Karlsruher Institut für Technologie

Target Value Criterion in Markov Decision Processes

Zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften

(Dr. rer. pol.)

von der
Fakultät für Wirtschaftswissenschaften
des Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Wi.-Ing. Lars Norman Moritz

Tag der mündlichen Prüfung:	4. Dezember 2014
Referent:	Prof. Dr. Karl-Heinz Waldmann
Korreferent:	Prof. Dr. Oliver Stein

Karlsruhe 2014

Abstract

During the past decade, risk-sensitive considerations have become more and more popular in the field of Markov Decision Processes. Most of the research focused on using special utility functions or mean-variance trade-offs to express risk-sensitivity.

In this thesis, we apply the Target Value Criterion to an MDP with a random planning horizon. In many decision problems appearing in innovative areas, the planning horizon of a project usually cannot be specified in advance. An approximation by an infinite planning horizon often turns out to be an oversimplification. A deterministic planning horizon considered so far in the literature is too restrictive for many applications. An estimation of the mean running time of the decision process, however, is usually possible due to historical data or expert knowledge. Also variability around the mean value can often be supposed to increase with the mean value. To take the random planning horizon into account, we use the geometric distribution.

Applying the Target Value Criterion means to minimize the probability that the total reward is below a predetermined target, referred to as the target value. We derive an optimality equation and prove the existence of an optimal stationary policy in a generalized state space, where the target space incorporates the realized one-stage rewards. The structure of the value function, that means the monotonicity and the asymptotic behavior is exploited to approximate the target space by a finite subset. Based on these structural results, upper and lower bounds are derived for the value function as well as nearly optimal policies. We show that the value function is continuous from the left side and we introduce an error integral to study the distance between the value function of the original problem and the one of the discretized problem. The discretization allows a decomposition of the problem, which is utilized to recursively determine its solution.

As an extension we combine the Total Reward Criterion and the Target Value Criterion in a penalty approach. The structure of the value function and the optimality of a stationary policy is proven. Moreover, we study the dependence of the optimal stationary policy on the penalty factor.

The thesis closes with a case study regarding an exemplary application.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

In the classical theory of Markov Decision Processes (MDPs) one of the most commonly used performance criteria is the Total Reward Criterion. Therein, a risk neutral decision maker is assumed, that concentrates on the maximization of expected revenues. However, in many applications, practitioners are concerned with the deviation of expected performance criteria and consider them too risky to adapt. Among these applications are portfolio management, revenue management, insurance and the management of energy systems. In these settings, the Total Reward Criterion is not appropriate to measure performance because it does not account for the involved risk of possible deviations from the expected value. Therefore, risk-sensitive decision maker may be interested in additional distributional properties of the Total Reward.

In the literature, risk aversion has, so far, been addressed by special utility function, certainty equivalent approaches, minmax approaches and mean-variance tradeoffs. For an exemplary application of the exponential utility function to address risk-aversion, we refer to Barz and Waldmann (2007) who considered a risk-averse capacity control problem in revenue management and to Bouakiz and Sobel (1992) who analyzed an inventory model in a risk-averse setting. Certainty equivalent approaches and mean-variance tradeoffs are dealt with in Van Dijk, Sladkỳ, et al. (2006),Sladkỳ and Sitavar (2004). For an overview, we refer to White (1988b), Howard and Matheson (1972). A drawback of the utilization of general utility functions is the enlargement of the state space, that makes the problems hard to solve from an algorithmical point of view. Moreover, the existence of an optimality equation, as well as, the

existence of deterministic policies is not guaranteed. Especially for practitioners it is difficult to implement policies that are randomized.

Utilizing the Target Value Criterion minimizes the probability of failing a predetermined target. This enables the decision maker to choose a target according to the risk attitude. Moreover, the existence of an optimal policy as well as the existence of an optimal deterministic policy can be proven.

1.1 Outline

Chapter 2 provides an introduction to Markov Decision Process (MDPs). We formally define sequential decision making under uncertainty in the MDP framework. As an optimization criterion, the Total Reward Criterion is introduced. Solving an MDP involves the determination of the optimal value function and the corresponding policy. A policy allows the control of the system at each point in time and is a sequence of decision roles. A decision rule prescribes the action the decision maker has to choose in each possible state of the system. In addition, structured policies are considered. The are desirable since they feature a simple structured form that can be exploited to efficiently calculate optimal decision rules. Next, MDPs with random planning horizon are discussed and the relationship to standard infinite MDPs is discussed. Section 2.4 introduces three methods for solving the optimality equation: Value Iteration including an efficient extrapolation method, Policy Iteration and Linear Programming. Chapter 2 closes with a motivating example that contains a simulation study illustrating the spreading of the realized total reward around the expected value.

Chapter 3 deals with the Target Value Criterion in MDPs with geometrically distributed planning horizon. First, an overview of the related literature is provided. Afterwards, the decision model is discussed. In order to apply the Target Value Criterion, the state space has to be extended by a second dimension - which we refer to as the target space - that incorporates the realized one-stage rewards and expresses the remaining contribution to the target that has to be achieved in the remaining

time of the planning horizon. Moreover, the definition of decision rules and policies have to be adopted to the new context. Finally, the existence of an optimality equation as well as the corresponding optimal stationary policy is proven. Section 3.3 introduces an equivalent model without discounting. For numerical calculations o the minimal probability of failing a predetermined target value, a discretization scheme for the target space is given. Since the target space is not a compact set, we develop a non-uniform discretization scheme that, based on an equidistant grid on the probability of failing a predetermined target and the exploitation of the structure of the geometrically distributed planning horizon, provides a possibility to recursively determine discretization points for the target space. Section 3.4.2 adopts the classical solution methods Value Iteration, Policy Iteration and Linear Programming to the new context. Upper and lower bounds of the value function are calculated. In order to evaluate the quality of the discretization scheme, an error integral based on the area between the lower and upper bound is proposed. Finally, a decomposition scheme of the target space into closed subsets is treated. Chapter 2 closes with numerical examples, concerning the proposed discretization scheme, the decomposition method and the evolution of the error integral dependent on the discretization step width.

Chapter 4 provides a combination of the Total Reward Criterion and the Target Value Criterion as an extension. Section 4.1 deals with a Chance Constrained approach, where the property that the probability of not achieving a predetermined Target Value is below a given threshold is treated as an additional constraint. In addition, a weighted sum approach, where the probability of not achieving a predetermined target value is treated as a penalty is proposed. Moreover, the existence of an optimality equation and the existence of a corresponding optimal stationary policy is proven. Section 4.2.1 adopts the discretization scheme to the new context. After that, an extrapolation method for the value Iteration Algorithm that speeds up the convergence is treated. Section 4.2.4 shows the limiting behavior of the objective function by varying the penalty parameter towards the Total Reward Criterion on the one hand, and towards the Target Value Criterion on the other hand. Section 4.3 provides a parametric programming approach for the penalty parameter. This chapter closes with numerical examples.

Chapter 5 provides a case study containing an application in inventory management.

Chapter 2

Total Reward Criterion in Markov Decision Processes

Markov Decision Processes (MDPs) provide a unified framework for the optimization of problems arising from sequential decision making under uncertainty. Applications cover a wide range of domains. To mention some of them, MDP models have been applied to revenue management, communication networks, inventory control, queuing systems, health care management, medical decision making and transportation science. For an overview, we refer to the survey papers of White (1985), White (1988a), White (1993b) and the textbook on methods and applications of Feinberg (2002). The recent survey of Altman (2001) on applications in communication networks contains nearly 200 references and emphasizes the importance of MDPs in that domain. For application examples in medical decision making we refer to Schaefer et al. (2005). A recommendation for the application of MDPs in medical decision making is given by the tutorial in Alagoz et al. (2010). For a variety of optimality criteria, these problems can be solved by dynamic programming. The main strength of this approach is that fairly general stochastic and nonlinear dynamics can be considered.

In the MDP framework, systems are characterized by collections of state variables evolving over time. State variables summarize the history of the system, containing all information that is relevant for describing future events. In formulating a problem as an MDP, state variables must be designed to satisfy the Markov property. That

means, conditioned on the current state of the system being known, its future is independent from its past.

The evolution of state variables depends on actions taken by the decision maker and on a probability distribution governing possible state transitions. We consider systems running in discrete time and occupy the discounted sum of rewards as an optimality criterion, which we refer to as the total reward criterion. Rewards accrue at each time step and depend on the state of the system and action being taken at that time.

The current chapter provides an introduction to risk-neutral MDPs. We formally define sequential decision-making under uncertainty in the MDP framework. We start with standard infinite horizon MDPs to introduce the theory, methodology and solution methods. A comprehensive introduction to MDPs is provided, e.g. by White (1993a), Puterman (2005) and Waldmann and Stocker (2013). General foundations of stochastic dynamical programming can, e.g. be found in Hinderer (1970). In the sequel, we allow MDPs to posses a structured state space, containing an absorbing set. That leads us to MDPs with random planning horizon that offer a framework for the risk-sensitive MDPs considered later on. MDPs with random planning horizon can be equivalently transformed to infinite horizon MDPs extending the state space by an absorbing state that indicates whether the process is still running or is already terminated. As a consequence, solution methods based on standard infinite horizon MDPs can be applied. We limit our discussion to finite state and action spaces.

2.1 The standard infinite horizon model

An infinite horizon MDP describes a stochastic system at discrete points in time $t \in \mathbb{N}_0$. At each point in time $t \in \mathbb{N}_0$ the system state s_t from the state space S is observed and a decision maker chooses an action a_t among the admissible actions $D(s)$. This action results in an immediate one-stage reward $r(s_t, a_t)$ and in a transition to system state s_{t+1} at time $t+1$ with probability $p_t(s_t, a_t, s_{t+1})$. In the case of a stationary MDP, the one-stage rewards, the actions and the stochastic transition law do not

depend on the decision epoch. In summary we can state the following definition of an infinite horizon MDP.

Definition 1. *An infinite horizon MDP consists of a five tuple* (S, A, p, r, α) *with*

(i) *a finite state space* S,

(ii) *a finite action space* A, *where* $D(s) \subset S$ *is the non-empty set of admissible actions in state* $s \in S$, *and the constrained set* $D := \{(s, a) | s \in S, a \in D(s)\}$,

(iii) *a stochastic transition law* $p : D \times S \to [0, 1]$, *that represents the probability* $p(s, a, s')$ *for a transition from state* $s \in S$ *to state* $s' \in S$ *for a given action* $a \in D(s)$,

(iv) *an one-stage reward function* $r : D \to \mathbb{R}$, *that represents the reward* $r(s, a)$ *for choosing action* a *in state* s,

(v) *one-stage discount factor* $\alpha \in (0, 1)$.

Notice that given s and a, $(p(s, a, s'), s' \in S)$ is a counting density on S. The discount factor α is a scalar between zero and one and represents inter-temporal preferences, indicating how rewards incurred at different time steps are combined in a single optimality criterion. In finance problems, α has a concrete interpretation: the same nominal value is worth less in the future than in the present, since in the latter case it can be invested for a risk-free return. Values of α greater than one can be allowed if the state space contains an absorbing set. For a detailed discussion of the critical discount factor we refer to Hinderer and Waldmann (2003).

The problem of sequential decision making amounts to the selection of a policy that optimizes a given criterion.

Definition 2. *A decision rule is a function* $f : S \to A$, *that specifies the action* $a = f(s)$ *in state* $s \in S$. *The set of all decision rules is denoted by* $F := \{f : S \times A | f(s) \in D(s) \text{ for all } s \in S\}$.

Definition 3. *A Markov policy* $\pi = (f_0, f_1, \ldots)$ *is a sequence of decision rules* $f \in F$, *specifying the action* $a_n = f_n(s_n)$ *chosen in state* $s_n \in S$ *at time* n. *The set of all stationary policies is denoted by* F^∞.

Mainly we are interested in stationary policies $\pi = (f, f, \ldots)$ for some $f \in F$.Given a stationary policy $\pi = (f, f, \ldots) \in F^\infty$, the dynamics of the system follows a Markov chain with transition probabilities $p(s, f(s), s'), s, s' \in S$. Due to our assumptions it suffices to consider stationary policies. Extending considerations to policies depending on the history of the system or randomization does not improve performance (e.g. Puterman (2005)). We employ the total reward criterion. The state process is denoted by $(\zeta_t)_{t\in\mathbb{N}_0}$.

For $\pi \in F^\infty$ and $s \in S$ let

$$V_\pi(s) := E_\pi \left[\sum_{t=0}^{\infty} \alpha^t r(\zeta_t, f_t(\zeta_t)) | \zeta_0 = s \right]$$

be the discounted expected total reward starting in state s and following policy π. Finally, we use

$$V(s) := \sup_{\pi \in F^\infty} V_\pi(s)$$

to denote the maximal expected total reward starting in state $s \in S$.

A policy π^* is called *reward*-optimal (r-optimal), if $V_{\pi^*}(s) = V(s)$ holds for all $s \in S$. We also say that a decision rule f^* is r-optimal, if the corresponding stationary policy $\pi^* = (f^*, f^*, \ldots)$ is r-optimal.

2.1.1 Structured Policies

Many applications comprise structures that correspond to optimal decision rules that can be exploited to efficiently calculate optimal policies. Consequently, using this structure, not all possible values of $s \in S$ have to be determined, since they feature a simply structured form. Some policies can even be completely characterized using only a few parameters.

A well-known example of a structured policy is the (s, S) policy in dynamic inventory problems. Its decision rule can be summarized as follows. If the inventory is above the level s then do not order, and if the inventory level is below the level s, order a quantity so that the inventory becomes S. In that example of a structured policy, only two values are necessary for the description of the policy.

A decision problem that provides an optimal structured policy is generally desirable. According to Powell (2011) the importance of identifying structured optimal policies is one of the most dramatic success stories from the study of Markov Decision Processes. Moreover, a structured policy is easily understood and implemented by end users which again increases the acceptance of the strategy. Finding structured optimal policies that can be computed efficiently and which are intuitive and exercisable in practice is one of the central challenges of dynamic optimization. For more examples on structured policies, we refer to Waldmann and Stocker (2013).

2.2 Markov Decision Processes with an Absorbing Set

An MDP with an absorbing set is a natural extension of the standard MDP, which allows a discount factor $\alpha = 1$, or more precisely, a discount factor α smaller than a critical discount factor $\alpha^* \geq 1$, resulting from both, the original discount factor α and the asymptotic behavior of the system.

It is realized by a structured state space S, containing an absorbing set $J_0 \subset S$, i.e. $\sum_{s' \in J_0} p(s, a, s') = 1$ with $r(s, a) = 0$ for $s \in J_0, a \in D(s)$. That means if the process enters an absorbing set, the evolution of the process is stopped and the rewards are equal to zero. For a more formal introduction, see Hinderer and Waldmann (2005) or Waldmann (2006).

Note that J_0 may be empty and need not be unique. In this section, however, we only consider $J_0 \neq \emptyset$ and assume J_0 to be arbitrary but fixed. The set $J := S \setminus J_0$ of transient states is called the essential state space because the behavior of the process is only of interest up to the entrance time into J_0 and not within J_0.

The absorbing property of J_0 is used there to find conditions that ensure the convergence of $V_\pi(s)$ for $s \in S$. Let $\tau := \inf\{t \in \mathbb{N} | \zeta_t \in J_0\} \leq \infty$ denote the entrance time into the absorbing set J_0, i.e. the first time that the state process (ζ_t) is in J_0, having started in some state $s \in S$. Note that using τ, the expected total reward is

$$V_\pi(s) = E_\pi \left[\sum_{n=0}^{\tau-1} \alpha^n r(\zeta_t, f(\zeta_t)) | \zeta_0 = s \right], \; s \in J.$$

Given a policy $\pi \in F^\infty$ and an initial state $s \in J$, the distribution of τ can be obtained by evaluating $P_\pi(\tau > t|\zeta_0 = s), t \in \mathbb{N}$ recursively.

In order to find an upper bound for $P_\pi(\tau > t|\zeta_0 = s)$, define an operator H on $\mathfrak{V}$, the set of all bounded functions on J with respect to the supremum norm, by

$$Hv(s) := \max_{a \in D(s)} \sum_{s' \in J} p(s, a, s')v(s'),$$

for $s \in J$ and $v \in \mathfrak{V}$. Let $H^{t+1}v = H(H^t v)$ and $H^0 v = v, v \in \mathfrak{V}, t \in \mathbb{N}_0$. Then,

$$H^t 1(s) = \sup_{\pi \in F^\infty} P_\pi(\tau > t|\zeta_0 = s),$$

with 1 denoting a vector with entries 1. Obviously, $\|H^t 1\|$ is an upper bound for the probability that the process has not entered the absorbing set J_0 at time t.

Hinderer and Waldmann (2005) show that the following equivalent assumptions ensure the existence of the maximal total expected reward.

