterdam, Math. Centre Tracts 93 (1977), 183-212.

Hinderer, K., Foundations of non-stationary dynamic programming with discrete time parameter.
Lecture Notes in Operations Research and Mathematical Economics 33, Springer Verlag (1970).

Howard, R.A., Dynamic programming and Markov processes.
Cambridge (Mass.) M.I.T. Press (1960).

Iglehart, D.L., The dynamic inventory problem with unknown demand distribution.
Management Sci. 10 (1964), 429-440.

Kushner, H., Introduction to stochastic control.
New York, Holt (1971).

MacQueen, J., A modified dynamic programming method for Markovian decision problems.
J. Math. Anal. Appl. 14 (1966), 38-43.

Mallows, C.L. and Robbins, H., Some problems of optimal sampling strategy.
J. Math. Anal. Appl. 8 (1964), 90-103.

Mandl, P., Estimation and control in Markov chains.
Advances in Appl. Probability 6 (1974), 40-60.

Mandl, P., On the adaptive control of countable Markov chains.
Unpublished report (1976).

Martin, J.J., Bayesian decision problems and Markov chains.
New York etc., Wiley (1967).

Neveu, J., Mathematical foundations of the calculus of probability.
San Francisco etc., Holden-day (1965).

Neveu, J., Martingales à temps discret.
Paris, Masson (1972).

van Nunen, J.A.E.E., Contracting Markov decision processes.
Amsterdam, Math. Centre Tracts 71 (1976).

van Nunen, J.A.E.E. and Wessels, J., The generation of successive approximations for Markov decision processes by using stopping times. In: Markov decision theory, ed. H.C. Tijms and J. Wessels.
Amsterdam, Math. Centre Tracts 93 (1977), 25-38.

Parthasarathy, K., Probability measures on metric spaces.
New York etc., Academic Press (1967).

Revuz, D., Markov chains.
Amsterdam etc., North-Holland/American Elsevier (1976).

Rieder, U., Bayessche dynamische Entscheidungs- und Stoppmodelle.
University of Hamburg (1972).

Rieder, U., Bayesian dynamic programming.
Advances in Appl. Probability 7 (1975), 330-348.

Rose, J.S., Markov decision processes under uncertainty-average return criterion.
Unpublished report (1975).

Ross, S.M., Arbitrary state Markovian decision processes.
Ann. Math. Statist. 39 (1968), 2118-2122.

Ross, S.M., Applied probability models with optimization applications.
San Francisco etc., Holden-day (1970).

Satia, J.K. and Lave, R.E., Markovian decision processes with uncertain transition probabilities.
Operations Res. 21 (1973), 728-740.

Scarf, H., Bayes solutions of the statistical inventory problem.
Ann. Math. Statist. 30 (1959), 490-508.

Schäl, M., Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal.
Z. Wahrscheinlichkeitstheorie und verw. Gebiete 32 (1975), 179-196.

Shapley, L.S., Stochastic games.
Proc. Nat. Acad. Sci. USA 39 (1953), 1095-1100.

Shiryaev, A.N., On the theory of decision functions and control of an observation process with incomplete data.
Trans. Third Prague Confer. on Inform. Theory etc. (1964), 657-682.

Shiryaev, A.N., Some new results in the theory of controlled random processes.
Trans. Fourth Prague Confer. on Inform. Theory etc. (1967), 131-203.

Shreve, S.E., Dynamic programming in complete separable spaces.
University of Illinois (1977).

Strauch, R.E., Negative dynamic programming.
Ann. Math. Statist. 37 (1966), 871-890.

Sworder, D.D., Optimal adaptive systems.
New York etc., Academic Press (1966).

Taylor, H.M., Optimal replacement under additive damage and other failure models.
Naval Res. Logist. Quart. 22 (1975), 1-18.

Veinott, A.F., Optimal policy for a multi-product, dynamic, nonstationary inventory problem.
Management Sci. 12 (1965), 206-222.

Wald, A., Sequential analysis.
New York, Wiley (1947).

Waldmann, K.H., Stationäre Bayessche Entscheidungsmodelle mit Anwendungen in der Lagerhaltungstheorie.
University of Technology Darmstadt, dissertation (1976).

Wessels, J., Decision rules in Markovian decision processes with incompletely known transition probabilities.
University of Technology Eindhoven, dissertation (1968).

Wessels, J., Inventory control with unknown demand distribution: a discrete time-discrete level case.
Colloquia Mathematica Societatis J̃anos Bolyai 7 (1971), 321-334.

Wessels, J., Inventory control with unknown demand distribution: a slow-mover case.
Statistica Neerlandica 26 (1972), 243-251.

Wessels, J., Stopping times and Markov programming.
Trans. Seventh Prague Confer. on Inform. Theory etc. (1974), 575-585.

Wessels, J., Markov programming and successive approximations with respect to weighted supremum norms.
J. Math. Anal. Appl. 58 (1977), 326-335.

Whitt, W., Approximations of dynamic programs.
Manuscript (1976) (to be published).

Yuschkevich, A.A., Reduction of a controlled Markov model with incomplete data to a problem with complete information in the case of Borel state and control spaces.
Theor. Probability Appl. 21 (1976), 153-158.

Zacks, S., The theory of statistical inference.
New York, Wiley (1971).

