HOW DOES THE SOLUTION OF A MARKOV DECISION PROCESS DEPEND ON THE TRANSITION PROBABILITIES?

ALFRED MÜLLER

Abstract

The present work deals with the comparison of the optimal value functions of (discrete time) Markov decision processes (MDPs), which differ only in their transition probabilities. We show that the solution of an MDP is monotone with respect to appropriately defined stochastic order relations. We also find conditions for continuity with respect to suitable probability metrics. The results are applied to some well known examples, including inventory control and optimal stopping.

Key words. Integral stochastic orders, probability metrics, Markov decision processes, sensitivity analysis.
AMS 1991 subject classification. Primary: 90C31, 90C39, 90C40; Secondary: 60B10, 60E15, 90B05.

1 Introduction.

It is well known that only a few simple Markov Decision Processes (MDPs) admit an “explicit” solution. Realistic models, however, are mostly too complex to be computationally feasible. Consequently, there is a continued interest in finding good approximations, even in times of rapidly growing computer power. The problem is to find approximate models with solutions that differ only slightly from the solution of the original problem. To handle this problem it is necessary to evaluate the difference of the solutions of two MDPs. In this paper we investigate the impact of replacing the transition probability distribution by another one. This is of special interest for several reasons. First of all, the distribution is very often unknown and has to be estimated by statistical methods. Furthermore, the distribution is sometimes taken from some parametric family, which exhibits an explicit solution. Last not least, the use of computers for numerical calculations requires discretizations.

There is a vast literature on approximations by discretization of state and action space. For a bibliography see Morin (1978). Most quantitative approaches involve the distance between the transition probabilities in total variation norm and require boundedness of the value functions, see e.g. Hinderer (1979) and Whitt (1978,1979). Our approach is more general, but also quantitative. We use the theory of so called *integral probability metrics* in combination with structural properties of the value functions. This includes the total variation as a special case. But to our experience, there are much more suitable metrics like the Kantorovich metric. An interesting qualitative investigation is given in Langen (1981), who utilizes the notion of *continuous convergence*.

In addition, we use integral stochastic orderings to prove a general monotonicity result. To our knowledge, this question has so far only been dealt with in the special case of convex order, see Hernandez-Lerma and Runggaldier (1993) and Rieder and Zagst (1994).

We restrict our attention to finite horizon problems, but most of our results can be extended to the case of infinite horizon by using the results of Schäl (1975).

Our paper is organized as follows. In the next section we give a formal characterization of the Markov Decision Process to be studied. We also introduce the concept of bounding functions and state a well known structure theorem, fundamental for the subsequent analysis. Section 3 is devoted to the theory of integral stochastic orders and probability metrics. We collect the most important facts of these theories and give several examples, which are relevant for our application. Section 4 contains the main results. We prove monotone and continuous dependency of the value function on the transition probability measures. In section 5 we apply these results to two classical models of MDPs. First we show how the solution of a problem of inventory control with setup costs depends on the demand distribution. The second example deals with the well known problem of optimal stopping, also known as secretary problem or job search problem. We will see that the solution of this problem is a monotone and continuous function of the distribution of the offers, if we use appropriate stochastic orders and probability metrics.

2 The Markov Decision Model.

Now we give a formal definition of our model for a (finite horizon, discrete time) Markov decision process. Similar models can be found in Hinderer (1970), Schäl (1975), Bertsekas and Shreve (1977) or Puterman (1994).

Definition 2.1 *A Markov decision process MDP is given by a tupel $(S, A, D, P, \tilde{r}, V_0, \beta)$ with the following meaning.*

1. *S and A are arbitrary nonempty sets, endowed with σ-algebras $\mathcal{S}$ and $\mathcal{A}$, respectively. S is the* **state space** *and A the* **action space**.
2. *$D \in \mathcal{S} \otimes \mathcal{A}$ is the* **restriction set**. *We assume that D contains the graph of a measurable mapping $f : S \to A$. $D(s) := \{a \in A : (s, a) \in D\}$ is the set of admissible actions if the system is in state s.*
3. *P is a transition probability measure from D into S. $P(s, a, \cdot)$ is the distribution of the state visited next if the system is in state s and action a is taken.*
4. *$\tilde{r}$ is a measurable mapping from $D \times S$ to $\mathbb{R}$ with the property that*
$$P\tilde{r}(s, a) := \int P(s, a, ds')\ \tilde{r}(s, a, s')$$
exists for all $(s, a) \in D$. $\tilde{r}$ is called **reward function**.
5. *The* **terminal reward function** *V_0 is a measurable mapping from S to $\mathbb{R}$.*
6. *$\beta \in \mathbb{R}_{>0}$ is the* **discount factor**.

Remark: Normally it is sufficient to consider the *reduced reward function* $r(s, a) := \int P(s, a, ds')\tilde{r}(s, a, s')$, but we want to compare MDPs which differ in their transition probabilities, so we must take into account that different transition probabilities P and Q lead to different reduced reward functions r_P and r_Q respectively.

Now we define the optimization problem.

Definition 2.2 *a) A measurable mapping $f : S \to A$ with graph $f \subset D$ is called* **decision rule**. *The set of all decision rules is denoted by F.*
*b) A sequence $\pi := (\phi_\nu)_{\nu=0}^{N-1}$ of $N \in \mathbb{N}$ decision rules is called N-***stage policy**.

To a given state $s \in S$ and a policy $\pi = (\phi_\nu)_{\nu=0}^{N-1} \in F^N$ we define on the measure space $(S^N, \otimes_1^N \mathcal{S})$ the canonical probability measure $P_{N\pi}(s, \cdot)$:
For $C \in \otimes_1^N \mathcal{S}$ let

$$P_{N\pi}(s, C) := \int P_{\phi_0}(s, ds_1) \int P_{\phi_1}(s_1, ds_2) ... \int P_{\phi_{N-1}}(s_{N-1}, ds_N) 1_C(s_1, ..., s_N),$$

where $P_f(s, \cdot) := P(s, f(s), \cdot)$ for $f \in F$.