(AS) $\alpha^t \|H^t 1\| < 1$ for some $t \in \mathbb{N}$,

(AS') $\alpha^t \|H^t 1\| \to 0$ as $t \to \infty$.

They prove the following theorem.

Theorem 1. *Given (AS) or (AS'),*

(i) The expected total reward

$$V(s) = \sup_{\pi \in F^\infty} E_\pi \left[\sum_{n=0}^{\infty} \alpha^n r(\zeta_n, f(\zeta_n)) | \zeta_0 = s \right],$$

is the unique bounded solution of the optimality equation $V = UV$,

$$V(s) = \max_{a \in D(s)} \left\{ r(s, a) + \sum_{s' \in J} \alpha p(s, a, s') V(s') \right\}, s \in J.$$

(ii) A decision rule f is r-optimal if and only if f is a maximizer of $U_f V$ (i.e. $UV(s) = U_f V(s)$ for all $s \in J$). Thus, there exists an r-optimal decision rule.

(iii) Value iteration works, i.e. for all $v_0 \in \mathfrak{V}$ *it holds that* $v_n := Uv_{n-1}, n \in \mathbb{N}$, *converges in norm to* V *(i.e.* $\|V - v_n\| \to 0$ *as* $n \to \infty$*).*

Given the assumption that the upper bound for the probability that the process has not entered the absorbing set J_0 at time n, $\alpha^n \|H^n 1\|$, converges to zero as n tends to infinity, value iteration can be used for finding the optimal policy and the associated maximal expected total reward. For $\alpha = 1$ this is equivalent to assuming that there is some $n' \in \mathbb{N}$ for which it is ensured that this upper bound of $P_\pi(\tau > n' | \zeta_0 = s)$ is smaller than 1.

2.3 Markov Decision Processes with random planning horizon

In the analysis of decisions concerning innovative products, the planning horizon of a project often cannot be specified exactly in advance. An approximation by an infinite planning horizon is often an oversimplification. Consequently, a deterministic planning horizon considered so far in the literature is too restrictive for many applications. An estimation of the mean running time, however, is often possible due to former experiences or expert knowledge. Also variability around the mean value can often be supposed to increase with the mean value. The geometric distribution, which will be used in the sequel, meets these requirements and distinguishes from other distributions by its mathematical simplicity of use due to the well known property of being memory-less.

It is well known (see e.g. Hinderer and Waldmann (2005), Ross (1995)) that the standard infinite-horizon MDP with discount factor $\alpha \in (0, 1)$ is equivalent to a finite horizon MDP with geometrically distributed planning horizon. The reformulation is essentially based on an extension of the state space by a stopping state, which is reached with transition probability $1 - \alpha$ from each of the (transient) states $s \in S$. Moreover, it is a special case of an MDP with an absorbing set.

To be more precise, we state the following definition.

Definition 4. *An MDP with random planning horizon, following a geometric distribution consists of a seven tuple* $(S, A, p, r, h, \alpha, 1-\beta)$ *with*

(i) *a finite state space* S*,*

(ii) *a finite action space* A*, where* $D(s) \subset A$ *is the non-empty set of admissible actions in state* $s \in S$*, and the constrained set* $D := \{(s,a)|s \in S, a \in D(s)\}$*,*

(iii) *a stochastic transition law* $p : D \times S \to [0,1]$*, that represents the probability* $p(s,a,s')$ *for a transition from state* $s \in S$ *to state* $s' \in S$ *for a given action* $a \in D(s)$*,*

(iv) *an one-stage reward function* $r : D \to \mathbb{R}$*, that represents the reward* $r(s,a)$*, for choosing action* a *in state* s*,*

(v) *a terminal reward function* $h : S \to \mathbb{R}$*, that represents the terminal reward* $h(s)$ *in state* s*,*

(vi) *one-stage discount factor* $\alpha \in (0,1)$*,*

(vii) *the planning horizon* τ*, following a geometric distribution on* $\mathbb{N}$ *with parameter* $1-\beta$*, i.e.* $P(\tau = n) = (1-\beta)\beta^{n-1}$ *for* $n \in \mathbb{N}$ *, (independent of the state process).*

For $\pi \in F^\infty$ and $s \in S$ let

$$V_\pi^\tau(s) := E_\pi \left[\sum_{t=0}^{\tau-1} \alpha^t r(\zeta_t, f_t(\zeta_t)) + \alpha^\tau h(\zeta_\tau) | \zeta_0 = s \right]$$

be the expected total reward starting in state s and following policy π. We use

$$V^\tau(s) := \sup_{\pi \in F^\infty} V_\pi^\tau(s)$$

to denote the maximal total reward starting in state $s \in S$. Finally, a policy $\pi^* \in F^\infty$, fulfilling $V_{\pi^*}^\tau(s) = V^\tau(s)$ for all $s \in S$, is called r-optimal. Extending the state space S by a stopping state s_{abs}, that indicates whether or not the process is still running, the problem can be reformulated as an MDP with an absorbing set in the sense of Hinderer and Waldmann (2005), with state space $S' = S \cup \{s_{abs}\}$, action space

$A' = A$, sets $D'(s) = D(s), s \in S, D'(s_{abs}) = A$ of admissible actions, transition probabilities $p'(s,a,s') = \beta p(s,a,s'), s' \in S$, $p(s,a,s_{abs}) = 1-\beta$ for $(s,a) \in D$ and $p(s_{abs},a,s_{abs}) = 1$ for $a \in A$, one-stage rewards $r'(s,a) = \overline{r}(s,a)$ on D, where

$$\overline{r}(s,a) := r(s,a) + \alpha(1-\beta)\sum_{s' \in S} p(s,a,s')h(s'),\ (s,a) \in D.$$

and $r' = 0$ otherwise, and, finally, discount factor $\alpha' = \alpha$. Then it follows from Theorem 3.1 in Hinderer and Waldmann (2005) that V^τ is the unique solution of the optimality equation

$$V^\tau(s) = \max_{a \in D(s)} \left\{ \overline{r}(s,a) + \alpha\beta \sum_{s' \in S} p(s,a,s')V^\tau(s') \right\}, s \in S \tag{2.1}$$

and that there exists an r-optimal decision rule $f^* \in F$ formed by actions $f^*(s) \in D(s)$ maximizing the right hand side of (2.1). Thus, a geometrically distributed planning horizon τ can be interpreted as a standard infinite-horizon MDP with one-stage rewards $\overline{r}(s,a)$ and discount factor $\alpha\beta$. Moreover, the efficient methods for solving the standard infinite-horizon model can be applied successfully.

2.4 Solution methods for the random horizon model

There are several iterative approaches for solving infinite horizon or random horizon problems. The first, Value Iteration, is the most widely used method. It involves iteratively estimating the value function. At each iteration the estimate of the value function determines which decisions will be made and, thus, defines a policy.

The second approach is the Policy Iteration . At each iteration a policy is defined and the according value function is calculated. Moreover, based on the test quantity the policy that is used in the next iteration is determined.

Finally, the third major iterative approach exploits the observation that the value function can be viewed as the solution to a specially structured linear programming problem.

The advantage of Value Iteration is the easy possibility of incorporating the structure of an optimal decision rule. This leads to a tremendous reduction of the calculation effort. As an example we refer to Grävenstein (2008) who exploited the structure of an optimal decision rule in a reservoir control problem.

Policy Iteration offers the possibility of applying a sensitivity analysis or parametric programming, as we will see later on. The linear programming method for discounted MDPs was proposed by Manne (1960). It is widely known that in the presence of a discount factor Howard's Policy Iteration routine corresponds precisely to block pivoting in a linear program. That was exhaustively studied in by d'Epenoux (1960).

2.4.1 Value iteration

The problem of finding an optimal policy can be converted into the problem of computing the maximum total reward $V(s)$, which we refer to as the value function. The set $\mathfrak{V}$ denotes the set of all bounded functions on S. Starting with an arbitrary starting point $v_0 \in \mathfrak{V}$, a sequence of approximations (v_n) is constructed that converges uniformly towards the value function. The algorithm can be found in Algorithm 1. For a proof and convergence properties we refer to Puterman (2005).

Algorithm 1 Value Iteration, risk neutral

Input: $n = 0, v_0 \in \mathfrak{V}, \varepsilon > 0$
repeat
  $n = n + 1$
  for all $s \in S$ **do**
    $$v_n(s) = \max_{a \in D(s)} \left\{ \overline{r}(s,a) + \alpha\beta \sum_{s' \in S} p(s,a,s') v_{n-1}(s') \right\}$$
    $$f_n(s) = \arg\max \{v_n(s)\}$$
  end for
until $\|v_n - v_{n-1}\| < \epsilon$
$V^\tau = v_n$
$f^* = f_n$
Output: approximation of the value function V^τ and a r-optimal decision rule f^*

Value Iteration with Extrapolation

The convergence of the basic Value Iteration algorithm is usually very slow. Combining the Value Iteration (v_n) with an extrapolation giving monotone upper and lower bounds

$$w_n^+(s) = v_n(s) + \frac{\alpha\beta}{1-\alpha\beta} \sup_{s' \in S} \{v_n(s) - v_{n-1}(s)\}$$
$$w_n^-(s) = v_n(s) + \frac{\alpha\beta}{1-\alpha\beta} \inf_{s' \in S} \{v_n(s) - v_{n-1}(s)\}, \; s \in S$$

for the value function at each step n of iteration, the quality of the approximation can usually be improved considerably. Moreover, a lower bound for the expected total reward associated with f_n can be given. The details of the approach credited to MacQueen for the standard infinite horizon model are summarized in the following theorem.

Theorem 2. *For all $n \in \mathbb{N}$ and all $s \in S$ it holds that*

(i) $w_n^-(s) \leq w_{n+1}^-(s) \leq V(s) \leq w_{n+1}(s) \leq w_n^+(s)$,

(ii) $\lim_{n\to\infty} w_n^-(s) = \lim_{n\to\infty} w_n^+(s) = V(s)$.

(iii) Let $f_n \in F$ with $v_n = Uv_{n-1}$. Then it holds that $V_{f_n} \geq w_n^-$.

Proof. The proof is contained in Waldmann and Stocker (2013). □

The algorithm is stated in Algorithm 2.

Algorithm 2 Value Iteration with Extrapolation, risk neutral

Input: $n = 0, v_0 \in \mathfrak{V}, \varepsilon > 0$
repeat
  $n = n + 1$
  for all $s \in S$ **do**
    $$v_n(s) = \max_{a \in D(s)} \left\{ \overline{r}(s,a) + \alpha\beta \sum_{s' \in S} p(s,a,s') v_{n-1}(s') \right\}$$
    $$f_n(s) = \arg\max \{v_n(s)\}$$
  end for
  for all $s \in S$ **do**
    $$w_n^-(s) = v_n(s) + \frac{\alpha\beta}{1-\alpha\beta} \inf_{s' \in S} \{v_n(s) - v_{n-1}(s)\}$$
    $$w_n^+(s) = v_n(s) + \frac{\alpha\beta}{1-\alpha\beta} \sup_{s' \in S} \{v_n(s) - v_{n-1}(s)\}$$
  end for
until $\|w_n^+ - w_n^-\| < 2\epsilon$
$V^\tau = (w_n^- + w_n^+)/2$
$f^* = f_n$
Output: approximation of the value function V^τ and a r-optimal decision rule f^*

2.4.2 Policy Iteration

While the Value Iteration is based on successive approximations of the value function, we have that the Policy Iteration constructs a sequence of policies converging towards a r-optimal policy π^*.

For the equivalence between Policy Iteration and the Newton-Kantorovich iteration procedure applied to the functional equation of dynamic programming, as well as proofs of the uniqueness of the solution and convergence rates in context with the standard MDP, we refer to Puterman and Brumelle (1979).

The Policy Iteration can be split into two steps: policy evaluation and policy improvement. Starting with an arbitrary decision rule $f_0 \in F$, the expected discounted total reward is calculated for the given decision rule. In a second step, based on the total reward associated with the decision rule, the optimality equation is used to decide whether or not the decision rule is r-optimal. If the maximizer can be chosen equal to the former decision rule, the algorithm terminates. In the other case, the maximizer is thought of as a new decision rule and is evaluated using the policy evaluation step. The properties of the policy iteration are summarized in the following theorem.
Part (i) ensures that f_n is optimal, if and only if $UV_{f_n} = U_{f_n}V_{f_n}$ holds;
part (ii) guarantees the monotonicity property, that is$V_{f_{n+1}} \geq V_{f_n}$.

Theorem 3.

(i) Let $f \in F$ with $U_f V_f(s) = UV_f(s)$ for all $s \in S$. Then it holds that

$$V_f(s) = V(s),\ s \in S.$$

(ii) Let $f, f' \in F$ with $U_{f'} V_f(s) = UV_f(s)$ for all $s \in S$.

$$V_{f'}(s) \geq V_f(s)\, s \in S.$$

Proof. For the proof we refer to Waldmann and Stocker (2013). □

The algorithm can be found in Algorithm 3.

Algorithm 3 Policy Iteration, risk neutral

Input: $n = 0, f_0 \in F$
Policy evaluation:
Calculate V_{f_n} as the unique solution of the linear system

$$V_{f_n}(s) = \overline{r}(s, f_n(s)) + \alpha\beta \sum_{s' \in S} p(s, f_n, s') V_{f_n}, s \in S$$

Policy improvement:
Calculate the test quantity

$$UV_{f_n}(s) = \max_{a \in D(s)} \left\{ \overline{r}(s,a) + \alpha\beta \sum_{s' \in S} p(s,a,s') V_{f_n}(s), s \in S \right\}$$

if f_n is maximizer of the test quantity UV_{f_n} for all $s \in S$ **then**
  f_n is optimal
else
  $n = n + 1$
  goto Policy evaluation
end if
$V^\tau = V_{f_n}$
$f^* = f_n$
Output: value function V^τ and a r-optimal decision rule f^*

2.4.3 Linear programming

The primal problem can be stated as follows.

$$\min \sum_{s \in S} w(s)$$

subject to

$$w(s) - \alpha\beta \sum_{s \in S} p(s,a,s') w(s') \geq \overline{r}(s,a),\ s \in S, a \in D(s).$$

The variables $w(s)$ of the primal problem correspond to the total reward $V(s)$.

Theorem 4. *It holds for the optimal solution $w^*(s), s \in S$ of the primal problem that*

$$w^*(s) = V(s), s \in S.$$

Proof. For the proof we refer to Waldmann and Stocker (2013). □

The dual problem can be stated as follows.

$$\max \sum_{s \in S} \sum_{a \in D(s)} \overline{r}(s,a) x(s,a)$$

subject to

$$\sum_{a \in D(s')} x(s',a) - \alpha\beta \sum_{s \in S} \sum_{a \in D(s)} p(s,a,s') x(s,a) = 1,\ s' \in S$$
$$x(s,a) \geq 0,\ s \in S,\ a \in D(s)$$

Theorem 5. *Let $x^*(s,a) \in D$ be an optimal solution of the dual problem. Then it holds that*

(i) For each $s \in S$ there exists exactly one $a \in D(s)$ with $x^*(s,a) > 0$. The remaining $x^*(s,a)$ equal zero.

(ii) The decision rule $f^* \in F$ resulting from the actions corresponding to $x^*(s,x)$ with $x^*(s,a) > 0$, is r-optimal.

Proof. For the proof we refer to Waldmann and Stocker (2013). □

The linear program has $|S \times A|$ inequality constraints. This formulation was viewed as primarily a theoretical result for many years, since it requires formulating a linear program where the number of constraints is equal to the number of states and actions. While even today this limits the size of problems it can solve, modern linear programming solvers can handle problems with tens of thousands of constraints without difficulty. This size is greatly expanded with the use of specialized algorithmic strategies, which are an active area of research.

2.5 Motivating example for alternative optimality criteria

The following example is a motivating example and illustrates that the policy that results from the application of the Total Reward criterion may lead to a high probability that the expected total reward will not be achieved.

Table 2.1 comprises the problem data. The first column contains the states, the second column the available actions in the corresponding state, the third column denotes the one-stage rewards and the following columns the transition probabilities to subsequent states. The discount factor α is set to one and the parameter $1-\beta$ of the geometric distribution is 0.2.

s	a	$r(s,a)$	$p(1,a,1)$	$p(1,a,2)$	$p(1,a,3)$
1	1	8	1/2	1/4	1/4
	2	11/4	1/16	3/4	3/16
	3	17/4	1/4	1/8	5/8
2	1	16	1/2	0	1/2
	2	15	1/16	7/8	1/16
3	1	7	1/4	1/4	1/2
	2	4	1/8	3/4	1/8
	3	9/2	3/4	1/16	3/16

Table 2.1: Motivating Example, problem data

The distribution function is determined by a simulation study. Notice that given a policy, the problem can be analyzed as a Markov Chain. For a detailed introduction to stochastic simulation we refer to Law and Kelton (2000) or Ross (2013). In total 1000 simulation runs are processed. The planning horizon is simulated according to a geometric distribution with parameter $1-\beta$ and the subsequent states according to the given transition probabilities. The realizations are determined by applying the Inverse Transform Method. Since the example serves as a motivating example we do not provide confidence intervals.