Zacks, S. and Fennel, J., Distribution of adjusted stock levels under statistical adaptive control procedures of inventory systems.
J. Amer. Statist. Assoc. 68 (1973), 88-91.

# LIST OF SYMBOLS

| symbols | page | sentence |
|---|---|---|
| A,A | 19 | 2.1 |
| $A_n, A_n(\omega)$ | 21 | 2.6 |
| $\alpha_n(B)$, $\alpha_{i,n}(B)$ | 27 | 2.21 |
| $\mathcal{B}_a(S)$, $\mathcal{B}_m(S)$ | 55 | 3.15 |
| B | 99 | 5.4 |
| $b_k(x,q)$ | 132 | 6.11 |
| C | 99 | 5.4 |
| c(x,a) | 98 | 5.3 |
| D(x) | 19 | 2.1 |
| $\Delta(\theta,\hat{\theta})$ | 139 | 6.15 |
| $\Delta(\beta,k,h,p,q)$ | 116 | 5.25 |
| $\mathbb{E}^{\pi}_{x,q}[\cdot]$ | 21 | 2.8 |
| E(q,ε,n) | 131 | 6.10 |
| $\varepsilon_k(x,q)$ | 132 | 6.11 |
| $F_n$ | 27 | 2.18 |
| $F,\tilde{F}$ | 121 | 6.1 |
| g(q) | 77 | 4.9 |
| g(x,q,π) | 22 | 2.13 |
| h | 109 | 5.13 |
| h(x,q) | 77 | 4.9 |
| $H_n$, $H_n$, $H_\infty$ | 20 | 2.3 |
| $H$ | 20 | 2.5 |
| $H_n(q)$ | 102 | 5.9 |
| $\theta$, $\theta_i$ | 19 | 2.1 |
| I | 19 | 2.1 |
| K,K | 19 | 2.1 |
| $K_n(i,j)$ | 102 | 5.9 |
| k | 109 | 5.13 |
| L(x,θ,a) | 76 | 4.6 |
| ℓ(x,q) | 122 | 6.4 |

| symbols | page | sentence |
|---|---|---|
| $(L_f b)(x,q)$ | 122 | 6.3 |
| $m_q(i)$, $M_q(i,j)$ | 99 | 5.5 |
| $\mathbb{N}$ | 11 | 1.1 |
| ν | 19 | 2.1 |
| $\mathbb{P}^{\pi}_{x,q}$ | 21 | 2.7 |
| $\tilde{\mathbb{P}}^{\tilde{\pi}}_{x,q}$ | 44 | 3.5 |
| P | 19 | 2.1 |
| $\bar{P}_\theta$ | 20 | 2.2 |
| $P(\cdot)$ | 11 | 1.10 |
| $p_i(y\|\theta_i)$ | 19 | 2.1 |
| $p_i(y,q)$ | 32 | 2.27 |
| p | 109 | 5.13 |
| $\pi,\pi_n,\Pi$ | 20 | 2.4 |
| $\tilde{\pi},\tilde{\pi}_n,\tilde{\Pi}$ | 44 | 3.3 |
| $Q_n, Q_n(B)(\omega)$ | 28 | 2.24 |
| $Q_{i,n}$, $Q_{i,n}(B)(\omega)$ | 28 | 2.25 |
| $q_n(q,x_0,a_0,y_1,\ldots,y_n)$ | 32 | 2.30 |
| R | 99 | 5.4 |
| $\mathbb{R}$ | 11 | 1.2 |
| r(x,a,y) | 20 | 2.1 |
| $\tilde{r}(x,q,a)$ | 43 | 3.1 |
| S,S | 54 | 3.11 |
| S | 99 | 5.4 |
| s(q), $\tilde{s}(q)$ | 109 | 5.18 |
| Σ | 55 | 3.14 |
| $\Sigma_q(i,j)$ | 99 | 5.5 |
| $\tau(i,n)(\omega)$ | 27 | 2.17 |
| $T_{i,y}(q)(\cdot)$ | 32 | 2.28 |
| $t_n(q,x_0,a_0,y_1,..,y_n,x_n)$ | 44 | 3.6 |
| $(U_\tau b)(x,q)$ | 66 | 3.27 |

## LIST OF SYMBOLS

| symbol | page | sentence |
|---|---|---|
| $v(x,q,\pi)$, $v(x,q)$ | 22 | 2.12 |
| $\hat{v}(x,q)$ | 110 | 5.19 |
| $\varphi_n(\theta,f)$, $\varphi_\infty(\theta,f)$ | 122 | 6.5 |
| $\phi(x,\theta,a)$, $\phi(x,q,a)$ | 77 | 4.8 |
| $w$, $W$ | 21 | 2.9 |
| $W_n(q)$ | 149 | 7.1 |
| $w(x,q)$ | 122 | 6.4 |
| $\psi$ | 55 | 3.13 |
| $x$, $X$ | 19 | 2.1 |
| $x_n$, $x_n(\omega)$ | 21 | 2.6 |
| $y$, $Y$ | 19 | 2.1 |
| $Y_n$, $Y_n(\omega)$ | 21 | 2.6 |
| $z$, $z(\omega)$ | 21 | 2.6 |
| $\Omega$ | 20 | 2.5 |