We define the (projection) random variables ζ_ν on S^N as

$$\zeta_\nu(s_1, ..., s_N) := s_\nu, \quad \nu = 1, ..., N.$$

Then $Y := (\zeta_1, ..., \zeta_N)$ is by construction of $P_{N\pi}$ a Markov chain. Defining $\zeta_0 := s$, the **total reward** $R_{N\pi}(s, Y)$ is given as

$$R_{N\pi}(s, Y) := \sum_{\nu=0}^{N-1} \beta^\nu \cdot \tilde{r}(\zeta_\nu, \phi_\nu(\zeta_\nu), \zeta_{\nu+1}) \ + \beta^N V_0(\zeta_N)$$

and, if the following integral exists, the **expected total reward** $V_{N\pi}(s)$ is given as

$$V_{N\pi}(s) := \int P_{N\pi}(s, dy) R_{N\pi}(s, y), \quad s \in S.$$

If $V_{N\pi}(s)$ exists for all $\pi \in F^N$ and $s \in S$, then the mapping

$$s \mapsto V_N(s) := \sup_{\pi \in F^N} V_{N\pi}(s)$$

is called the **value function** (for horizon N). When we want to emphasize the dependency on the transition probability distribution P we write V_n^P, $V_{n\pi}^P$, etc.

A policy π^* is ε-**optimal** for $\varepsilon \geq 0$ if $V_{N\pi^*}(s) \geq V_N(s) - \varepsilon$ for all $s \in S$. A 0-optimal policy is simply called optimal.

Before we collect some important facts about MDP's, we define some useful abbreviations. Let $\mathfrak{V}_0$ be the set of all functions $v : S \to \mathbb{R}$, which are $P(s, a, \cdot)$-integrable for all $(s, a) \in D$. For $v \in \mathfrak{V}_0$ we introduce the Markov operator

$$Pv(s, a) := \int P(s, a, ds') \ v(s')$$

and we use the same notation $Pf(s,a) := \int P(s,a,ds')\ f(s,a,s')$ also for functions $f : D \times S \to \mathbb{R}$, if $f(s,a,\cdot) \in \mathfrak{V}_0$ for all $(s,a) \in D$.

Now we define for an arbitrary MDP the operators L, U_f and U on $\mathfrak{V}_0$:

$Lv(s,a) := P\tilde{r}(s,a) + \beta Pv(s,a), \quad (s,a) \in D,$

$U_f v(s) := Lv(s, f(s)), \quad s \in S,\ f \in F,$

$Uv(s) := \sup_{a \in D(s)} Lv(s,a) = \sup_{f \in F} U_f v(s), \quad s \in S.$

Finally, a decision rule $f \in F$ is called **ε-maximizer** of Lv, if $U_f v(s) \geq Uv(s) - \varepsilon$ for all $s \in S$. Using these notations we can define the fundamental recursive scheme of MDPs, the so called value iteration, in the following short form.

Definition 2.3 *For a MDP the* **value iteration (VI)** *holds, if V_n exists for all $n \in \mathbb{N}$, belongs to $\mathfrak{V}_0$ and fulfills $V_n = UV_{n-1}$.*

The following conditions for the existence of $V_{N\pi}$ and V_N are well known and can be found e.g. in Wessels (1977) or Puterman (1994), p. 231ff.

Definition 2.4 *A measurable function $b : S \to [1, \infty)$ is a* **bounding function** *for a MDP, if there is a constant $\delta > 0$ such that the following holds:*

(i) $\int P(s,a,ds')\ |\tilde{r}(s,a,s')| \leq \delta \cdot b(s)$ *for all* $(s,a) \in D$.

(ii) $|V_0(s)| \leq \delta \cdot b(s)$ *for all* $s \in S$.

(iii) $\int P(s,a,ds')\ b(s') \leq \delta \cdot b(s)$ *for all* $(s,a) \in D$.

Remarks: 1. Mostly one only demands $b \geq 0$ for a bounding function, but the requirement $b \geq 1$ is no real restriction. If $b \geq 0$ is a bounding function, then $b+1$ is also one (with respect to the constant $1+\delta$).

2. We define the weighted supremum norm

$$\|f\|_b := \sup_{s \in S} \frac{|f(s)|}{b(s)},$$

and we denote the set of all measurable functions with $\|f\|_b < \infty$ by $\mathfrak{B}_b$. Then b is a bounding function, if $P\tilde{r}(\cdot,a),\ a \in A$ and V_0 are in $\mathfrak{B}_b$ and if the Markov operator P maps $\mathfrak{B}_b$ into $\mathfrak{B}_b$.

3. If $\tilde{r}$ and V_0 are bounded, then $b \equiv 1$ is a bounding function.

If a MDP has a bounding function then it is easy to see that $V_{N\pi}$ and V_N exist. In fact, we have the following result.

Theorem 2.5 *If MDP has a bounding function b, then $V_{N\pi}(s)$ and $V_N(s)$ exist for all $N \in \mathbb{N}$, $\pi \in F^N$ and $s \in S$, and $\|V_n\|_b < \infty$ for all $n \in \mathbb{N}$.*

Now we are able to formulate an important tool for proving the validity of the value iteration. It seems to be due to Porteus (1975). Similar results can also be found in Dynkin/Yushkevich (1979), p. 57, and Puterman (1994), p. 225ff. From now on we will refer to it as the *structure theorem.*

Theorem 2.6 *Assume that MDP has a bounding function b and there is a set of functions $\mathfrak{V} \subset \mathfrak{B}_b$ with the following properties:*
(S1) For all $v \in \mathfrak{V}$ and $\varepsilon > 0$ there is an ε-maximizer of Lv.
(S2) For all $v \in \mathfrak{V}$ we have $Uv \in \mathfrak{V}$.
(S3) $V_0 \in \mathfrak{V}$.