Figures 2.1 to 2.3 depict the distribution function for each possible initial state of the Markov Chain, following the risk-neutral policy. Notice that the distribution function

is continuous from the left side. The dashed vertical line depicts the expected total reward that results from applying a risk-neutral policy. $\Phi(s, x)$ depicts the simulated distribution function when the process starts in ininital state $s \in S$. The probability that the expected total reward is not achieved is about 70 per cent independent of the initial state.

A risk averse decision maker could prefer a lower expected total reward that will be achieved with a higher probability. This idea is connected with the satisficing approach that was first published in Simon (1955). According to behavioral science, risk averse decision makers tend to choose low targets that should be achieved with a high probability. Risk loving decision makers want to realize high total rewards that bear the risk of not being achieved with a high probability. For a more detailed discussion we refer to Simon (1957).

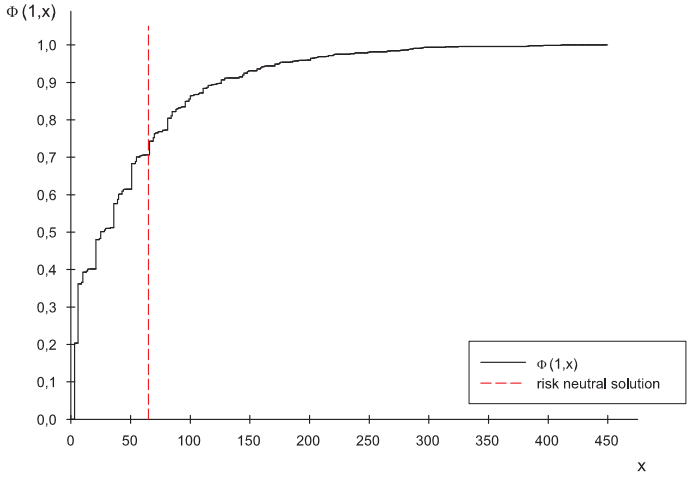

Figure 2.1: Motivating Example, simulated distribution function for state 1

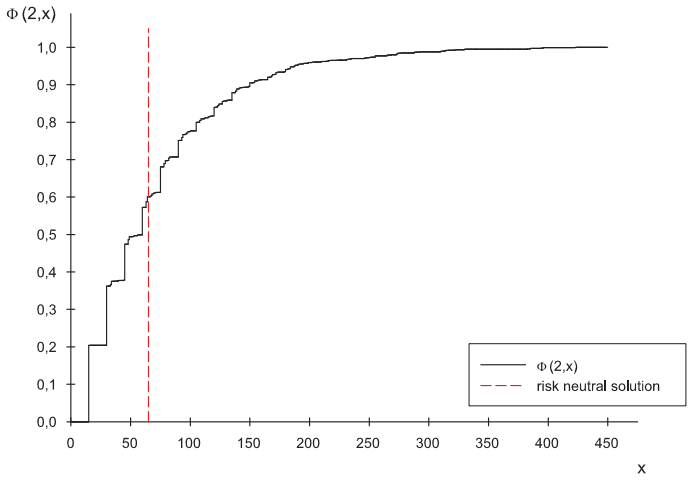

Figure 2.2: Motivating Example, simulated distribution function for state 2

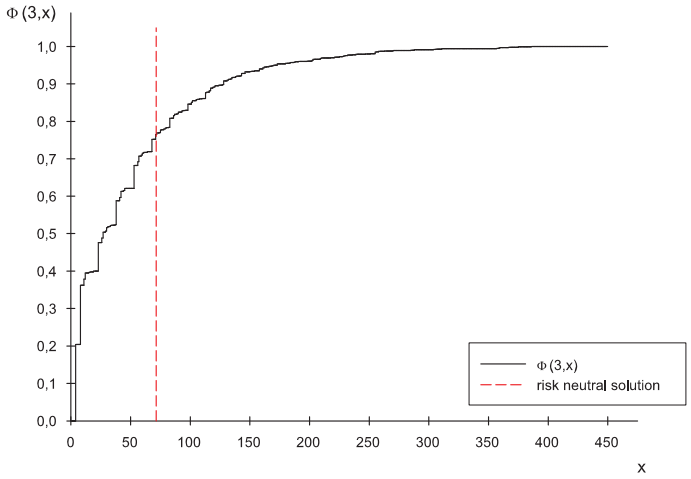

Figure 2.3: Motivating Example, simulated distribution function for state 3

Chapter 3

Target Value Criterion in Markov Decision Processes

In this chapter, we apply the Target Value criterion to an MDP with geometrically distributed planning horizon. The chapter is organized as follows. Section 3.1 provides an overview of the related literature. In Section 3.2 we prove the existence of an optimal stationary policy $\delta = (g, g, \ldots)$ determined by a decision rule g specifying action $g(s, x) \in D(s)$ to be taken in the generalized state $(s, x) \in S \times X$. In Section 3.3 the structure of the value function, e.g. the monotonicity as well as the asymptotic behavior is exploited to approximate the target space by a finite subset. Based on these structural results, upper and lower bounds are derived for the value function as well as nearly optimal policies. Since the value function is left continuous in x only, an error integral is introduced to study the area between the value function of the discretized and the original problem. The error integral can be used to evaluate the quality of the approximation. The discretization allows a decomposition of the problem, which is utilized to recursively determine its solution. Section 3.5 contains numerical results that demonstrate the efficiency of the method. Moreover, the proposed discretization scheme is compared with a uniform discretization scheme indicating that the number of essential discretization points is less in the proposed discretization scheme to achieve a comparable approximation accuracy. Finally we show that the error integral tends to zero as the discretization step width Δ tends to zero.

3.1 Related Literature

Within the setting of a finite and infinite-horizon MDP, several authors study the probability that the total discounted reward does not exceed a target value (cf. White (1993b),Bouakiz (1995),Wu and Lin (1999),Ohtsubo (2003)). This is done by extending the state space S by a set $X := \mathbb{R}$ of target values (target space), where $x \in X$ describes the total discounted reward to be received in the remaining time, and results in policies depending on both the actual state and the actual target value. Bouakiz (1995) derive an optimality equation and prove various properties of the value function. Wu and Lin (1999) additionally prove the existence of an optimal policy for the finite-horizon model and give sufficient conditions for the optimality of an infinite-horizon policy. Finally, in Ohtsubo and Toyonaga (2002) it is shown that an optimal policy is right continuous in the target value. We also refer to Koenig and Meissner (2009), who consider an application of the target value criterion in revenue management and provide a comparison of different risk-sensitive policies.

3.2 The decision model

Recall that the planning horizon τ of the MDP is geometrically distributed with parameter $1 - \beta$. To derive an optimal policy we extend the state space S and the set F^∞ of policies by incorporating the discounted sum of realized one-stage rewards. In particular, at time $t \in \mathbb{N}_0$ the generalized state space (s_t, x_t) and the action depend on the actual state $s_t \in S$ as well as on the updated target value $x_t \in X := \mathbb{R}$,

$$x_t := \frac{x_{t-1} - r(s_{t-1}, a_{t-1})}{\alpha},$$

to be realized in the remaining running time of the process. The adopted definition of a decision rule and a policy are stated below.

Definition 5. *A decision rule with respect to the target value criterion is a function $g : S \times X \to A$, that specifies the action $a = g(s, x)$ in state $(s, x) \in S \times X$. The set of all decision rules is $G := \{g : S \times X \times A | g(s, x) \in D(s)$ for all $(s, x) \in S \times X\}$.*

Definition 6. *A policy $\delta = (g_0, g_1, \ldots)$ with respect to the target value is a sequence of decision rules specifying the action $a_t = g_t(s_t, x_t)$ to be taken in the generalized state (s_t, x_t) at time $t \in \mathbb{N}_0$. The corresponding set of policies is denoted by G^∞.*

Now, for policy $\delta = (g_0, g_1, \ldots) \in G^\infty$ and initial state $(s, x) \in S \times X$ let

$$\Phi_\delta(s,x) := P_\delta \left(\sum_{t=0}^{\tau-1} \alpha^t r(\zeta_t, g_t(\zeta_t, \xi_t)) + \alpha^\tau h(\zeta_\tau) < x | \zeta_0 = s, \xi_0 = x \right) \tag{3.1}$$

be the probability that the total discounted reward x will not be achieved starting in state s and applying policy δ.

A policy δ^* is called optimal with respect to the target value criterion (t-optimal), if

$$\Phi_{\delta^*}(s,x) = \Phi(s,x) := \inf_{\delta \in G^\infty} \Phi_\delta(s,x)$$

holds for all $(s, x) \in S \times X$. We also call a decision rule g^* t-optimal, if the associated stationary policy $\delta^* = (g^*, g^*, \ldots)$ is t-optimal.

To determine $\Phi(s, x)$ and a t-optimal decision rule g^*, let $\mathfrak{V}$ be the set of all bounded Borel-measurable functions on $S \times X$. In order to simplify notation, set

$$c(s,x,a) := (1-\beta) 1_{(0,\infty)}(x - r(s,a) - \alpha \sum_{s' \in S} p(s,a,s') h(s')).$$

On $\mathfrak{V}$ define the operators $U_g, g \in G$, and U by

$$\begin{aligned} U_g v(s,x) &:= c(s,x,g(s,x)) + \beta \sum_{s' \in S} p(s, g(s,x), s') v(s', \frac{x - r(s, g(s,x))}{\alpha}) \\ U v(s,x) &:= \min_{a \in D(s)} \left\{ c(s,x,a) + \beta \sum_{s' \in S} p(s,a,s') v(s', \frac{x - r(s,a)}{\alpha}) \right\} \end{aligned}$$

and, for $n \in \mathbb{N}$, let $U^n v := U(U^{n-1} v)$, where $U^0 v = v$. Define $U_g^n, g \in G$ analogously. The operator U has the property of being monotone, i.e. $v \leq v'$ implies $Uv \leq Uv'$. The same holds for $U_g, g \in G$.

Moreover, the probability of failing a target value for a given policy g fulfills the

functional equation $\Phi_g = U_g\Phi_g$. The solution can be obtained by the method of successive approximations. These results are summarized in Proposition 1.

Proposition 1. *Let $g \in G$. It holds that*

(i) Φ_g is the unique solution to $\Phi_g = U_g\Phi_g$ in $\mathfrak{V}$,

(ii) $\Phi_g = \lim_{n\to\infty} U_g^n v, v \in \mathfrak{V}$.

Proof.

(i) For notational convenience, let $P_{g,s,x} := P_g(\cdot|\zeta_0 = s, \xi_0 = x)$ and $\gamma_k := (1-\beta)\beta^{k-1}$. Fix $(s,x) \in S \times X$. Set $a := g(s,x)$. Conditioning on $\tau = 1$ and $\tau > 1$, we then obtain

$$\begin{aligned}
\Phi_g(s,x) &= P_{g,s,x}\left[\sum_{t=0}^{\tau-1} r(\zeta_t, g(\zeta_t,\xi_t)) + \alpha^\tau h(\zeta_\tau)|\tau = 1\right](1-\beta) \\
&\quad + P_{g,s,x}\left[\sum_{t=0}^{\tau-1} \alpha^t r(\zeta_t, g(\zeta_t\xi_t)) + \alpha^\tau h(\zeta_\tau|\tau > 1\right]\beta \\
&= P_{g,s,x}\left[r(s,a) + \alpha\sum_{s'\in S} p(s,a,s')h(s') < x\right](1-\beta) \\
&\quad + \sum_{k=2}^{\infty} P_{g,s,x}\left[\sum_{t=0}^{k-1} \alpha^t r(\zeta_t, g(\zeta_t,\xi_t)) + \alpha^k h(\zeta_k) < x\right]\gamma_k \\
&= c(s,x,a) \\
&\quad + \sum_{k'=1}^{\infty} P_{g,s,x}\left[r(s,a) + \sum_{t'=0}^{k'-1} \alpha^{t'+1} r(\zeta_{t'+1}, g(\zeta_{t'+1},\xi_{t'+1})) \right. \\
&\quad \left. + \alpha^{k'+1} h(\zeta_{k'+1}) < x\right]\beta\gamma_{k'} \\
&= c(s,x,a) + \beta\sum_{s'\in S} p(s,a,s')\Phi_g(s', \frac{x - r(s,a)}{\alpha}) \\
&= U_g\Phi_g(s,x).
\end{aligned}$$

Thus, Φ_g is a fixed point of $\Phi_g = U_g\Phi_g$. Moreover, by applying the contraction mapping theorem, it follows that Φ_g is the unique fixed point of $\Phi_g = U_g\Phi_g$ in $\mathfrak{V}$.

(ii) Applying the contraction mapping theorem once again, if follows that $\Phi_g = \lim_{n\to\infty} U_g^n v, v \in \mathfrak{V}$.

□

Now we are in a position to minimize the probability of failing a target value. We state the following theorem, that regards the optimality equation, the corresponding decision rules, convergence properties of successive approximations and properties of the value function.

Theorem 6.

(i) Φ *is the unique solution in* $\mathfrak{V}$ *to the optimality equation* $\Phi = U\Phi$*, i.e. we have for all* $s \in S, x \in X$

$$\Phi(s,x) = \min_{a\in D(s)} \left\{ c(s,x,a) + \beta \sum_{s'\in S} p(s,a,s')\Phi(s', \frac{x - r(s,a)}{\alpha}) \right\}. \tag{3.2}$$

(ii) Each decision rule $g^* \in G$ *formed by actions* $g^*(s,x)$*, minimizing the right hand side of* (3.2) *(i.e., for which* $U\Phi = U_{g^*}\Phi$ *holds) is t-optimal.*

(iii) $\Phi = \lim_{n\to\infty} U^n v, v \in \mathfrak{V}$.

(iv) $\Phi(s,\cdot), s \in S$*, is increasing and left continuous in* x*.*

(v) For $s \in S$*, the smallest minimizer* $g^* \in G$ *of* (3.2) *is left continuous in* x*.*

Proof. Using the contraction mapping theorem, there exists a unique $v^* \in \mathfrak{V}$ such that $v^* = Uv^*$. Moreover, $v^* = \lim_{n\to\infty} U^n v_0$ for any $v_0 \in \mathfrak{V}$.

To verify $v^* = \Phi$, fix $\delta = (g_0, g_1, \ldots) \in G^\infty, z_n := (s_0, x_0, \ldots, s_n, x_n) \in (S \times X)^{n+1}$. Let $N \in \mathbb{N}, n \leq N$. Set $r_t = r$ and $h_{t+1} = h$ for $t < N$ and $r_t = 0$, $h_{t+1} = 0$ for $t \geq N$. Then, for

$$\Phi_{\delta,n}^N(z_n) := P_\delta \left(\sum_{t=n}^{\tau-1} r_t(\zeta_t, g_t(\zeta_t, \xi_t)) + \alpha^{\tau-n} h_t(\zeta_\tau) < x_n | \tau > n, \zeta_0 = s, \ldots, \xi_n = n \right),$$

it follows by induction on $n = N-1, N-2, \ldots, 0$ that

$$\Phi^N_{\delta,n}(z_n) = U_{g_n}\Phi^N_{\delta,n+1}(z_n) \geq U^n 0(s_n, x_n).$$

Thus $\Phi^N_{\delta,0} \geq U^N 0$. Finally, by letting $N \to \infty, \Phi_\delta = \lim_{N\to\infty} \Phi^N_{\delta,0} \geq \lim_{N\to\infty} U^N 0 = v^*$, which implies $\Phi = \inf_{\delta\in G^\infty} \Phi_\delta \geq v^*$.

On the other hand, since $D(s)$ is finite for all $s \in S$, there exists $g^* \in G$ such that $v^* = U_{g^*}v^*$. By Proposition 1, $\Phi_{g^*} = v^*$. Thus $\Phi \leq \inf_{g\in G} \Phi_g \leq \Phi^*_g$, which completes the proof of (i)-(iii).

Recall that $D(s), s \in S$, is finite. Starting value iteration with $v_0 = 1$, it easily follows on induction on n that $v_n = Uv_{n-1}$ is increasing and left continuous in x. Thus (iv) holds. Since $D(s), s \in S$, is finite, (v) is an immediate consequence of (iv) and (3.2). □

Note that $\Phi(s,\cdot), s \in S$, is only left continuous but not continuous, since, for $r = 1$ and $h = 0$, we have $\Phi(s,x) = 0$ for $x \leq 1$ and $\Phi(s,x) = 1-\beta^j$ for $j < x \leq j+1, j \in \mathbb{N}$.

3.3 An equivalent model without discounting

The optimality equation (3.2) depends on the parameter $1-\beta$ of the geometric distribution of τ as well as on the discounting factor α. In this section, we show that under certain conditions, the model can be equivalently transformed to a model without discounting factor.

In the sequel, we use the well known result that the standard infinite-horizon MDP with discount factor $\alpha < 1$ can be reduced to an MDP with discount factor $\alpha = 1$ by extending the state space S by an absorbing state $s_{abs} \notin S$. To exploit the approach for the target value criterion, we introduce an MDP$'$ with state space $S' = S \cup \{s_{abs}\}$, action space $A' = A$, sets $D'(s) = D(s), s \in S$ and $D(s_{abs}) = A$ of admissible actions, transition probabilities $p'(s,a,s') = \alpha p(s,a,s'), s \in S$, $p'(s,a,s_{abs}) = 1-\alpha$ for $(s,a) \in D$ and $p'(s_{abs},\cdot,s_{abs}) = 1$, one-stage rewards $r'(s,a) = r(s,a), (s,a) \in D$,

and $r(s_{abs}, a) = 0$ for $a \in A$, and $h(s_{abs}) = 0$, discount factor $\alpha' = 1$, and planning horizon $\tau = \tau'$.

Each decision rule $g \in G$ can be extended to a decision rule $g' \in G'$ in MDP$'$ such that $g(s, x) = g'(s, x), (s, x) \in S \times X$. Moreover, the restriction of a decision rule $g' \in G'$ to $S \times X$ coincides with a decision rule $g \in G$. Finally, it is easily verified that $\Phi_g(s, x) = \Phi'_{g'}, (s, x) \in S \times X$. In summary, we immediately obtain the following theorem.

Theorem 7. *We have $\Phi_g(s, x) = \Phi'_{g'}(s, x), (s, x) \in S \times X$. Further, the extension (reduction) of a t-optimal decision rule in MDP (MDP$'$) is t-optimal in MDP$'$ (MDP).*

Applied to MDP$'$, optimality equation (3.2) is modified to

$$\begin{aligned} \Phi'(s,x) &= \min_{a \in D'(s)} \Bigg\{ (1-\beta) 1_{(0,\infty)}(x - r'(s,a) - \sum_{s' \in S} p'(s,a,s') h(s')) \\ &\quad + \beta \sum_{s' \in S} p'(s,a,s') \Phi'(s', x - r'(s,a)) \Bigg\}. \end{aligned} \tag{3.3}$$

Inserting $\Phi'(s_{abs}, x) = 1_{(0,\infty)}(x), x \in X$, into (3.3), we obtain for $s \in S$

$$\begin{aligned} \Phi'(s,x) &= \min_{a \in D(s)} \Bigg\{ (1-\beta) 1_{(0,\infty)}(x - r(s,a) - \alpha \sum_{s' \in S} p(s,a,s') h(s')) \\ &\quad + \beta(1-\alpha) 1_{(0,\infty)}(x - r(s,a)) + \alpha\beta \sum_{s' \in S} p(s,a,s') \Phi'(s', x - r(s,a)) \Bigg\}. \end{aligned} \tag{3.4}$$

For the special case that there is no terminal reward, that is $h = 0$, we arrive at

$$\Phi'(s,x) = \min_{a \in D(s)} \Bigg\{ (1 - \alpha\beta) 1_{(0,\infty)}(x - r(s,a)) + \alpha\beta \sum_{s' \in S} p(s,a,s') \Phi(s', x - r(s,a)) \Bigg\},$$

which implies that the planning horizon is the minimum of two geometric distributions, the original one (with parameter $1 - \beta$) on the one hand and a second one (with parameter $1 - \alpha$) resulting from discount factor α on the other hand.

Hence, the results of this section show:

(a) MDP is equivalent to an MDP′ with discount factor $\alpha' = 1$ and extended state space $S \cup \{s_{abs}\}$.

(b) If $h = 0$, then MDP is equivalent to an MDP′ with discount factor $\alpha' = 1$ and a planning horizon τ', which follows a geometric distribution with parameter $\alpha\beta$.

To simplify our approach, we therefore suppose an MDP with discount factor $\alpha = 1$ in the sequel.

3.4 Discretization of the target space

For numerical calculations it is necessary to approximate the target space X by a finite set X_Δ. Instead of using a discretization scheme with equidistant points on a finite subinterval of X, we use a non-equidistant scheme which results in a natural way from a simplified reward structure. Therefore, we first derive upper and lower bounds for $\Phi(\cdot, x)$ by exploiting the properties of the geometric distribution. These bounds enable us to determine X_Δ as a finite set of representatives of X which are equidistant with respect to a monotone function.

Since r and h are bounded, there exist $r^\pm, h^\pm \in \mathbb{R}$ such that, for all $(s, a) \in D$, hold:

(a) $r^- \leq r(s, a) \leq r^+$;

(b) $h^- \leq \sum_{s' \in S} p(s, a, s')h(s') \leq h^+$.

Based on $r^\pm, h^\pm$ and $\rho_x^\pm : \mathbb{N}_0 \to \mathbb{R}, x \in \mathbb{R}$, defined by $\rho_x^\pm(t) := x-(t+1)r^\mp-h^\mp, t \in \mathbb{N}_0$, introduce $\Psi^\pm : X \to [0, 1]$,

$$\Psi^\pm(x) := (1 - \beta) \sum_{t=0}^{\infty} \beta^t 1_{(0,\infty)}(\rho_x^\pm(t)),\ x \in X.$$

It is easily verified that $\Psi^+(x)$ and $\Psi^-(x)$ are upper and lower bounds to $\Phi_\delta(\cdot, x), \delta \in G^\infty$. In fact, for $(s, x) \in S \times X$,

$$\begin{aligned}\Phi_\delta(s,x) &\leq P_\delta\left(\sum_{t=0}^{\tau-1}\alpha^t r^- + \alpha^\tau h^- < x | \zeta_0 = s, \xi_0 = x\right)\\ &= \sum_{\nu=1}^{\infty}(1-\beta)\beta^{\nu-1}1_{(0,\infty)}(\rho_x^+(\nu-1))\\ &= \Psi^+(x).\end{aligned}$$

Since the bounds are independent of δ, the same holds for Φ, that is

$$\Psi^-(x) \leq \Phi(s,x) \leq \Psi^+(x),\ (s,x) \in S \times X. \tag{3.5}$$

To simplify the calculation of $\Psi^\pm(x)$, first introduce $t_0^\pm(x) \in \mathbb{N}_0$ and $t_0^\pm(x) \leq t_1^\pm(x) \in \mathbb{N}_0 \cup \{\infty\}$, defined by

$$\begin{aligned}t_0^\pm(x) &:= \inf\left\{t \in \mathbb{N}_0 | \rho_x^\pm(t) > 0\right\}\\ t_1^\pm(x) &:= \sup\left\{t \in \mathbb{N} | \rho_x^\pm(t-1) > 0\right\},\end{aligned}$$

where $\inf\emptyset = 0, \sup\emptyset = 0$. Then, based on the sign of $r^\pm$, the following properties of $\rho_x^\pm(\cdot)$ are easily verified to hold.

Lemma 1. *For all $x \in X$, the map $t \to \rho_x^\pm(t)$ is affine and it holds that*

(i) If $r^\pm \geq 0$, then $\rho_x^\pm(\cdot)$ is decreasing. Further, $\rho_x^\pm(t) > 0$ on $\left\{t_0^\pm(x), \cdots, t_1^\pm(x) - 1\right\}$ and $\rho_x^\pm(t) \leq 0$, otherwise, where $t_0^\pm(x) = 0$.

(ii) If $r^\pm \leq 0$, then $\rho_x^\pm(\cdot)$ is increasing. Further, $\rho_x^\pm(t) > 0$ on $\left\{t_0^\pm(x), \cdots, t_1^\pm(x) - 1\right\}$ and $\rho_x^\pm(t) \leq 0$, otherwise, where $t_1^\pm(x) = \infty$ in case of $t_0^\pm(x) > 0$.

By Lemma 1, $\Psi^\pm(x)$ is positive on the interval $\left\{t_0^\pm(x), \ldots, t_1^\pm(x) - 1\right\}$ only, which allows us to rewrite $\Psi^\pm(x)$ as

$$\Psi^\pm(x) = (1-\beta)\sum_{t=t_0^\pm(x)}^{t_1^\pm(x)-1}\beta^t = \beta^{t_0^\pm(x)} - \beta^{t_1^\pm(x)}. \tag{3.6}$$

By combining (3.5) with (3.6) we immediately get

$$\beta^{t_0^-(x)} - \beta^{t_1^-(x)} \leq \Phi(s,x) \leq \beta^{t_0^+(x)} - \beta^{t_1^+(x)}. \tag{3.7}$$

One easily verifies that $t_0^{\pm}(x) \to 0$ and $t_1^{\pm}(x) \to \infty$ as $x \to \infty$, from which we conclude, together with (3.7), that $\lim_{x\to\infty} \Phi(s,x) = 1, s \in S$. Similarly we also obtain $\lim\limits_{x\to-\infty} \Phi(s,x) = 0, s \in S$.

It is convenient to work with the following more detailed presentation of the bounds (3.7).

Proposition 2. *It holds:*

(i) If $r^- \geq 0$, *then* $1-\beta^{t_1^-(x)} \leq \Phi(\cdot,x) \leq 1-\beta^{t_1^+(x)}$ *for* $x > r^- + h^-$ *and* $\Phi(\cdot,x) = 0$, *otherwise.*

(ii) If $r^+ \leq 0$, *then* $\beta^{t_0^-(x)} \leq \Phi(\cdot,x) \leq \beta^{t_0^+(x)}$ *for* $x \leq r^+ + h^+$ *and* $\Phi(\cdot,x) = 1$, *otherwise.*

(iii) If $r^- < 0 < r^+$, *then* $1 - \beta^{t_1^-(x)} \leq \Phi(\cdot,x) \leq \beta^{t_0^+(x)}$ *for* $x \in X$.

Additionally we have $\lim\limits_{x\to-\infty} \Phi(s,x) = 0$ *and* $\lim\limits_{x\to\infty} \Phi(s,x) = 1$ *for* $s \in S$.

Proof. (i)-(iii) follow from (3.7) by specifying the constants $t_0^{\pm}(x), t_1^{\pm}(x)$ and concluding $\Phi(\cdot,x) = 0$ (resp. $\Phi(\cdot,x)$=1) from $\Psi^+(x) = 0$ (resp. $\Psi^-(x) = 1$). □

According to Proposition 2, $\Phi(\cdot,x)$ is close to one for sufficiently large values of x and close to zero for sufficiently small values of x. Therefore we can restrict attention to a subinterval of X in order to calculate Φ approximately. Within this subinterval we select a finite number of representatives. These representatives, extended by $+\infty$ and $-\infty$, then form the target space in the discretized version of the MDP. To be more precise, fix $\Delta > 0, k = k(\Delta) \in \{2,3,\ldots\}$ such that $k\Delta = 1$. Depending on Δ and k, we now choose the representatives $x_j \in X_\Delta$, say, such that $\Phi(\cdot,x_j)$ is nearly equal to j/k for $j = 0,\ldots,k$. The technical details are given in the following Proposal.

3.4.1 Proposal to construct X_Δ

Determine constants $\Delta > 0, k = k(\Delta) \in \{2, 3, \ldots\}$ such that $k\Delta = 1$. Let $\varepsilon > 0$ arbitrary.

(1) $r^- \geq 0$.
By Proposition 2(i), $\Phi(\cdot, x) = 0$ for $x \leq x_{\min} := r^- + h^-$. Therefore choose $X_\Delta = \{-\infty, x_0, x_1, \ldots, x_{k-1}, \infty\}$, where, for $j = 0, \ldots, k-1$,

$$x_j = x_{\min} + \frac{\ln(1 - j/k)}{\ln(\beta)} \cdot r^+, \tag{3.8}$$

where x_j results from both $1 - \beta^{t_j} = j/k$, which is equivalent to $t_j = \ln(1 - j/k)/\ln(\beta)$, and $x_j - r^- - h^- - t_j r^+ = x_j - x_{\min} - t_j r^+ = 0$.

(2) $r^+ \leq 0$.
By Proposition 2(ii), $\Phi(\cdot, x) = 1$ for $x \geq x_{\max} := r^+ + h^+ + \varepsilon$. Therefore choose $X_\Delta = \{-\infty, x_{-k+1}, \ldots, x_{-1}, x_0, \infty\}$, where, for $j = 1, \ldots, k$,

$$x_{-k+j} = x_{\max} + \frac{\ln(j/k)}{\ln(\beta)} \cdot r^-, \tag{3.9}$$

where x_{-k+j} results from both $\beta^{t_j} = j/k$, which is equivalent to $t_j = \ln(j/k)/\ln(\beta)$, and $x_{-k+j} - x_{\max} - t_j r^- = 0$.

(3) $r^- < 0 < r^+$.
First apply (3.9) with $x_{\max} = 0$ to get $X_\Delta^- = \{-\infty, x_{-k+1}, \ldots, x_{-1}\}$, where, for $j = 1, \ldots, k-1$,

$$x_{-k+j} = \frac{\ln(j/k)}{\ln(\beta)} \cdot r^-. \tag{3.10}$$

Then apply (3.8) with $x_{\min} = 0$ to obtain $X_\Delta^+ = \{0, x_1, \ldots, x_{k-1}, \infty\}$, where, for $j = 1, \ldots, k-1$,

$$x_j = \frac{\ln(1 - j/k)}{\ln(\beta)} \cdot r^+. \tag{3.11}$$

Finally, set $X_\Delta = X_\Delta^- \cup X_\Delta^+$.

The idea of our Proposal can be generalized in the following way: For some $k \in \mathbb{N}$ define a set $X_{1/k}$ of representatives of X by

$$X_{1/k} := \left\{x_0, \ldots, x_k \in I | x_j = \eta^{-1}(j/k) \text{ for } j = 0, \ldots, k\right\} \cup \{\pm\infty\},$$

where I is a closed interval in $\mathbb{R} \cup \{\pm\infty\}$ such that $\Phi(\cdot, x) \in \{0, 1\}$ for $x \notin I$ and $\eta : I \to [0, 1]$ is a surjective map, which is strongly increasing.

It is easily verified that the representatives (3.8) result from $\eta_2(z) := 1 - \beta^{(z - x_{\min})/r^+}$ for $z \in I = [x_{\min}, \infty]$. Further (3.9) results from $\eta_1(z) := \beta^{(z - x_{\max})/r^-}$ for $z \in I = [-\infty, x_{\max}]$. Finally, (3.10) and (3.12) then result from $\eta(z) := 0.5\eta_1(z)$ for $z \leq 0$ and $\eta(z) := 0.5(1 + \eta_2))$ for $z \geq 0$.

After having made a proposal for selecting the set X_Δ of representatives of X, we next look at discretized versions of the MDP with finite target space. Let X_Δ be a finite subset of $\mathbb{R} \cup \{\pm\infty\}$ (not necessarily resulting from our proposal). For all $x \in \mathbb{R}$, introduce $\lceil x \rceil_\Delta$ (resp. $\lfloor x \rfloor_\Delta$) to be the smallest (resp. largest) $x_k \in X_\Delta$ such that $x \leq x_k$ (resp. $x \geq x_k$). For notational convenience we also use $\lceil x \rfloor_\Delta^+$ (resp. $\lceil x \rfloor_\Delta^-$) in case of $\lceil x \rceil_\Delta$ (resp. $\lfloor x \rfloor_\Delta$). Let $\mathfrak{V}_\Delta$ be the set of all bounded functions on $S \times X_\Delta$. On $\mathfrak{V}_\Delta$ introduce operators U_Δ^+ and U_Δ^-, defined by

$$U_\Delta^\pm v(s, x) := \min_{a \in D(s)} \left\{ c(s, x, a) + \beta \sum_{s' \in S} p(s, a, s') v(s', \lceil x - r(s, a) \rfloor_\Delta^\pm) \right\} \tag{3.12}$$

for all $(s, x) \in S \times X_\Delta$, where $v(\cdot, \infty) = 1, v(\cdot, -\infty) = 0$, and $\lceil \pm\infty \rfloor_\Delta^\pm = \pm\infty$.

The operators $U_\Delta^\pm$ fulfill the assumptions of the contraction mapping theorem, from which we conclude that there exist $\Phi_\Delta^+, \Phi_\Delta^- \in \mathfrak{V}_\Delta$ such that $\Phi_\Delta^\pm = U_\Delta^\pm \Phi_\Delta^\pm$ and $\Phi_\Delta^\pm = \lim_{n \to \infty} (U_\Delta^\pm) v, v \in \mathfrak{V}_\Delta$. Further, exploiting monotonicity of U_Δ^+ and U_Δ^-, it easily follows by induction on n that $\Phi_\Delta^-(s, x) \leq \Phi_\Delta^+(s, x), (s, x) \in S \times X_\Delta$. Finally, interpret $G_\Delta := \{g : S \times X_\Delta | g(s, x) \in D(s)\}$ as the corresponding set of decision rules. In particular, a minimizer of $\Phi_\Delta^+ = U_\Delta^+ \Phi_\Delta^+$ (resp. $\Phi_\Delta^- = U_\Delta^- \Phi_\Delta^-$) is called t_Δ^+-optimal (resp. t_Δ^-optimal).