Then the following holds:
a) $V_n \in \mathfrak{V}$ for all $n \in \mathbb{N}$.
b) The value iteration holds: $V_n = UV_{n-1}$, $n \in \mathbb{N}$.
c) For all $N \in \mathbb{N}$ and $\varepsilon > 0$ there is an ε-optimal policy for MDP.
d) If $f(v)$ is an ε-maximizer of Lv, $v \in \mathfrak{V}$, then $(f(V_{n-1}))_N^1$ is a δ_N-optimal policy for MDP, where

$$\delta_N := \varepsilon \cdot \sum_{i=0}^{N-1} \beta^i, \quad N \in \mathbb{N}.$$

3 Stochastic Orders and Probability Metrics.

The main objective of this paper is to show monotonicity and continuity results of the functions $P \to V_n^P$. For this we need appropriate concepts of stochastic order relations and probability metrics. It turns out that so called *integral* stochastic orders and probability metrics are best suited for our purpose. The basic idea is to use the bounding function and the function classes $\mathfrak{V}$ of Theorem 2.6 for the definition.

For a given bounding function b we denote by $\mathbb{P}_b$ the set of all probability measures P with $\int b \, dP < \infty$. It is easy to see that $\int f \, dP$ exists for all $f \in \mathfrak{B}_b$ and $P \in \mathbb{P}_b$. Hence all integrals in the following definition exist.

Definition 3.1 *Let b be a bounding function and $\mathfrak{F} \subset \mathfrak{B}_b$ an arbitrary set of b-bounded functions. Then we define on $\mathbb{P}_b$*

a) the **integral stochastic order** $\leq_{\mathfrak{F}}$ *by*

$$P \leq_{\mathfrak{F}} Q \quad \textit{iff} \quad \int f\, dP \;\leq\; \int f\, dQ \quad \textit{for all } f \in \mathfrak{F}.$$

b) the **integral probability metric** $d_{\mathfrak{F}}$ *by*

$$d_{\mathfrak{F}}(P,Q) \;:=\; \sup_{f \in \mathfrak{F}} \left| \int f\, dP - \int f\, dQ \right|.$$

Remarks: 1. Obviously $\leq_{\mathfrak{F}}$ is reflexive and transitive. Hence it is a pre-order. If $\mathfrak{F}$ separates points in $\mathbb{P}_b$, i.e. if $\int f\, dP = \int f\, dQ\ \forall f \in \mathfrak{F}$ implies $P = Q$, then $\leq_{\mathfrak{F}}$ is also antisymmetric, hence a (partial) order.

2. A general theory of integral stochastic orders can be found in Stoyan (1983), Marshall (1991) or Müller (1996a). Several examples are given below.

3. $d_{\mathfrak{F}}$ is non-negative, symmetric and fulfils the triangle inequality. Hence it is a semimetric. It is a metric, if $\mathfrak{F}$ separates points in $\mathbb{P}_b$. As usual in the theory of probability metrics, we allow $d_{\mathfrak{F}}$ to assume infinite values, see e.g. Rachev (1991), p. 10ff.

4. Integral probability metrics are sometimes called *metrics with a ζ-structure*, see e.g. Zolotarev (1983). Many examples and properties of these metrics are given there, in Rachev (1991) and in Müller (1996b).

5. Sometimes it is more convenient to formulate properties of stochastic orders and probability metrics in terms of random variables X, Y or distribution functions F, G. The meaning of notations like $F \leq_{\mathfrak{F}} G$ or $d_{\mathfrak{F}}(X, Y)$ should be obvious.

There may be different classes of functions, which generate the same order (metric). For checking $P \leq_{\mathfrak{F}} Q$ and evaluating $d_{\mathfrak{F}}(P,Q)$ it is desirable to have "small" generators. For our applications in the next section, however, we are interested in "large" generators. The *maximal generators* have been characterized in Müller (1996a,b). We do not need these characterizations here. For our applications it is sufficient to know the following facts. We omit the easy proof.

Theorem 3.2 *a) If $\mathfrak{F}$ is an arbitrary generator of an integral stochastic order, then the convex cone spanned by $\mathfrak{F}$ generates the same stochastic order.*

b) If $\mathfrak{V}$ is an arbitrary generator of an integral probability metric, then the balanced convex hull spanned by $\mathfrak{V}$ generates the same probability metric.

The structural properties of the value function that can be proven most often are *monotonicity, convexity* and *Lipschitz continuity.* This is due to the fact that these properties are preserved under the typical operations in the value iteration, namely under mixture, addition and formation of suprema. Some general results about these structures of the value functions are given in Hinderer (1984).

Therefore we are especially interested in orders and metrics with generators $\mathfrak{V}$ consisting of functions with some of these properties.

Example 3.1. The most important integral stochastic orders for our purpose are the following well known relations, see e.g. Stoyan (1983) or Shaked and Shanthikumar (1994).
a) Let P and Q be probability measures on an arbitrary ordered space S. Then P is said to be *stochastically smaller* then Q (written $P \leq_{st} Q$), if $\int f\ dP \ \leq\ \int f\ dQ$ for all measurable bounded increasing functions $f : S \to \mathbb{R}$.
b) Let P and Q be probability measures with finite expectation on some euclidian space S. Then P is said to be smaller than Q in *increasing convex order* (written $P \leq_{ic} Q$), if $\int f\ dP \ \leq\ \int f\ dQ$ for all increasing convex functions $f : S \to \mathbb{R}$, for which the integrals exist.
c) Let P and Q be probability measures with finite expectation on some euclidian space S. Then P is said to be smaller than Q in *convex order* (written $P \leq_c Q$), if $\int f\ dP \ \leq\ \int f\ dQ$ for all convex functions $f : S \to \mathbb{R}$, for which the integrals exist.