We are now in a position to present the desired upper and lower bounds to Φ and nearly optimal decision rules on the basis of X_Δ.

Theorem 8. *Let $g \in G, g(s,x) := g^+_\Delta(s, \lceil x \rceil_\Delta)$, be an extension of an t^+_Δ-optimal decision rule $g^+_\Delta \in G_\Delta$. Then it holds that*

$$\Phi^-_\Delta(s,x) \leq \Phi(s,x) \leq \Phi_g(s,x) \leq \Phi^+_\Delta(s, \lceil x \rceil_\Delta)$$

for all $(s,x) \in S \times X$.

Proof. Recall that $\Phi = \lim_{n\to\infty} v_n, \Phi^\pm_\Delta = \lim_{n\to\infty} v^\pm_n$, where $v_n = Uv_{n-1}, v^\pm_n = U^\pm_\Delta v^\pm_{n-1}$. Starting value iteration with $v_0 = 0$ and $v^\pm_0 = 0$, respectively, and exploiting that c is increasing in x, it follows by induction on n that $v^-_n(\cdot, \lfloor x \rfloor) \leq v_n(\cdot, x) \leq v^+_n(\cdot, \lceil x \rceil), x \in X$. Finally, by letting $n \to \infty$, we get $\Phi^-_\Delta(s, \lceil x \rceil_\Delta) \leq \Phi(s,x) \leq \Phi^+_\Delta(s, \lceil x \rceil_\Delta), (s,x) \in S \times X$.

Let $v(\cdot, x) := \Phi^+_\Delta(\cdot, \lceil x \rceil_\Delta), x \in X$. Then, for all $(s,x) \in S \times X$, using that c and v are increasing in x, it follows that

$$\begin{aligned}
U_g v(s,x) &= c(s,x,g(s,x)) + \beta \sum_{s' \in S} p(s, g(s,x), s') v(s', x - r(s, g(s,x)) \\
&\leq c(s, \lceil x \rceil_\Delta, g^+_\Delta(s, \lceil x \rceil_\Delta)) \\
&\quad + \beta \sum_{s' \in S} p(s, g^+_\Delta(s, \lceil x \rceil_\Delta), s') v(s', \lceil \lceil x \rceil_\Delta - r(s, g^+_\Delta(s, \lceil x \rceil_\Delta)) \rceil_\Delta) \\
&= \Phi^+_\Delta(s, \lceil x \rceil_\Delta) \\
&= v(s,x),
\end{aligned}$$

which is well known to imply $\Phi_g(s,x) \leq \Phi^+_\Delta(s, \lceil x \rceil_\Delta), (s,x) \in S \times X$. □

3.4.2 Solution Methods

Value Iteration, Target Value

In order to solve the optimality equation, Value Iteration, Policy Iteration and Linear Programming can be used. This section starts with the presentation of the Value Iteration algorithm. With ϕ_n, we denote the iterates. The convergence results can be found in Proposition 1. The algorithm can be found in Algorithm 4.

Algorithm 4 Value Iteration, Target Value

Input: $n = 0, \phi_0 \in \mathfrak{V}, \varepsilon > 0$
repeat
$n = n + 1$
for all $(s, x) \in S \times X_\Delta$ **do**

$$\begin{aligned} \phi_n(s,x) &= \min_{a \in D(s)} \big\{(1-\alpha\beta)1_{(0,\infty)}(x - r(s,a)) \\ &\quad +\alpha\beta \sum_{s' \in S} p(s,a,s')\phi_{n-1}(s', x - r(s,a))\big\} \end{aligned}$$

$g_n(s,x) = \arg min \{v_n(s,x)\}$
end for
until $\|\phi_n - \phi_{n-1}\| < \epsilon$
$\Phi = \phi_n$
$g^* = g_n$
Output: approximation of the value function Φ and a t-optimal decision rule g^*

Policy Iteration, Target Value

The Policy Iteration algorithm is stated in Algorithm 5.

Algorithm 5 Policy Iteration, Target Value

Input: $n = 0, g_0 \in G$
Policy evaluation:
Calculate Φ_{f_n} as the unique solution of the linear system

$$\begin{aligned}\Phi_{g_n}(s) &= (1-\alpha\beta)1_{(0,\infty)}(x - r(s, g(s,x))) \\ &\quad +\alpha\beta \sum_{s' \in S} p(s, g_n(s,x), s')\Phi_{g_n}(s', x - r(s, x - g(s,x)), (s,x) \in S \times X_\Delta\end{aligned}$$

Policy improvement:
Calculate the test quantity

$$\begin{aligned}U\Phi_{f_n}(s) &= \min_{a \in D(s)} \big\{(1-\alpha\beta)1_{(0,\infty)}(x - r(s,a)) \\ &\quad +\alpha\beta \sum_{s' \in S} p(s,a,s')\Phi_{f_n}(s', x - r(s,a)), s \in S\big\}\end{aligned}$$

if g_n is maximizer of the test quantity UV_{g_n} for all $(s,x) \in S \times X_\Delta$ **then**
  g_n is optimal
else
  $n = n + 1$
  goto Policy evaluation
end if
$\Phi = \Phi_{g_n}$
$g^* = g_n$
Output: value function Φ and a t-optimal decision rule g^*

3.4.3 Error Integral

In order to quantify the discretization error, we determine $\sigma_s(\Delta)$, that is the area between $\Phi(s,\cdot)$ and its discretized versions $\Phi^{\pm}_{\Delta}(s,\cdot)$. Furthermore, we show that $\sigma_s(\Delta) \to 0$ as $\Delta \to 0$.

We first introduce for $x \in \mathbb{R}$

$$\underline{\Psi}(x) := \int_{-\infty}^{x} \Psi^{+}(u)du, \quad \overline{\Psi}(x) := \int_{x}^{\infty} (1 - \Psi^{-}(u))du$$

in order to state the following proposition.

Proposition 3.

(i) Let $x - r^+ - h^+ > 0$. If $r^+ > 0$, then

$$0 \leq \overline{\Psi}(x) = r^+ \beta^{\lceil (x-h^+)/r^+ \rceil - 1} \left(\lceil \frac{x-h^+}{r^+} \rceil - \frac{x-h^+}{r^+} + \frac{\beta}{1-\beta} \right).$$

Otherwise (i.e. $r^+ \leq 0$) we have $\overline{\Psi}(x) = 0$.

(ii) Let $x - r^- - h^- \leq 0$. If $r^- < 0$, then

$$0 \leq \underline{\Psi}(x) = -r^- \beta^{\lceil (x-h^-)/r^- \rceil - 1} \left(\lceil \frac{x-h^-}{r^-} \rceil - \frac{x-h^-}{r^-} + \frac{\beta}{1-\beta} \right).$$

Otherwise (i.e. $r^- \geq 0$) we have $\underline{\Psi}(x) = 0$.

Proof. Let $x - r^+ - h^+ > 0$. If $r^+ \leq 0$, then $x - r^+ - h^+ - tr^+ > 0$ for all $t \in \mathbb{N}_0$ and

$$\overline{\Psi}(x) = (1-\beta) \sum_{t=0}^{\infty} \beta^t \int_{x}^{\infty} (1 - 1_{(0,\infty)}(u - r^+ - h^+ - tr^+))du = 0.$$

Therefore let $r^+ > 0$. Then there exists a smallest $N = N(x) \in \mathbb{N}$ such that

$x - r^+ - h^+ - tr^+ \leq 0$ for $t \geq N$ and we get

$$\begin{aligned}
\overline{\Psi}(x) &= (1-\beta)\sum_{t=0}^{\infty}\beta^t \int_x^{\infty}(1 - 1_{(0,\infty)}(u - r^+ - h^+ - tr^+))du \\
&= (1-\beta)\sum_{t=N}^{\infty}\beta^t \int_x^{\infty}(1_{(-\infty,0]}(u - r^+ - h^+ - tr^+)du \\
&= (1-\beta)\beta^N \sum_{t=0}^{\infty}\int_x^{(t+1+N)r^+ + h^+} 1du \\
&= \beta^N\left(\frac{r^+}{1-\beta} + Nr^+ - (x - h^+)\right) \\
&= r^+\beta^{\lceil (x-h^+)/r^+ \rceil - 1}\left(\lceil \frac{x-h^+}{r^+} \rceil - \frac{x-h^+}{r^+} + \frac{\beta}{1-\beta}\right).
\end{aligned}$$

Since $\overline{\Psi}(x)$ is nonnegative, assertion (i) holds. (ii) can be shown analogously. □

Proposition 3(ii) corresponds to the left tail of the value function and shows that the area, i.e. the area between the x-axis and $\Phi(s,\cdot)$, restricted to the interval $(-\infty, x)$, converges to zero as $x \to -\infty$. Moreover, regarding Proposition 3(i), an analogous result also holds for the asymptotic behavior of $\Phi(s,\cdot)$ as $x \to +\infty$, i.e. the right tail of the value function.

For notational convenience, let $X_\Delta := \{-\infty, x_0, \ldots, x_k, \infty\}$ (with $x_j \in \mathbb{R}, x_j \leq x_{j+1}$). Using the asymptotic behavior of $\Phi(s,\cdot)$ resulting from Proposition 3 and utilizing the bounds $\Phi^-_\Delta(s, \lfloor x \rfloor_\Delta) \leq \Phi(s,x) \leq \Phi^+_\Delta(s, \lceil x \rceil_\Delta)$, $(s,x) \in S \times X$, stated in Theorem 8, we consider

$$\begin{aligned}
\sigma_s(\Delta) &:= \int_{x_0}^{x_k}(\Phi^+_\Delta(s, \lceil x \rceil_\Delta) - \Phi^-_\Delta(s, \lfloor x \rfloor_\Delta))dx + \underline{\Psi}(x_0) + \overline{\Psi}(x_k) \\
&= \sum_{j=0}^{k-1}(\Phi^+_\Delta(s, x_{j+1}) - \Phi^-_\Delta(s, x_j))(x_{j+1} - x_j) + \underline{\Psi}(x_0) + \overline{\Psi}(x_k),\ s \in S,
\end{aligned}$$

in the sequel to measure the area between $\Phi(s,\cdot)$ and its discretized versions $\Phi^\pm_\Delta(s,\cdot)$. If $\sigma_s(\Delta) \to 0$ as $\Delta \to 0$, then the jump discontinuities (and jump heights) of the discretized versions of $\Phi(s,\cdot)$ converge to the original ones.

3.4.4 Decomposition of the discretized MDP

In order to improve the efficiency of the optimization procedure, we present a decomposition scheme for the discretized MDP. Therefore, we introduce the concept of $\lceil\cdot\rceil$ and $\lfloor\cdot\rfloor$-closeness to partition the target space. Based on the partition, we identify subsets of the target space, where the value function is already known in advance. As a result, we only need to use successive approximations in the single subset, that is neither $\lceil\cdot\rceil_\Delta$-1-absorbing nor $\lfloor\cdot\rfloor_\Delta$-0-absorbing.

Definition 7. *A subset $J \subset S\times X_\Delta$ fulfilling $\sum_{s'\in S} p(s,a,s')1_J(s', \lceil x-r(s,a)\rceil_\Delta) = 1$ for $(s,x) \in J$ is said to be $\lceil\cdot\rceil_\Delta$-closed. $\lfloor\cdot\rfloor_\Delta$-closeness is defined analogously (with $\lceil\cdot\rceil_\Delta$ replaced by $\lfloor\cdot\rfloor_\Delta$). By construction, the sets $S\times\{-\infty\}$ and $S\times\{+\infty\}$ are both $\lceil\cdot\rceil_\Delta$-closed and $\lfloor\cdot\rfloor_\Delta$-closed.*

Proposition 4. *Let J be a $\lceil\cdot\rceil_\Delta$-closed (resp. $\lfloor\cdot\rfloor_\Delta$-closed) subset of $S\times X_\Delta$. If $c(s,x,a) = (1-\beta)\gamma$ for all $(s,x)\in J, a\in D(s)$, and some $\gamma\in\{0,1\}$, then $\Phi^+_\Delta(s,x) = \gamma$ (resp. $\Phi^-_\Delta(s,x)=\gamma$) for all $(s,x)\in J$.*

Proof. Let J be $\lceil\cdot\rceil_\Delta$-closed. Then value iteration (v_n) based on (3.10) can be restricted to J. Starting with $v_0 = 0$, then $\Phi^+_\Delta = \lim\limits_{n\to\infty} v_n = \lim\limits_{n\to\infty}(1-\beta^n)\gamma = \gamma$. The second case follows analogously. □

A $\lceil\cdot\rceil_\Delta$-closed subset J of $S\times X_\Delta$, for which Φ^+_Δ is known, is said to be $\lceil\cdot\rceil_\Delta$-absorbing. In particular, a $\lceil\cdot\rceil_\Delta$-closed subset J is called $\lceil\cdot\rceil_\Delta$-1-absorbing (resp. $\lceil\cdot\rceil_\Delta$-0-absorbing), if Φ^+_Δ takes on the constant value 1 (resp. 0) on J. By construction, the sets $S\times\{-\infty\}$ and $S\times\{+\infty\}$ are 0-absorbing and 1-absorbing, with respect to both $\lceil\cdot\rceil_\Delta$ and $\lfloor\cdot\rfloor_\Delta$. Moreover, in case of $r\le 0$ we may expect that $S\times\{x\}$ is $\lceil\cdot\rceil_\Delta$-1-absorbing for large enough $x\in X_\Delta$. Some details will be given in the following examples.

In the sequel, we will simply speak of a closed (absorbing,...) set, if the statement is true for both types of rounding.

The discretization of the target space leads to a partition $J_1\cup\ldots\cup J_m$ of the state space $S\times X_\Delta$ consisting of $m\in\mathbb{N}$ closed subsets $J_1,\ldots,J_m$. The partition allows us to determine $\Phi^\pm_\Delta$ on $J_1,\ldots,J_m$, separately by restricting the corresponding optimality

equations to the closed subset under consideration. Some of these subsets are also absorbing for which, using Proposition 4, $\Phi^{\pm}_{\Delta}$ is already known in advance.

Moreover, a closed subset J may have an absorbing subset $K \subset J$. In this case, we may restrict our calculation to $J \setminus K$, the set of essential states of J. To be more precise, first recall that $\Phi^{\pm}_{\Delta}$ is known on K. Then, for $(s,x) \in J \setminus K, a \in D(s)$ and $v : J \to \mathbb{R}$, we may introduce

$$\begin{aligned} c^{\pm}_{J\setminus K}(s,x,a) &:= c(s,x,a) + \beta \sum_{s' \in S} p(s,a,s')(1_K \cdot \Phi^{\pm}_{\Delta})(s', \lceil x - r(s,a) \rfloor^{\pm}_{\Delta}) \\ L^{\pm}_{J\setminus K} v(s,x,a) &:= \sum_{s' \in S} p(s,a,s')(1_{J\setminus K} \cdot v)(s' \lceil x - r(s,a) \lfloor^{\pm}_{\Delta}) \\ U^{\pm}_{J\setminus K} &:= \min_{a \in D(s)} \left\{ c^{\pm}_{J\setminus K}(s,x,a) + \beta L^{\pm}_{J\setminus K} v(s,x,a) \right\} \end{aligned} \tag{3.13}$$

in order to obtain $\Phi^{\pm}_{\Delta}$ as the unique fixed point of $v^* = U^{\pm}_{J\setminus K} v^*$ on $J \setminus K$. To hold the notation simple, we do not distinguish between $1_{J\setminus K} \cdot v$ and the restriction of v to $J \setminus K$; in other words we also interpret $U^{\pm}_{J\setminus K} v$ as a function on $J \setminus K$ with extension $1_{J\setminus K} U^{\pm}_{J\setminus K} v$ to J.

The following example illustrates the reduction in the computational effort resulting from the use of the additional structure. Suppose we are interested in calculating Φ^{+}_{Δ} on the basis of X_{Δ} introduced in the proposal. Let $\beta > 0.5$. We consider two cases:

(1) $0 \le r^- < r^+$.
Recall that $X_{\Delta} = \{-\infty, x_0, x_1, \ldots, x_{k-1}, \infty\}$ with representatives x_j defined by (3.8). It is easily verified that

$$x_{k-j} - x_{k-j-1} = (\ln \beta)^{-1} \ln(j/(j+1)) \cdot r^+,$$

is decreasing in j $(j = 1, \ldots, k-1)$. Since $\beta > 0.5$, there exists

$$j^* := \max \{ j \in \{1, \ldots, k-1\} \,|\, \ln(j/(j+1)) < \ln \beta \}. \tag{3.14}$$

For $j \le j^*$,

$$\lceil x_{k-j} - r(s,a) \rceil_{\Delta} - x_{k-j-1} \ge (\ln \beta)^{-1} \ln(j/(j+1)) \cdot r^+ - r^+ > 0,$$

which implies that $S \times \{x_{k-j^*}\}, S \times \{x_{k-j^*+1}\}, \ldots, S \times \{x_{k-1}\}, S \times \{\infty\}$ are $\lceil\cdot\rceil_\Delta$-closed subsets of $S \times X_\Delta$. If $h = 0$, the these sets are also $\lceil\cdot\rceil_\Delta$-1-absorbing. Set $x_k := \infty$. Further, let

$$j^{**} := \max\left\{j \in \{0, 1, \ldots, j^*\} \,|x_{k-j} - r^+ - h^+ > 0\right\}.$$

Then the set $S \times \{x_{k-j^{**}}\}, \ldots, S \times \{x_{k-1}\}, S \times \{\infty\}$ are $\lceil\cdot\rceil_\Delta$-1-absorbing. On the other hand, since $r \geq 0$, we have $\lceil x_{k-j^*-1} - r(s,a)\rceil_\Delta \leq x_{k-j^*-1}$ for $(s,a) \in D$. Hence, $S \times \{x_0, \ldots, x_{k-j^*-1}\}$ is $\lceil\cdot\rceil_\Delta$-closed. Finally, let $S \times \{-\infty, x_0\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing. These observations allow us to decompose the optimization problem: First solve $\Phi^+_\Delta = U^+\Phi^+_\Delta$ on $S \times \{x_0, \ldots, x_{k-j^*-1}\}$ using that $S \times \{x_0\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing. Then, for $j^{**} < j \leq j^*$, solve $\Phi^+_\Delta = U^+\Phi^+_\Delta$ on $S \times \{x_{k-j}\}$. On the remaining subset $S \times \{x_{k-j^{**}}, \ldots, \infty\}$, Φ^+_Δ is $\lceil\cdot\rceil_\Delta$-1-absorbing and thus already known.

(2) $r^- < 0 < r^+$.
Recall that $X_\Delta = \{-\infty, x_{-k+1}, \ldots, x_{-1}, 0, x_1, \ldots, x_{k-1}, \infty\}$ with representatives x_j defined by (3.9) and (3.10). Define j^* and j^{**} as in (1). Then $S \times \{x_{k-j^{**}}, \ldots, \infty\}$ is $\lceil\cdot\rceil_\Delta$-1-absorbing, and $S \times \{x_{k-j^*}, \ldots, x_{k-1}, \infty\}$ is $\lceil\cdot\rceil_\Delta$-closed.

Since we round up, an analogous line of argumentation is not possible for negative values of X_Δ. However, for small enough j, that is for $1 \leq j^{***} \leq k$ such that $r^- \ln(j/(j+1)) < r^+ \ln\beta$, the sets $S \times \{x_{-k+j^{***}}, \ldots, x_{k-1}, \infty\}$ is $\lceil\cdot\rceil_\Delta$-closed. Finally, set $S \times \{-\infty\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing.

These observations allow us to decompose the optimization problem: First solve (3.12) on the subset $S \times \{x_{k-j^*}, \ldots, \infty\}$ using that $S \times \{x_{k-j^{**}}, \ldots, \infty\}$ is $\lceil\cdot\rceil_\Delta$-1-absorbing. Then solve (3.12) on the subset $S \times \{x_{-k+j^{***}}, \ldots, \infty\}$ using that Φ^+_Δ is already known on $S \times \{x_{k-j^*}, \ldots, \infty\}$ using that Φ^+_Δ is already known on $S \times \{x_{j-j^*}, \ldots, \infty\}$. Finally, for $j < j^{***}$, based on the knowledge of Φ^+_Δ on $S \times \{x_{-k+j+1}, \ldots, \infty\}$ solve (3.12) on $S \times \{x_{-k+j}, \ldots, \infty\}$.

Algorithm 6 contains the Decomposition scheme for case 1.

Algorithm 6 Value Iteration with Decomposition case 1, Target Value

Input: $n = 0, \phi_0 \in \mathfrak{V}, \varepsilon > 0$
for all $(s, x) \in S \times \{-\infty, x_0\}$ **do**
  $\phi_n(s, x) = 0$
end for
repeat
  $n = n + 1$
  for all $(s, x) \in S \times \{x_0, \ldots, x_{k-j^*-1}\}$ **do**
    $\phi_n^+ = U^+ \phi_n^+$
    $g_n(s, x) = \arg\min \{\phi_n(s, x)\}$
  end for
until $\|\phi_n - \phi_{n-1}\| < \epsilon$
$n = 0$
repeat
  $n = n + 1$
  for all $(s, x) \in S \times \{x_{j^*}, \ldots, x_{k-j^{**}}\}$ **do**
    $\phi_n^+ = U^+ \phi_n^+$
    $g_n(s, x) = \arg\min \{\phi_n(s, x)\}$
  end for
until $\|\phi_n - \phi_{n-1}\| < \epsilon$
$\Phi = \phi_n$
$g^* = g_n$
Output: approximation of the value function Φ and a t-optimal decision rule g^*

Algorithm 7 contains the Decomposition scheme for case 2.

Algorithm 7 Value Iteration with Decomposition case 2, Target Value

Input: $n = 0, \phi_0 \in \mathfrak{V}, \varepsilon > 0$
for all $(s, x) \in S \times \{x_{k-j^{**}}, \ldots, \infty\}$ **do**
  $\phi_n(s, x) = 1$
end for
repeat
  $n = n + 1$
  for all $(s, x) \in S \times \{x_{k-j^{*}}, \ldots, \infty\}$ **do**
    $\phi_n^+ = U^+ \phi_n^+$
    $g_n(s, x) = \arg\min \{\phi_n(s, x)\}$
  end for
until $\|\phi_n - \phi_{n-1}\| < \epsilon$
$n = 0$
repeat
  $n = n + 1$
  for all $(s, x) \in S \times \{x_{-k+j^{***}}, \ldots, \infty\}$ **do**
    $\phi_n^+ = U^+ \phi_n^+$
    $g_n(s, x) = \arg\min \{\phi_n(s, x)\}$
  end for
until $\|\phi_n - \phi_{n-1}\| < \epsilon$
$\Phi = \phi_n$
$g^* = g_n$
Output: approximation of the value function Φ and a t-optimal decision rule g^*

3.5 Numerical Examples

This section provides numerical examples for the methods discussed so far. Example 1 compares the upper and lower bounds that result from a uniform discretization of the target space with those resulting from the application of the proposal.

Example 2 examines the number of discretization points needed to achieve a similar value of the error integral when using the uniform discretization instead of the discretization resulting from the proposal.

Example 3 illustrates that the error integral tends to zero as the discretization step width tends to zero.

Example 4 investigates the reduction in computation effort that results from the utilization of the decomposition approach.

The following table contains the transition probabilities and the one-stage rewards for all examples. Notice that there are no terminal rewards nor discounting.

s	a	$r(s,a)$	$p(s,a,1)$	$p(s,a,2)$	$p(s,a,3)$
1	1	8	1/2	1/4	1/4
	2	11/4	1/16	3/4	3/16
	3	17/4	1/4	1/8	5/8
2	1	16	1/2	0	1/2
	2	15	1/16	7/8	1/16
3	1	7	1/4	1/4	1/2
	2	4	1/8	3/4	1/8
	3	9/2	3/4	1/16	3/16

Table 3.1: Example 1 to 3, input data

Example 1. We apply the discretization step width $\Delta = 0.01$ and $1 - \beta = 0.9$. The resulting number of discretization points is 102. Figure 3.1 illustrates the distribution of the discretization points resulting from the application of the proposal. The

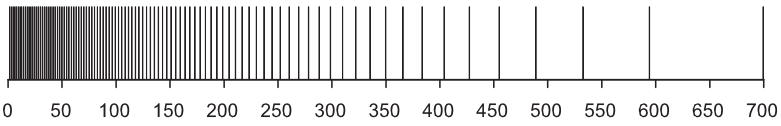

Figure 3.1: Example 1, discretization according to the proposal with $\Delta = 0.01$

increasing gaps between the discretization points result from the use of the structure resulting from the geometrical distributed planning horizon.

Figure 3.2 depicts the uniform discretization scheme. The 102 discretization points are distributed uniformly between the minimal and maximal discretization point of the discretization scheme according to the proposal.

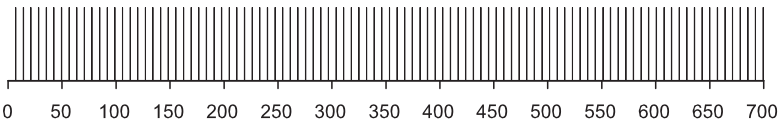

Figure 3.2: Example 1, discretization according to a uniform discretization scheme with $\Delta = 0.01$

Table 3.2 shows the number of iterations needed for the lower and upper bound on the value function resulting from the utilization of both discretization schemes. The application of the proposal leads to a reduced number of iterations for the lower bound.

discretization scheme	iterations lower bound	iterations upper bound
uniform	38	45
proposal	32	45

Table 3.2: Example 1, comparison of number of iterations

Figures 3.3 to 3.5 illustrate the upper and lower bounds on the value function resulting from both discretization schemes for each state of the state space. The red and dark green colored distribution functions correspond to the upper and lower bound on the value function resulting from the discretization scheme of the proposal. The blue and light green colored distribution functions belong to the upper and lower bound on the value function resulting from the uniform discretization scheme.

The upper and lower bounds of the discretization schemes resulting from the proposal are tighter than those resulting from the uniform discretization scheme.

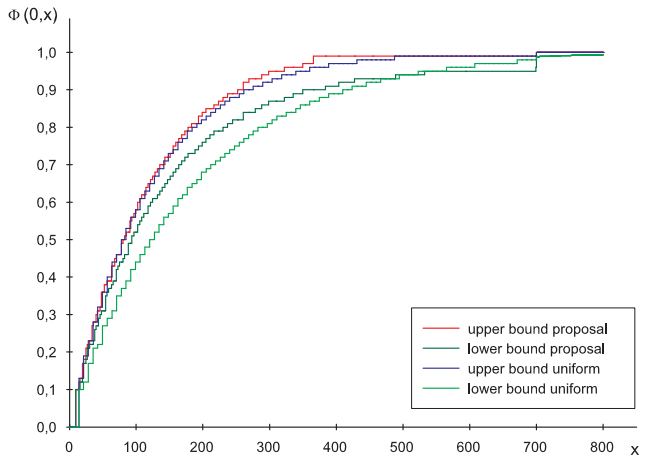

Figure 3.3: Example 1, comparison of the bounds for state $s = 0$

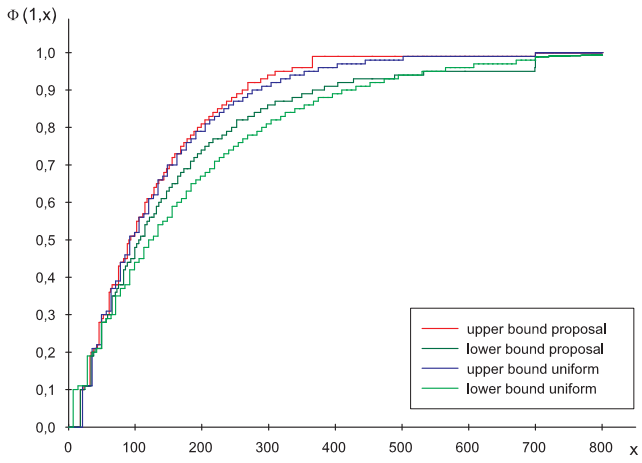

Figure 3.4: Example 1, comparison of the bounds for state $s = 1$

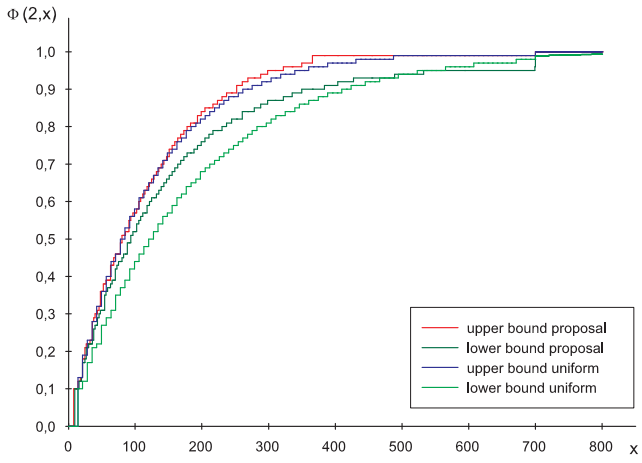

Figure 3.5: Example 1, comparison of the bounds for state $s = 2$

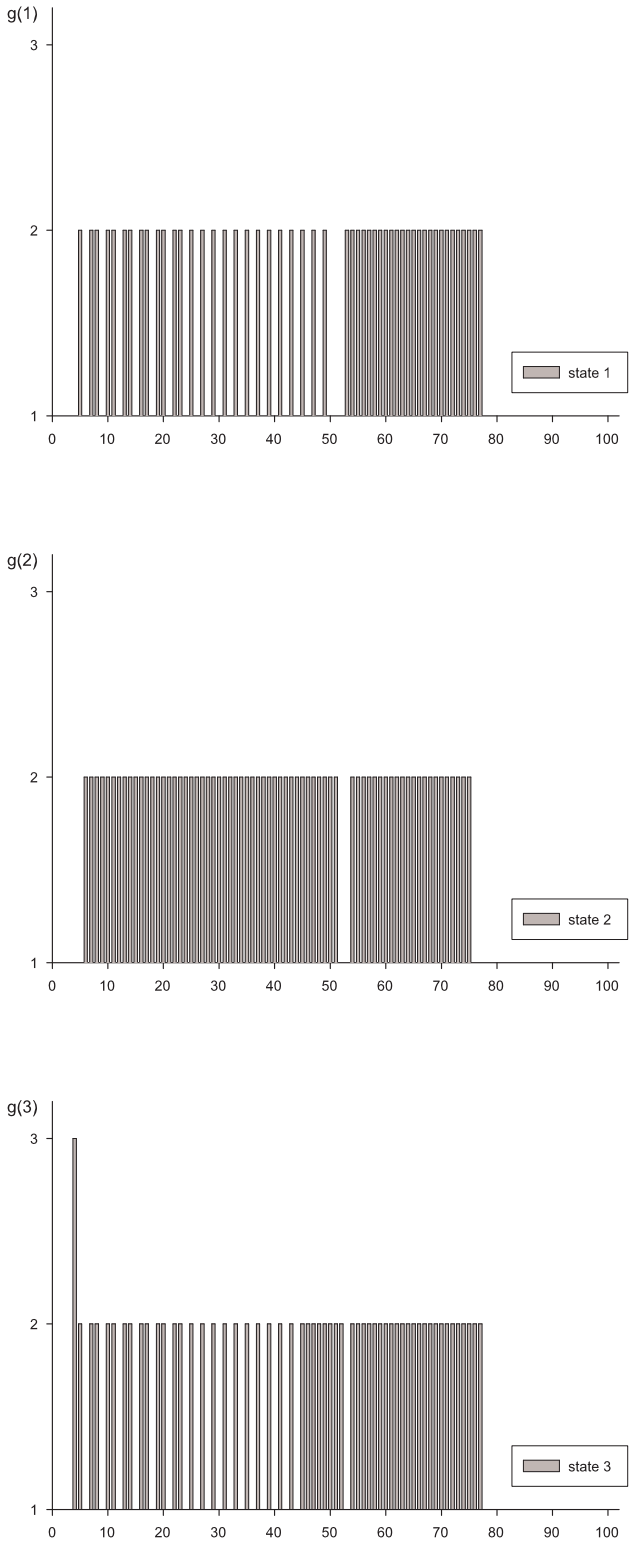

Figure 3.6: Example 1, policies according to the proposal

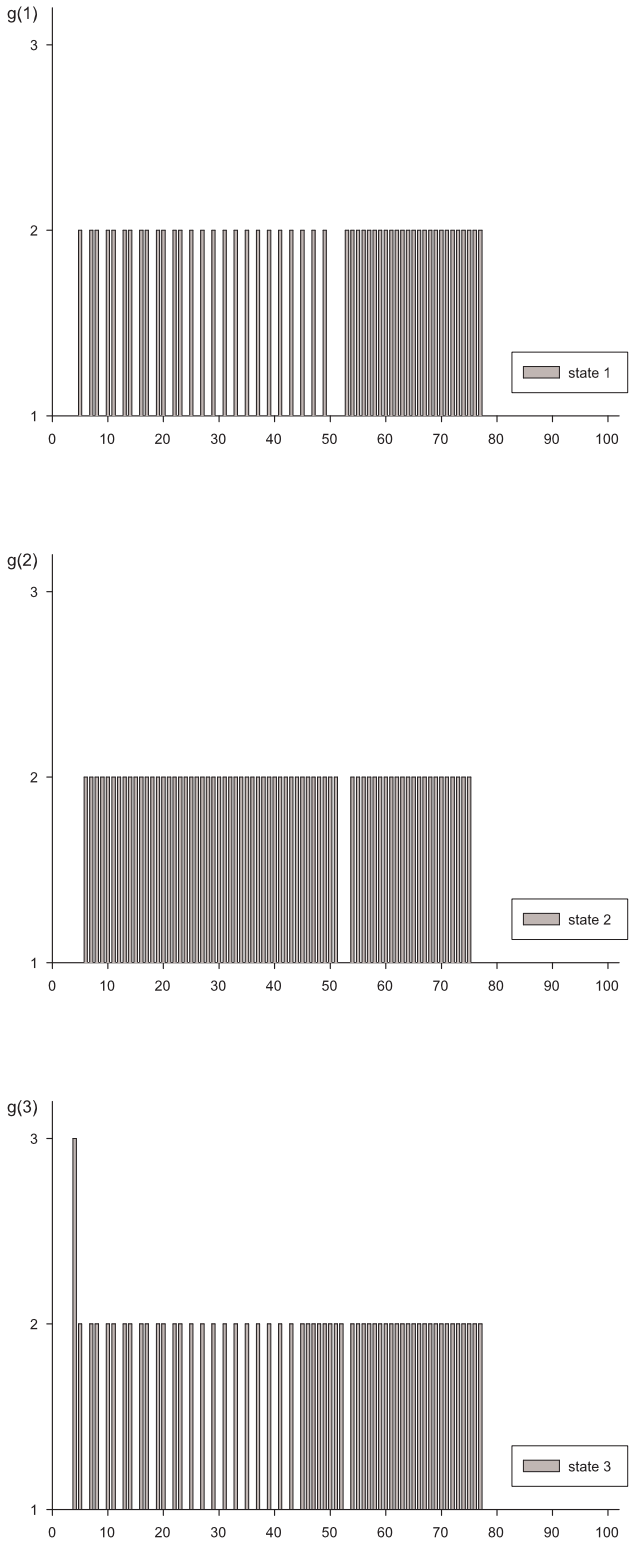

Figure 3.7: Example 1, policies according to a uniform discretization

Table 3.3 shows that the area between the lower and the upper bound of the discretized version are smaller for the discretization scheme resulting from the utilization of the proposal.