Example 3.2. The most interesting integral probability metrics with applications to our object are the following ones.
a) The presumably best known probability metric is the *Kolmogorov distance* $\rho(F,G) := \sup_{t\in\mathrm{R}} |F(t) - G(t)|$. Since $F(t) = \int 1_{(-\infty,t]}\ dF$, this metric is generated by the integrals of the set of functions $\mathfrak{V} = \{1_{(-\infty,t]},\ t \in \mathbb{R}\}$.
b) The *total variation metric* $\sigma(P,Q) := \sup_{A\in\mathcal{S}} |P(A) - Q(A)|$ is also an integral probability metric. One has to choose $\mathfrak{V}$ as the set of all indicator functions of measurable sets.
c) Let (S,d) be a metric space and let $\mathfrak{L}_1$ be the set of all Lipschitz functions

$f: S \to \mathbb{R}$ with Lipschitz-constant 1, i.e. the set of all functions with

$$\|f\|_L := \sup_{s \neq t} \frac{|f(s) - f(t)|}{d(s,t)} \;\; \leq \; 1.$$

The integral probability metric ζ generated by $\mathfrak{L}_1$ is called *Kantorovich metric*. In case $S = \mathbb{R}$ the Kantorovich metric can easily be evaluated by the following well known formula (see e.g. Dudley (1989), p. 333):

$$\zeta(X,Y) = \int |F_X(t) - F_Y(t)| \; dt.$$

d) Another interesting probability metric has been defined by Rachev and Rüschendorf (1990). They introduce the so called *stop-loss metric* d_{sl} for real-valued random variables with finite expectation. It is given by

$$d_{sl}(X,Y) := \sup_{t \in \mathbb{R}} |E(X-t)_+ - E(Y-t)_+|,$$

where $x_+ := \max\{x, 0\}$. This is obviously an integral probability metric generated by the functions $x \mapsto (x-t)_+$, $t \in \mathbb{R}$. A larger generator of this metric is given by the set of all increasing convex functions with $\|f\|_L \leq 1$.

4 Main results.

The theory of integral stochastic orders can now be used to show that the value function of a MDP depends in a monotonic way on the transition probabilities.

Theorem 4.1 *Let MDP(P) and MDP(Q) be two Markov Decision Processes, which differ only in their transition probabilities P and Q, respectively. Assume that there is a bounding function b and a class $\mathfrak{V}$ of functions, such that the assumptions of Theorem 2.6 are fulfilled for both MDPs. Assume further, that for all $(s,a) \in D$ we have*

(i) $P(s,a,\cdot) \leq_{\mathfrak{V}} Q(s,a,\cdot)$,

(ii) $\tilde{r}(s,a,\cdot) \in \mathfrak{V}$.

Then $V_n^P(s) \;\leq\; V_n^Q(s)$ for all $n \in \mathbb{N}$ and $s \in S$.

Moreover, if f_n is an ε_n-maximizer of LV_{n-1}^P, $n \in \mathbb{N}$, and $\pi := (f_n)_N^1$, then

$$V_{N\pi}^Q(s) \;\geq\; V_N^P(s) - \delta_N, \quad n \in \mathbb{N}, \tag{4.1}$$

where for δ_n, $n \in \mathbb{N}$, the following recursion holds:
$\delta_1 := \varepsilon_1$, $\delta_{n+1} := \varepsilon_{n+1} + \beta\delta_n$.

Proof. We proceed by induction on $n \in \mathbb{N}$. For $n = 0$ there is nothing to show, as $V_0^P = V_0^Q$ by assumption. Hence assume that $V_n^P(s) \leq V_n^Q(s)$ holds for all $s \in S$. By Theorem 2.6 we have $V_n^P \in \mathfrak{V}$ and thus (ii) implies that the function

$$W_n^P(s, a, \cdot) := \tilde{r}(s, a, \cdot) + \beta V_n^P(\cdot)$$

is in the convex cone spanned by $\mathfrak{V}$. Combining this with the induction hypothesis and (i) we get

$$\begin{aligned} V_{n+1}^P(s) &= \sup_{a \in D(s)} \left\{ \int P(s, a, ds')\ W_n^P(s, a, s')) \right\} \\ &\leq \sup_{a \in D(s)} \left\{ \int Q(s, a, ds')\ W_n^P(s, a, s')) \right\} \\ &\leq \sup_{a \in D(s)} \left\{ \int Q(s, a, ds')\ W_n^Q(s, a, s')) \right\} = V_{n+1}^Q(s), \end{aligned}$$

when taking into consideration that, by Theorem 3.2, the convex cone spanned by $\mathfrak{V}$ generates the same integral stochastic order as $\mathfrak{V}$.

To prove the second part of the theorem, let f_1 be an ε_1-maximizer of LV_0 in MDP(P). Then (i) and (ii) imply

$$\begin{aligned} V_{1f}^Q(s) &= \int Q_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta V_0(s')] \\ &\geq \int P_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta V_0(s')] \\ &= V_{1f}^P(s)\ \geq\ V_1^P(s) - \varepsilon_1. \end{aligned}$$

Hence the assertion holds for $N = 1$. Now assume the induction hypothesis holds for $\sigma := (f_n)_N^1 \in F^N$, i.e. $V_{N\sigma}^Q(s)\ \geq\ V_N^P(s) - \delta_N$, and let $f := f_{N+1}$ be an ε_{N+1}-maximizer of LV_N^P. Then we get for $\pi := (f, \sigma)$

$$\begin{aligned} V_{N+1,\pi}^Q(s) &= \int Q_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta V_{N\sigma}^Q(s')] \\ &\geq \int Q_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta \cdot (V_N^P(s') - \delta_N)] \\ &= -\beta\delta_N + \int Q_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta V_N^P(s')] \\ &\geq -\beta\delta_N + \int P_f(s, ds')\ [\tilde{r}(s, f(s), s') + \beta V_N^P(s')] \\ &\geq -\beta\delta_N + V_{N+1}^P(s) - \varepsilon_{N+1}, \quad \text{(since } f \text{ is an } \varepsilon_{N+1}\text{-maximizer)} \\ &= V_{N+1}^P(s) - (\varepsilon_{N+1} + \beta\delta_N). \end{aligned}$$

□

Remarks: 1. If in the second part of Theorem 4.1 $\varepsilon_n = 0$ for all n, then π is an optimal policy for MDP(P) and $V^Q_{N\pi} \geq V^P_N$.
2. If $\tilde{r}(s, a, s')$ does not depend on s', then we can dispense with assumption (ii).

Very often it is difficult to specify the transition probability measure P or it is impossible to evaluate V^P_n. Then the above theorem can be applied in the following way:
Let Q_1 and Q_2 be "lower" and "upper" bounds for P, i.e.

$$Q_1(s, a, \cdot) \leq_{\mathfrak{V}} P(s, a, \cdot) \leq_{\mathfrak{V}} Q_2(s, a, \cdot),$$

and assume that assumption (ii) of Theorem 4.1 is fulfilled. If we can evaluate $V^{Q_1}_n$ and $V^{Q_2}_n$ explicitly, then we have lower and upper bounds for V^P_n. Moreover, by the second statement of the theorem, a "good" policy for MDP(Q_1) or MDP(Q_2) is also a "good" policy for MDP(P).

Next we want to show how it is possible to use integral probability metrics for sensitivity analysis of Markov Decision Processes. It can happen that the value function V_n is not in a class $\mathfrak{V}$, which defines a useful probability metric. But it may be that there is some constant c with $V_n/c \in \mathfrak{V}$. Notice that here we can not assume w.l.o.g. $\mathfrak{V}$ to be a convex cone, we can only assume $\mathfrak{V}$ to be balanced and convex. We define the Minkowski functional (see e.g. Rudin (1973))

$$\mu_{\mathfrak{V}}(f) := \inf\{t > 0 : t^{-1} f \in \mathfrak{V}\},$$

and $[\mathfrak{V}]$ shall be the vector space spanned by $\mathfrak{V}$. If $\mathfrak{V}$ is balanced (which is always the case if $\mathfrak{V}$ is the maximal generator), then $[\mathfrak{V}]$ is the set of all functions with $\mu_{\mathfrak{V}}(f) < \infty$. It is easy to see, that

$$\left| \int f \, dP - \int f \, dQ \right| \ \leq \ \mu_{\mathfrak{V}}(f) \cdot d_{\mathfrak{V}}(P, Q). \tag{4.2}$$

For many familiar integral probability metrics this Minkowski functional leads to well known expressions if we consider the maximal generators of the metrics. From Müller (1996b) the following examples can be deduced.

Examples: 1. For the Kolmogorov metric we get $\mu_{\mathfrak{V}}(f) = \mathrm{Var}(f)$, where $\mathrm{Var}(f)$ is the total variation of the function f.
2. For the total variation metric we have $\mu_{\mathfrak{V}}(f) = \sup(f) - \inf(f)$.
3. For the Kantorovich metric obviously $\mu_{\mathfrak{V}}(f) = \|f\|_L$.

In the following result we give conditions for a Markov Decision Process to imply that the value function depends continuously on the transition probabilities.

Theorem 4.2 *Let MDP(P) and MDP(Q) be two Markov Decision Processes, which differ only in their transition probabilities P and Q, respectively. Assume that for both of them the value iteration holds and that there is a class of functions $\mathfrak{V}$ with $V_n^P \in [\mathfrak{V}]$, $n \in \mathbb{N}_0$. We define the functions*

$$\eta(s) := \sup_{a \in D(s)} |P\tilde{r}(s,a) - Q\tilde{r}(s,a)|, \quad s \in S,$$
$$\delta(s) := \sup_{a \in D(s)} \delta(s,a) := \sup_{a \in D(s)} d_{\mathfrak{V}}(P(s,a,\cdot), Q(s,a,\cdot)), \quad s \in S,$$

and for functions $v : S \to \mathbb{R}$ the operator

$$H_Q v(s) := \sup_{a \in D(s)} |Qv(s,a)|, \quad s \in S.$$

Then we have for all $n \in \mathbb{N}_0$ and $s \in S$:

$$|V_n^P(s) - V_n^Q(s)| \le g_n(s),$$

where g_n satisfies the recursion

$$g_0(s) := 0, \quad g_{n+1}(s) := \eta(s) + \beta \cdot \mu_{\mathfrak{V}}(V_n^P) \cdot \delta(s) + \beta \cdot H_Q g_n(s). \qquad (4.3)$$

Proof. We proceed by induction on n. For $n = 0$ the assertion is trivial. Hence assume $|V_n^P(s) - V_n^Q(s)| \le g_n(s)$. Since $|\sup f - \sup g| \le \sup |f - g|$, we can deduce for all $s \in S$:

$$|V_{n+1}^P(s) - V_{n+1}^Q(s)|$$
$$= \Big| \sup_{a \in D(s)} \{P\tilde{r}(s,a) + \beta P V_n^P(s,a)\} - \sup_{a \in D(s)} \{Q\tilde{r}(s,a) + \beta Q V_n^Q(s,a)\}\Big|$$
$$\le \sup_{a \in D(s)} \Big|P\tilde{r}(s,a) + \beta P V_n^P(s,a) - Q\tilde{r}(s,a) - \beta Q V_n^Q(s,a)\Big|$$
$$\le \sup_{a \in D(s)} |P\tilde{r}(s,a) - Q\tilde{r}(s,a)| \; + \; \beta \sup_{a \in D(s)} |P V_n^P(s,a) - Q V_n^Q(s,a)|$$
$$\le \eta(s) + \beta \sup_{a \in D(s)} \underbrace{|P V_n^P(s,a) - Q V_n^P(s,a)|}_{\le \mu_{\mathfrak{V}}(V_n^P) \cdot \delta(s,a)} + \beta \sup_{a \in D(s)} \underbrace{|Q V_n^P(s,a) - Q V_n^Q(s,a)|}_{\le Q g_n(s,a)}$$
$$\le \eta(s) + \beta \mu_{\mathfrak{V}}(V_n^P) \delta(s) + \beta H_Q g_n(s) \; = \; g_{n+1}(s).$$

□

Remarks: 1. Very often the class $\mathfrak{V}$ in Theorem 4.2 does not coincide with the class of functions in Theorem 2.6. This is why we included here the value iteration as an assumption and did not refer to Theorem 2.6.