discretization scheme	$\sigma_0(\Delta)$	$\sigma_1(\Delta)$	$\sigma_2(\Delta)$
proposal	45.87	45.30	45.86
uniform	66.84	67.32	76.50

Table 3.3: Example 1, discretization error with $1 - \beta = 0.1$

Example 2. Now, we set the parameter of the geometrical distribution to $1 - \beta = 0.04$. Both discretization schemes are constructed in the same way as in Example 1. Figure 3.8 depicts the discretization scheme utilizing the proposal and Figure 3.9 illustrates the uniform discretization scheme.

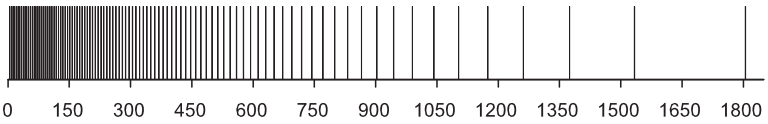

Figure 3.8: Example 2, discretization according to the proposal with $\Delta = 0.01$

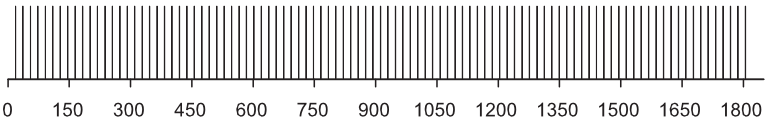

Figure 3.9: Example 2, discretization according to the proposal with $\Delta = 0.01$

Table 3.4 shows the necessary iterations for the upper and lower bounds.

discretization scheme	iterations lower bound	iterations upper bound
proposal	56	92
uniform	92	92

Table 3.4: Example 2, number of iterations for $1 - \beta = 0.04$

Figures 3.10 to 3.12 illustrate the upper and lower bounds on the value function resulting from both discretization schemes. Each figure corresponds to a state of the original problem.

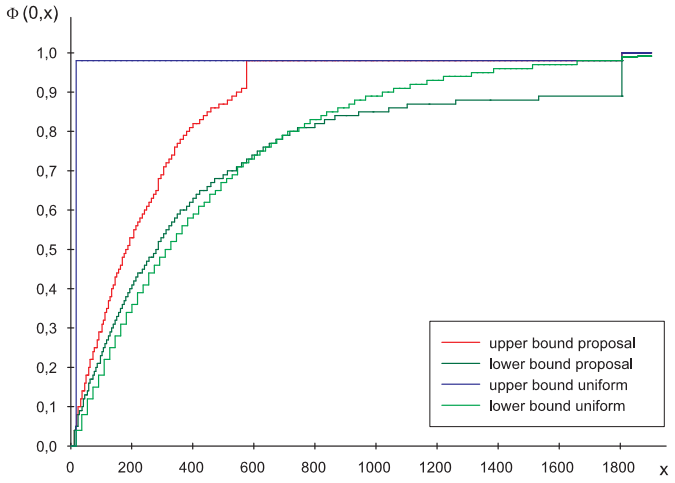

Figure 3.10: Example 2, comparison of the bounds for state $s = 0$

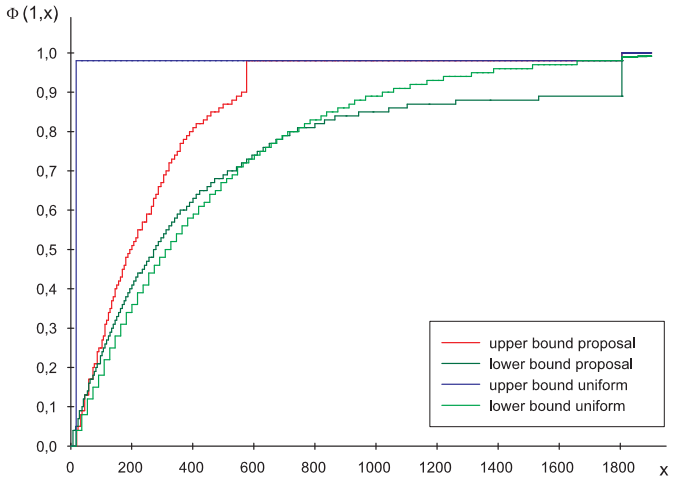

Figure 3.11: Example 2, comparison of the bounds for state $s = 1$

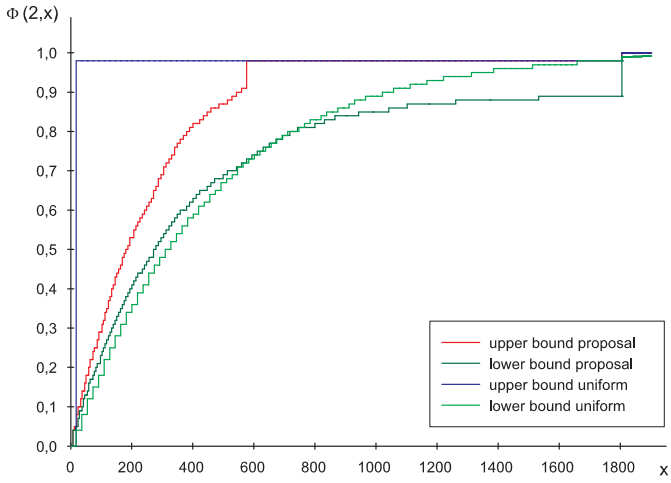

Figure 3.12: Example 2, comparison of the bounds for state $s = 2$

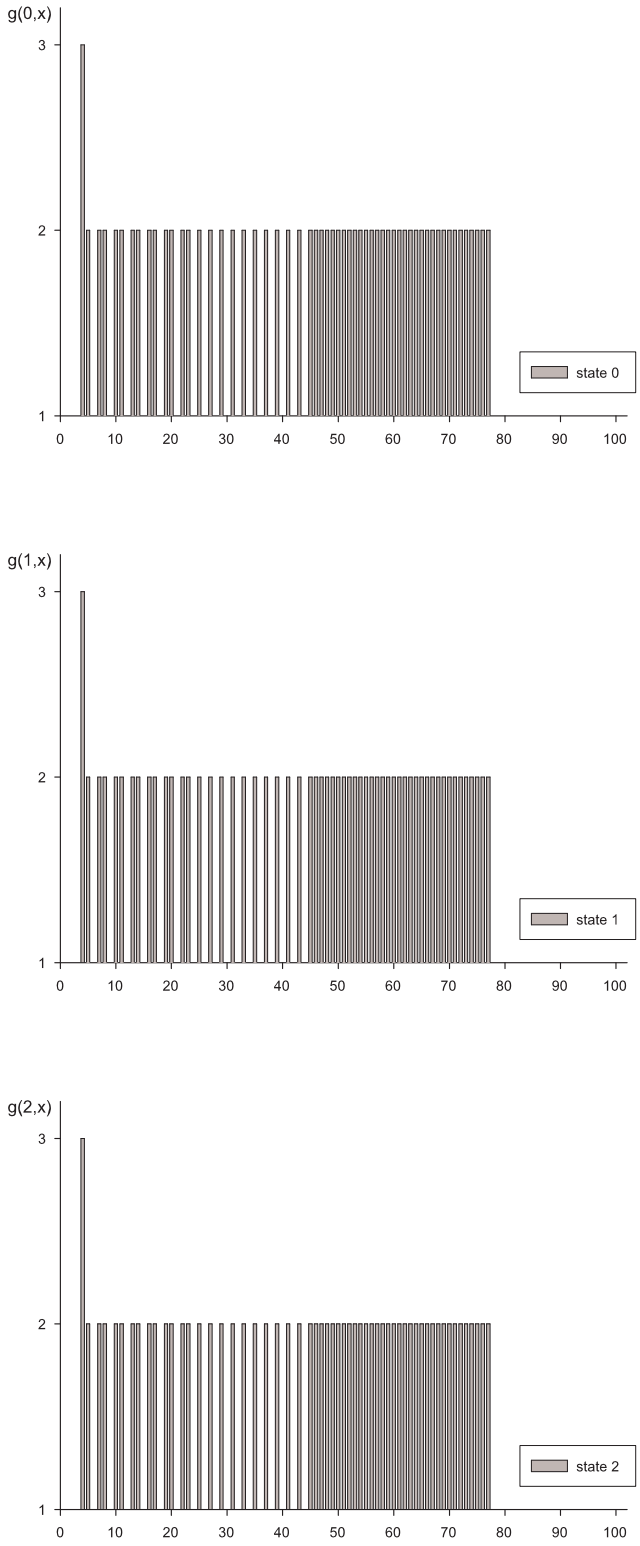

Figure 3.13: Example 2, policies according to the proposal

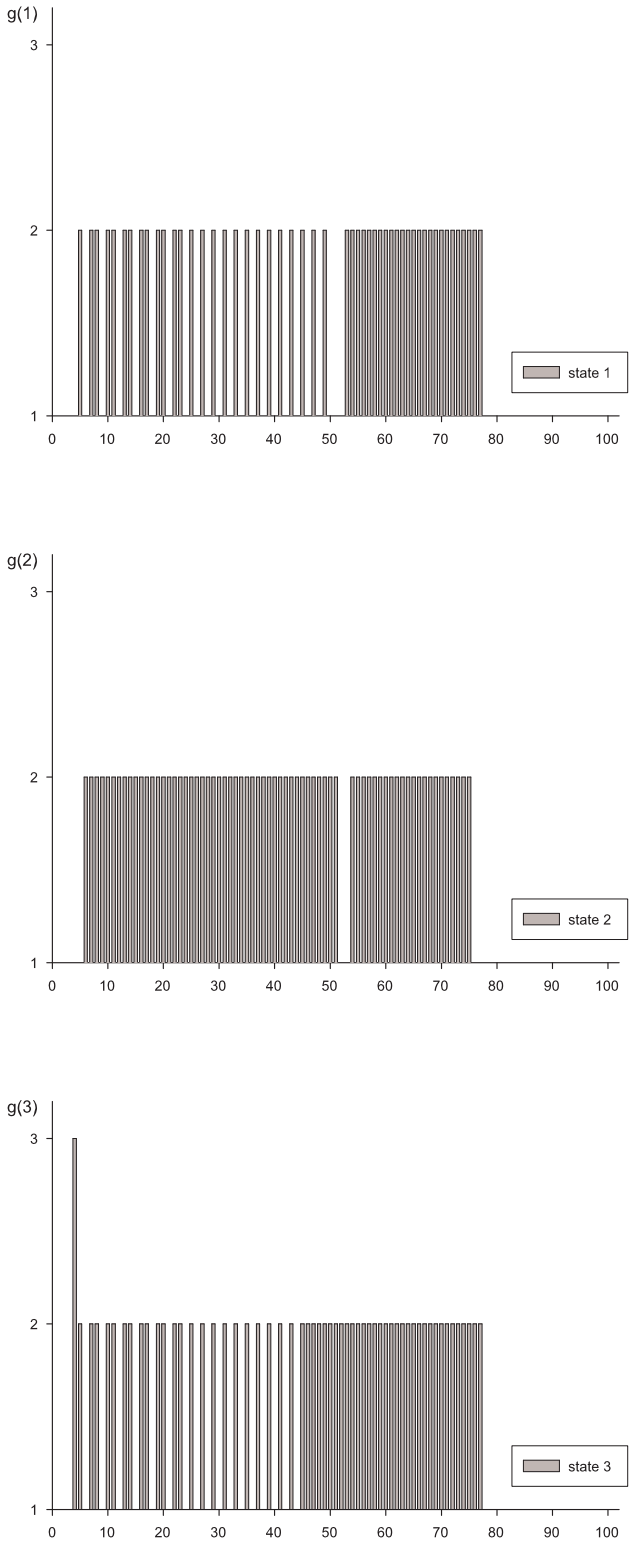

Figure 3.14: Example 2, policies according to uniform discretization

Table 3.5 depicts the discretization error for both discretization schemes. To get a

discretization scheme	$\sigma_0(\Delta)$	$\sigma_1(\Delta)$	$\sigma_2(\Delta)$
proposal	242.23	240.20	242.39
uniform	409.92	409.92	409.92

Table 3.5: Example 2, discretization error with $1 - \beta = 0.04$

comparable discretization error in the second example, we need 202 discretization points instead of 102. The resulting bounds are plotted in Figures 3.15 to 3.17.