2. If we have $\eta \in [\mathfrak{V}]$, $\delta \in [\mathfrak{V}]$ and $H_Q([\mathfrak{V}]) \subset [\mathfrak{V}]$, then g_n is finite and $g_n \in [\mathfrak{V}]$. This follows easily from (4.3) and the sublinearity of the Minkowski functional.

5 Applications.

A. Inventory Control with Setup Costs.

One of the best known applications of the theory of MDPs is inventory control. Especially interesting is the optimization of models with setup costs. In their seminal work, Scarf (1960), Veinott (1966) and Schäl (1976) gave conditions for the optimality of structured policies, so called (s, S)-policies. In practice, deterministic models are often preferred since most practitioners are uncertain how to determine the distribution of the (random) demand, and how sensitive the solution of the problem is to errors in the elicitation of this distribution.

It will now be shown that the value function of the corresponding MDP depends continuously (with respect to the Kantorovich metric) on the distribution of the demands. We will restrict our investigation to an easy model with nonstationary data and proportional holding and back order costs. It should be mentioned, however, that these assumptions are only made to keep the notation simple. The results can easily be extended to more complicated models. A detailed examination of our proof shows that the only crucial assumption we need is Lipschitz continuity of the cost functions.

We consider the following model (cp. Heyman and Sobel (1984), p. 306ff): Let s_ν be the inventory level at the beginning of period ν and let a_ν be the inventory level after ordered goods (if any) are delivered. Therefore $a_\nu - s_\nu$ is the ordered quantity. Let $D_1, ..., D_N$ be the i.i.d. demands with probability distribution P. We assume that excess demand is backlogged, hence $s_{\nu+1} = a_\nu - D_\nu$.

We assume setup costs $K > 0$, inventory costs c_1 per unit and back order costs c_2 per unit. Summing up, we get the one period cost function

$$\tilde{c}(s,a,d) := K \cdot H(a-s) + c_1 \cdot (a-d)_+ + c_2 \cdot (d-a)_+,$$

where H is the Heavyside function, i.e. $H(u) = 1$ if $u > 0$ and $H(u) = 0$ otherwise. The terminal reward shall be $V_0(s) \equiv 0$.

The solution of this problem is given by the value iteration

$$V_{n+1}^P(s) \;=\; \inf_{a \geq s} \left\{ \int P(dx)\; [\tilde{c}(s,a,x) + \beta \cdot V_n^P(a-x)] \right\}, \quad n \in \mathbb{N}_0, \qquad (5.1)$$

and it is well known that in this case an optimal policy of (s,S)-type exists. The proof uses a version of the structure theorem 2.6 with $\mathfrak{V}$ the set of all K-convex functions, see e.g. Scarf (1960). But this class of functions is not useful for our purposes. If we define a stochastic order relation $\leq_K$ generated by the K-convex functions, then $P \leq_K Q$ implies $P \leq_{ic} Q$ as well as $P \geq_{st} Q$. This follows from the fact that all increasing convex functions are K-convex and that all decreasing functions with range $[0,K]$ are K-convex. Hence $P \leq_K Q$ iff $P = Q$.

Therefore we have to look for another structural property of the value functions V_n. It turns out that one can show that V_n is Lipschitz continuous and hence we can apply Theorem 4.2 with $\mathfrak{V} = \mathfrak{L}_1$. We need the following properties of the Lipschitz functional $\|\cdot\|_L$. The proof is straightforward and therefore omitted.

Lemma 5.1 *Let $\mathfrak{L}$ be the set of all Lipschitz functions on the real line. Then the following holds:*

a) $\|\cdot\|_L$ is a seminorm.

b) $f, g \in \mathfrak{L} \;\Rightarrow\; f \circ g \in \mathfrak{L}$ and $\|f \circ g\|_L \leq \|f\|_L \cdot \|g\|_L$.

c) If I is an arbitrary index set and $f_i \in \mathfrak{L}$ for all $i \in I$, then

$$\|\sup f_i\|_L \leq \sup \|f_i\|_L \quad \textit{and} \quad \|\inf f_i\|_L \leq \sup \|f_i\|_L.$$

d) If $f \in \mathfrak{L}$ and $g(x) := \inf_{t \geq x} f(t)$ is finite, then $\|g\|_L \leq \|f\|_L$.

e) Let $(\Omega, \mathfrak{A}, P)$ be an arbitrary probability space and let $f : \Omega \times \mathbb{R} \to \mathbb{R}$ be such that $\|f(\omega,\cdot)\|_L \leq \alpha$ for all $\omega \in \Omega$. Then $g(s) := \int P(d\omega)\; f(\omega,s)$ fulfils $\|g\|_L \leq \alpha$.

Using these properties of $\|\cdot\|_L$, the next Lemma follows easily by induction.

Lemma 5.2 *The value function V_n^P defined in equation (5.1) is Lipschitz continuous with $\|V_n^P\|_L \leq \gamma \cdot \sigma_n(\beta)$, where $\gamma := \max\{c_1, c_2\}$ and*

$$\sigma_n(\beta) := \sum_{i=0}^{n-1} \beta^i.$$

Now we are ready to prove the main result. Though we deal with a cost minimization problem here, we will apply Theorem 4.2, which is formulated for maximization problems. This makes no difficulties since it is well known that cost minimization problems can be transferred into maximization problems by regarding costs as negative rewards.

Theorem 5.3 *Let P and Q be two different demand distributions with finite mean. Then*

$$|V_n^P(s) - V_n^Q(s)| \;\leq\; \alpha_n \cdot \zeta(P,Q),$$

where α_n satisfies the following recursion:
$\alpha_0 := 0,\ \ \alpha_{n+1} := \beta\alpha_n + \gamma \cdot \sigma_{n+1}(\beta)$.

Proof. We will apply Theorem 4.2 with $\mathfrak{V} = \mathfrak{L}_1$. Then $d_{\mathfrak{V}}$ is the Kantorovich metric ζ and for the Minkowski functional we get $\mu_{\mathfrak{V}}(f) = \|f\|_L$. We have to define $P(s,a,B) := P(a-B),\ B \in \mathcal{S}$ and $P\tilde{r}(s,a) := -\int P(dx)\ \tilde{c}(s,a,x)$. The corresponding quantities in MDP(Q) shall be defined analogously.

Next we want to give an upper bound for $\eta(s)$. Applying (4.2) yields

$$\begin{aligned}\eta(s) \;&:=\; \sup_{a\geq s}\left|\int P(dx)\ \tilde{c}(s,a,x) - \int Q(dx)\ \tilde{c}(s,a,x)\right| \\ &\leq\; \sup_{a\geq s}\Big|\|\tilde{c}(s,a,\cdot)\|_L \cdot \zeta(P,Q)\Big| \;\leq\; \gamma \cdot \zeta(P,Q). \qquad (5.2)\end{aligned}$$

As the Kantorovich metric is invariant under translations, we get

$$d_{\mathfrak{V}}(P(s,a,\cdot), Q(s,a,\cdot)) = d_{\mathfrak{V}}(P,Q) = \zeta(P,Q)$$

for all $(s,a) \in D$. Hence $\delta(s) = \zeta(P,Q)$ for all $s \in S$.

Now we are ready to prove the assertion of the theorem by induction on n. The case $n = 0$ is trivial. Hence assume $|V_n^P(s) - V_n^Q(s)| \;\leq\; g_n(s) \;\leq$

$\alpha_n \cdot \zeta(P, Q)$. Inserting (5.2) and the result of Lemma 5.2 into (4.3) yields

$$\begin{aligned} g_{n+1}(s) &:= \eta(s) + \beta \cdot \mu_{\mathfrak{V}}(V_n^P) \cdot \delta(s) + \beta \cdot H_Q g_n(s) \\ &\leq \gamma \cdot \zeta(P, Q) + \beta \cdot \|V_n^P\|_L \cdot \zeta(P, Q) + \beta \cdot \alpha_n \cdot \zeta(P, Q) \\ &\leq \zeta(P, Q) \cdot (\gamma + \beta\gamma\sigma_n(\beta) + \beta\alpha_n) \;=\; \zeta(P, Q) \cdot \alpha_{n+1}. \end{aligned}$$

In the first inequality we made use of the monotonicity of the operator H_Q.

□

B. Optimal Stopping.

Let $X_1, ..., X_N$ be a sequence of i.i.d. random variables with distribution P, which can be observed sequentially at a cost c per observation. If the decision maker stops after the kth observation, he receives an immediate reward of $\max\{X_1, ..., X_k\}$. We are looking for an optimal stopping rule. This is a familiar problem of optimal stopping that can be solved by backward induction, see e.g. Chow, Robbins and Siegmund (1971). It occurs in some classical search problems such as the "secretary problem" or the "job search" problem (with recall), see. e.g. Ferguson (1989) or Lippman and McCall (1976).

It is well known that the solution of this problem is given by the following value iteration.

$$V_{n+1}(s) = \max\left\{s, -c + \beta \int P(dx)\ V_n(\max\{s, x\})\right\}.$$

Here $V_n(s)$ is the optimal expected reward, if there are still n possible observations and s is the best offer so far. The data of the underlying MDP have to be defined as follows: $A := \{0, 1\}$, where action 1 means "stop" and action 0 means "continue". The reward function is given by $\tilde{r}(s, 0, s') = -c$ and $\tilde{r}(s, 1, s') = s$. We have to define the transition probabilities as $P(s, 1, \cdot) := \varepsilon_0$ (with $s = 0$ as absorbing state) and

$$P(s, 0, \cdot) := \int P(dx)\ \varepsilon_{\max\{s,x\}}(\cdot),$$

where ε_x denotes the one point measure in x.

It is easy to show that V_n is increasing, convex and Lipschitz continuous with $\|V_n\|_L \leq 1$. Hence we get the following results, if we apply Theorem 4.1 and 4.2 with $\mathfrak{V}$ the set of all increasing convex functions with $\|v\|_L \leq 1$.

Theorem 5.4 *Let P and Q be probability measures with finite mean, such that $P \leq_{ic} Q$. Then $V_n^P(s) \leq V_n^Q(s)$ for all $s \in S$ and $n \in \mathbb{N}_0$.*

Proof. This follows immediately from Theorem 4.1. □

Theorem 5.5 *Let P, Q be two probability measures with finite mean. Then*

$$|V_n^P(s) - V_n^Q(s)| \;\leq\; \beta\sigma_n(\beta) \cdot d_{sl}(P, Q)$$

for all $s \in S$ and $n \in \mathbb{N}_0$.

Proof. We will apply Theorem 4.2 with the set $\mathfrak{V}$ as defined above. The metric generated by $\mathfrak{V}$ is the stop-loss metric d_{sl}, see Example 3.2. For the Minkowski functional we get $\mu_{\mathfrak{V}}(V_n^P) \leq 1$. The reward function $\tilde{r}$ is independent of s'. This yields $\eta(s) \equiv 0$. From the definition of the transition probabilities we obtain $d_{\mathfrak{V}}(P(s,1,\cdot), Q(s,1,\cdot)) = 0$ and hence

$$\begin{aligned}
\delta(s) &= d_{\mathfrak{V}}(P(s,0,\cdot), Q(s,0,\cdot)) \\
&= \sup_{f \in \mathfrak{V}} \left| \int P(dx)\; f(\max\{s,x\}) - \int Q(dx)\; f(\max\{s,x\}) \right| \\
&= d_{\mathfrak{V}}(P, Q) \;=\; d_{sl}(P, Q).
\end{aligned}$$

The third equality holds, because $f \in \mathfrak{V}$ iff $x \mapsto f(\max\{s,x\}) \in \mathfrak{V}$ for all $s \in \mathbb{R}$.

Now we can prove the assertion by induction on n. The case $n = 0$ is trivial. Hence assume $g_n(s) \leq \beta\sigma_n(\beta) \cdot d_{sl}(P, Q)$. Then (4.3) implies

$$\begin{aligned}
g_{n+1}(s) &:= \eta(s) + \beta \cdot \mu_{\mathfrak{V}}(V_n^P) \cdot \delta(s) + \beta \cdot H_Q g_n(s) \\
&\leq \beta \cdot [d_{sl}(P,Q) + \beta \cdot \sigma_n(\beta) \cdot d_{sl}(P,Q)] \\
&= \beta\sigma_{n+1}(\beta) \cdot d_{sl}(P,Q).
\end{aligned}$$

□

Remark: Let us assume that the distribution of the observations involves a parameter θ, which is unknown to the decision maker, and the decision maker follows a Bayesian approach. Then we get a generalization of the optimal stopping problem, which can be solved by Bayesian Dynamic Programming. This model has been considered in Müller (1995). By using similar methods as here, the dependency of the solution on parameters of the prior distribution has been investigated there.

Acknowledgements

Some parts of this paper are content of my dissertation. I wish to express my sincere gratitude to Prof. Hinderer for his expert guidance and encouragement in the course of the development of that work.

References

BERTSEKAS D.P., S.E. SHREVE (1978). *Stochastic Optimal Control: The Discrete Time Case.* Academic Press, New York.

CHOW, Y.S., H. ROBBINS AND D. SIEGMUND (1971). *Great Expectations: The Theory of Optimal Stopping.* Houghton Mifflin, Boston.

DUDLEY R. M. (1989). *Real Analysis and Probability.* Wadsworth & Brooks.

DYNKIN E. B., A.A. YUSHKEVICH (1979). *Controlled Markov Processes.* Springer, Berlin.

FERGUSON, T.S. (1989) Who solved the secretary problem? *Statistical Sciences* **4**, 282-296.

HERNANDEZ-LERMA, O. AND W.J. RUNGGALDIER (1993). Monotone Approximations for convex stochastic control problems. *J. Math. Syst., Estimation and Control* **3**.

HEYMAN, D.P. AND M.J. SOBEL (1984). *Stochastic Models in Operations Research, Volume II.* McGraw-Hill.

HINDERER K. (1970). *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter.* Lecture Notes in Oper. Res. and Math. Syst. **33**. Springer, Berlin.

HINDERER K. (1979). On approximate solutions of finite-stage dynamic programs. *Dynamic Programming and its Applications* (M. Puterman ed.) 289-317. Academic Press, New York.

HINDERER K. (1984). On the structure of solutions of stochastic dynamic programs. *Proc. 7th Conf. on Probability Theory,* Brasov 1984 (M. Iosifescu ed.) 173-182.

Langen H.J. (1981). Convergence of dynamic programming models. *Math. Oper. Res.* **6**, 493 - 512.

Lippman, A. and J.J. McCall (1976). Job Search in a Dynamic Economy. *J. Econ. Theory* **12** 365-390.

Marshall A. W. (1991). Multivariate Stochastic Orderings and generating cones of functions. In Mosler K. und Scarsini M. (eds.), *Stochastic Orders and Decision under Risk.* IMS Lecture Notes - Monograph Series, Volume **19**, 231 - 247 .

Morin, T.L. (1978). Computational Advances in Dynamic Programming. In *Dynamic Programming and its Applications* (M. Puterman ed.) 53-90. Academic Press, New York.

Müller, A. (1995). Optimal Selection from Distributions with Unknown Parameters: Robustness of Bayesian Models. Technical Report, WIOR-464, University of Karlsruhe. To appear in *ZOR - Math. Meth. Oper. Res.*

Müller, A. (1996a). Stochastic Orders generated by Integrals: A Unified Study. Technical Report, WIOR-462, University of Karlsruhe. To appear in *Adv. Appl. Prob.*

Müller, A. (1996b). Integral Probability Metrics and their Generating Classes of Functions. Technical Report, WIOR-463, University of Karlsruhe. To appear in *Adv. Appl. Prob.*

Porteus E.L. (1975). On the optimality of structured policies in countable stage decision processes. *Man. Sci.* **22**, 148 - 157.

Puterman M.L. (1994). *Markov Decision Processes.* Wiley, New York.

Rachev S.T. (1991). *Probability Metrics and the Stability of Stochastic Models.* Wiley, New York.

Rachev S.T., L. Rüschendorf (1990). Approximation of sums by compound poisson distributions with respect to stop-loss distances. *Adv. Appl. Prob.* **22**, 350 - 374.

Rieder U., R. Zagst (1994). Monotonicity and bounds for convex stochastic control models. *Zeitschrift f. Oper. Res.* **39**, 187 - 207.

Rudin, W. (1973). *Functional Analysis.* McGraw-Hill.

SCARF, H. (1960). The Optimality of (s, S) Policies in Dynamic Inventory Models. in *Mathematical Methods in the Social Sciences 1959.* Stanford University Press.

SCHÄL, M. (1975). Conditions for Optimality in Dynamic Programming and for the Limit of n-stage Optimal Policies to be Optimal. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 179-196.

SCHÄL, M. (1976). On the Optimality of (s, S) Policies in Dynamic Inventory Models with Finite Horizon. *SIAM J. Appl. Math.* **30**, 528-537.

SHAKED, M., J.G. SHANTHIKUMAR (1994). *Stochastic Orders and their Applications.* Academic Press, London.

STOYAN D. (1983). *Comparison Methods for Queues and Other Stochastic Models.* Wiley.

VEINOTT, A.F. (1966). On the Optimality of (s, S) Inventory Policies: New Conditions and a New Proof. *SIAM J.* **14**, 1067-1083.

WESSELS J. (1977). Markov programming by successive approximation with respect to weighted supremum norms. *J. Math. Anal. Appl.* **58**, 326 - 335.

WHITT W. (1978). Approximations of dynamic programs, I. *Math. Oper. Res.* **3**, 231 - 243.

WHITT W. (1979). Approximations of dynamic programs, II. *Math. Oper. Res.* **4**, 179 - 185.

ZOLOTAREV V.M. (1983). Probability metrics. *Theory Prob. Appl.* **28**, 278-302.

Alfred Müller, Institut für Wirtschaftstheorie und Operations Research, Universität Karlsruhe, Kaiserstr. 12, D-76128 Karlsruhe, Germany.