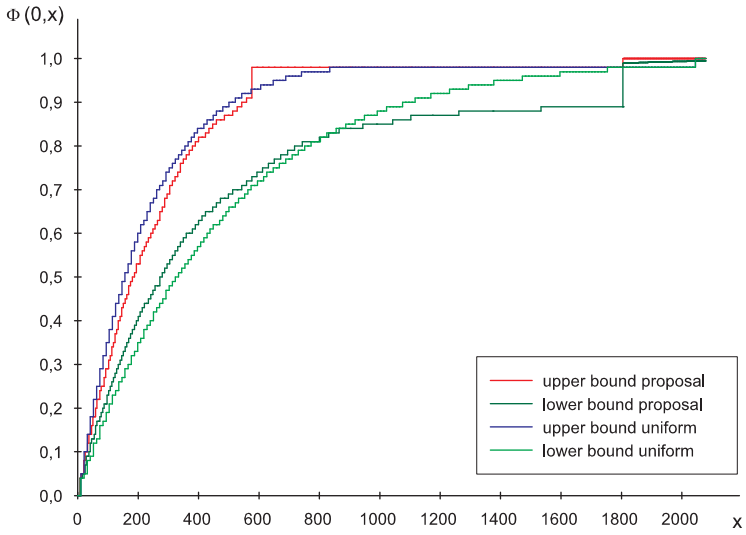

Figure 3.15: Example 2, modified comparison of the bounds for state $s = 0$

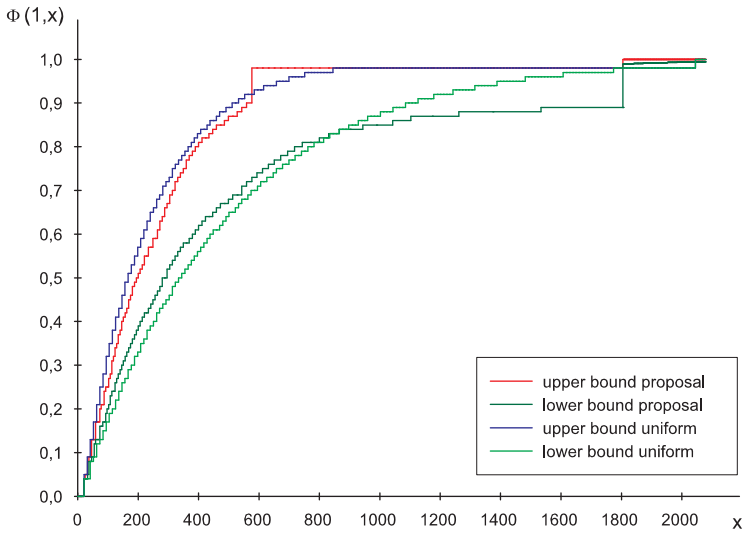

Figure 3.16: Example 2, modified comparison of the bounds for state $s = 1$

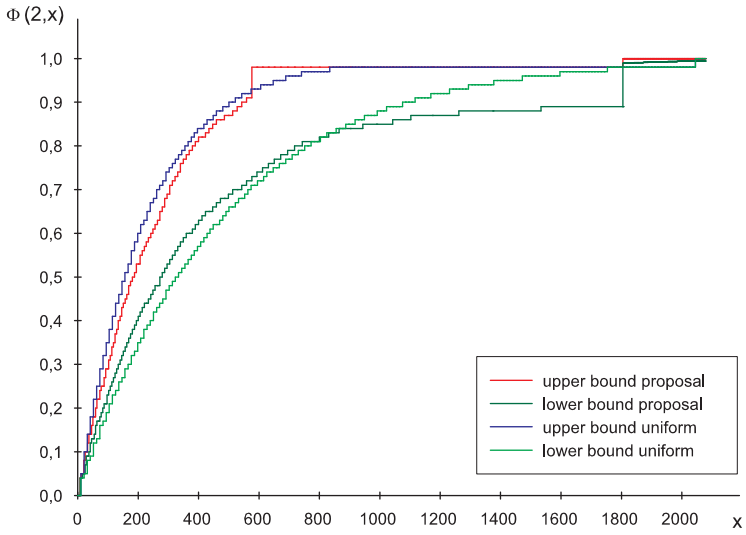

Figure 3.17: Example 2, modified comparison of the bounds for state $s = 2$

discretization scheme	iterations lower bound	iterations upper bound
proposal	56	92
uniform	92	92

Table 3.6: Example 2, number of iterations $1 - \beta = 0.04$

Example 3. In this example, we vary the discretization step width Δ and show that the smaller the discretization step width is, the smaller the value of the error integral is. Figures 3.18 to 3.19 illustrate the upper and lower bounds for $\Delta = 0.0125$ and $\Delta = 0.00625$, respectively.

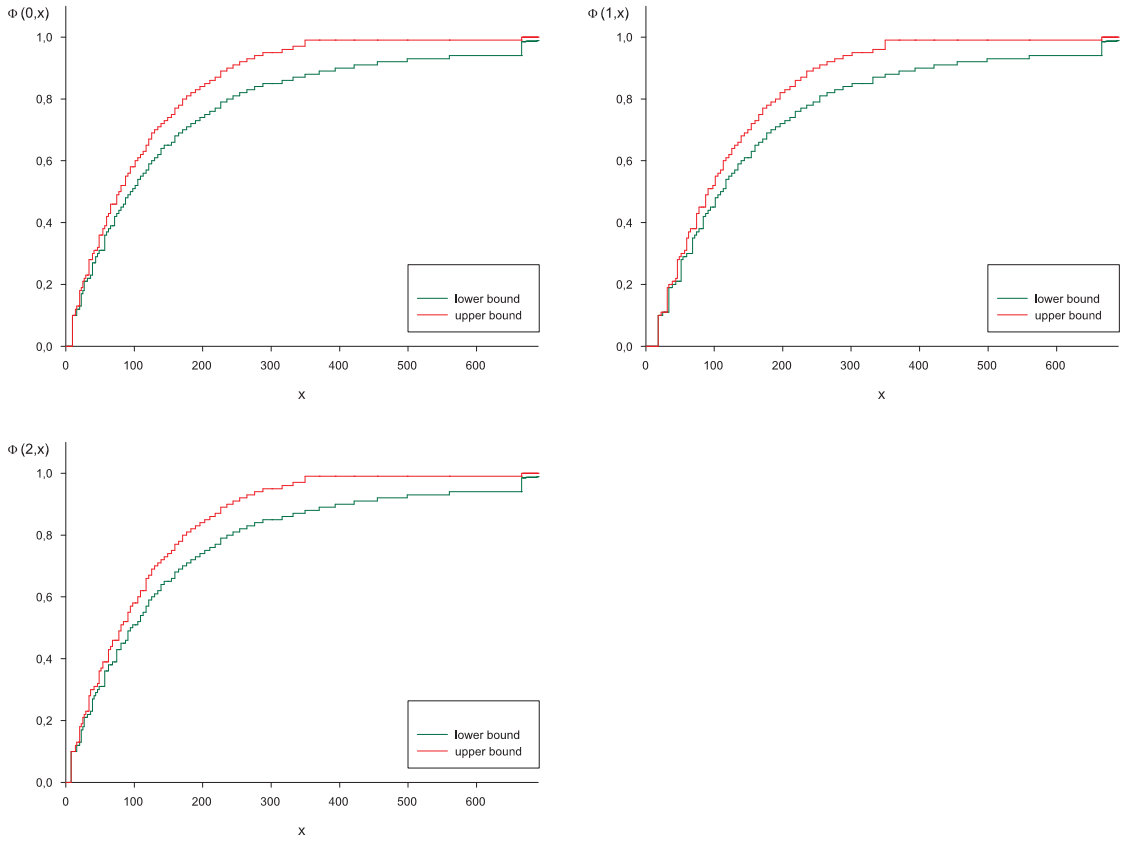

Figure 3.18: Example 3, upper and lower bounds for Δ=0.0125

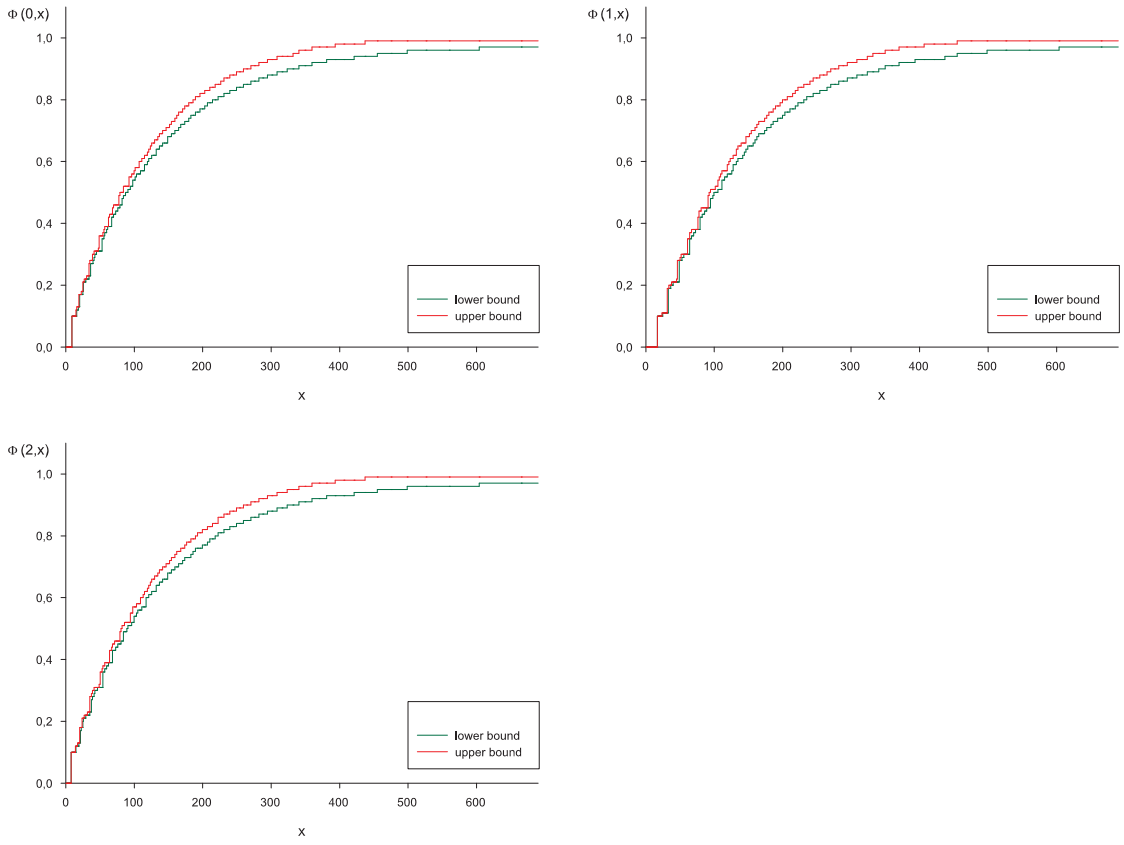

Figure 3.19: Example 2, upper and lower bounds for Δ=0.00625

Δ	iterations lower bound	iterations upper bound
0.0125	56	92
0.00625	92	92

Table 3.7: Example 2, number of iterations for each Δ with $1 - \beta = 0.04$

The evolution of the discretization error is depicted in Table 3.8.

Δ	$\sigma_0(\Delta)$	$\sigma_1(\Delta)$	$\sigma_2(\Delta)$	iter. lower bound	iter. upper bound
0.01	242.23	240.20	242.40	56	92
0.005	130.67	129.77	130.80	74	92
0.0025	60.93	60.61	60.95	90	92
0.00125	28.36	28.37	28.37	92	92
0.000625	14.04	14.07	14.03	92	92
0.0003125	6.85	6.86	6.86	92	92

Table 3.8: Example 2, evolution of the discretization error with $1 - \beta = 0.04$

Example 4. This example shows the effects of the decomposition scheme. Figure 3.20 shows the results of the decomposition scheme for $1 - \beta = 0.1$ and Figure the results for $1 - \beta = 0.04$.

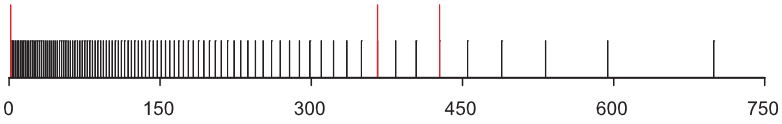

Figure 3.20: Example 4, discretization according to the proposal with $\Delta = 0.01$ and closed sets

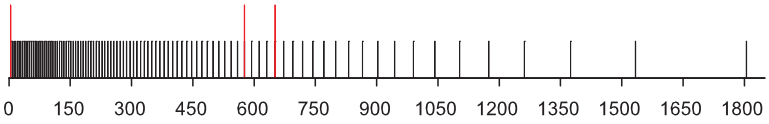

Figure 3.21: Example 4, discretization according to the proposal with $\Delta = 0.01$ and closed sets

Table 3.9 shows the reduction of iterations by utilizing the decomposition approach.

$1 - \beta$	iteration lower bound	iteration upper bound
0.1	28	40
0.04	50	70

Table 3.9: Example 4, reduction of computational effort

Chapter 4

Extensions of the model

It is common to address risk by considering special utility functions (e.g. the exponential utility function) or by penalizing deviations from the mean value in the objective function. The main problem of using standard deviation as a measure of risk is that fluctuations above and below the mean are treated in the same way. For example, an investor aims to maximize the expected return subject to the constraint that the probability of achieving a return below a predefined target level is below a certain threshold. This constraint is called shortfall-constraint and reflects the desire of risk-averse investors to limit the maximum likely loss. We assume that investors have in mind some disaster level of returns and that they aim to minimize the probability of disaster.

In this chapter we combine the risk-neutral approach form Chapter 2 with the Target Value Criterion discussed in Chapter 3. Section 4.1 deals with the Chance-Constrained Programming formulation, where the probability of failing a predetermined target value has to be below some predefined threshold and is treated as an additional constraint. Chance-Constrained Programming dates back to the original paper of Charnes and Cooper (1959) and is primarily used in the domain of Stochastic Programming. In Section 4.2 we combine both criteria in a hybrid approach as a weighted sum in the objective function. The optimality equation and structural properties are discussed. In order to derive a numerical solution Section 4.2.1 contains a discretization scheme of the target space based on the discretization approach contained in Chapter 3. In Section 4.2.2 we discuss a decomposition method

that reduces the computational effort. Section 4.2.3 deals with solution methods and especially an extrapolation method for the Value Iteration Algorithm. Section 4.2.4 contains the limit behavior of the hybrid model that results from varying the penalty cost k. Section 4.3 provides an algorithm that treats the penalty cost k in a parametric programming manner. The value functions and corresponding policies in dependence on k are calculated in an efficient way. The chapter closes with several numerical examples.

4.1 Chance Constraint Approach

Chance-Constrained Programming belongs to the common approaches for treating random parameters in optimization problems. Typical application domains are engineering and finance, where uncertainties like product demand, meteorological or demographic conditions, currency exchange rates, enter the inequalities describing the proper working of a system under consideration. Often a constrained violation can be compensated. For example power generating companies can buy energy on the liberalized market if they are faced with unforeseen peaks of electrical load. As long as the costs of compensating decisions are known, these may be considered as a penalization for constrained violation.

In our model, we assume that the cost of compensating decisions are known. That is why we state the Chance-Constrained Programming formulation in a brief way, but turn our attention towards the hybrid approach in Section 4.2.

In the case of $\alpha = 1$, given initial state s and target value x, maximize the total expected reward

$$E_\delta \left[\sum_{t=0}^{\tau-1} r(\zeta_t, g_t(\zeta_t, \xi_t)) + h(\zeta_\tau) | \zeta_0 = s, \xi_0 = x \right]$$

with respect to $g \in G^\infty$ subject to subject to

$$P_\delta \left(\sum_{t=0}^{\tau-1} r(\zeta_t, g_t(\zeta_t, \xi_t)) + h(\zeta_\tau) < x | \zeta_0 = s, \xi_0 = x \right) \leq \overline{\Phi}$$

for some predetermined constant $\overline{\Phi} \in [0,1]$. The problem can be used by applying standard methods of linear programming.

4.2 Multi-criteria approach

Mixed criteria are linear combinations of standard criteria that cannot be represented as standard criteria. We consider linear combinations of both the target value criterion and the total reward criterion as a weighted sum. The penalty cost $k \in \mathbb{R}, k \geq 0$ quantify the costs of compensating a constrained violation. That means, the expected discounted total reward is maximized an the probability of missing a predetermined target is penalized.

Again, we allow the control to depend on the actual state $s_t \in S$ and the updated target $x_t \in X := \mathbb{R}$, to be realized in the remaining time. The decision rules and stationary policies are defined analogously to the previous chapter. Now, for policy $\delta = (g_0, g_1, \ldots) \in G^\infty$, initial value $(s,x) \in S \times X$ and penalty cost $k \in \mathbb{R}, k \geq 0$, let

$$\begin{aligned} \mathcal{V}_\delta(s,x) &= E_\delta \left[\sum_{t=0}^{\tau-1} r(\zeta_t, g_t(\zeta_t, \xi_t)) | \zeta_0 = s, \xi_0 = x \right] \\ &\quad -kP_\delta \left(\sum_{t=0}^{\tau-1} r(\zeta_t, g_t(\zeta_t, \xi_t)) < x | \zeta_0 = s, \xi_0 = x \right) \end{aligned}$$

be the objective function. Using the reformulation without discounting, the formulation can be simplified by using the results of the former chapter.

First we show that the expected discounted reward for a given policy g fulfills the functional equation $\mathcal{V}_g = \mathcal{U}_g \mathcal{V}_g$ and that the solution can be obtained by the method of successive approximations. These results are summarized in Proposition 5.

Proposition 5. *Let $g \in G$ and $h = 0$. Then $\mathcal{V}_g$ is the unique solution to $\mathcal{V}_g = \mathcal{U}_g \mathcal{V}_g$ and we have that $\mathcal{V}_g = \lim_{n\to\infty} \mathcal{U}_g^n \nu, \nu \in \mathfrak{V}$.*

Proof. For notational convenience, let $E_{g,s,x} := E_g(\cdot|\zeta_0 = s, \xi_0 = x)$ and $P_{g,s,x}(\cdot|\zeta_0 = s, \xi_0 = x)$, respectively. Set $a := g(s,x)$. Then we have

$$\begin{aligned}
\mathcal{V}_g(s,x) &= E_{g,s,x}\left[\sum_{t=0}^{\tau-1}\alpha^t r(\zeta_t, g(\zeta_t,\xi_t))\right] - k\cdot P_{g,s,x}\left[\sum_{t=0}^{\tau-1}\alpha^t r(\zeta_t, g(\zeta_t,\xi_t)) < x\right] \\
&= V_g(s,x) - k\cdot\Phi_g(s,x) \\
&= r(s,a) + \alpha\beta\sum_{s'\in S} p(s,a,s')V_g(s', x - r(s,a)) \\
&\quad -k\cdot\left[(1-\alpha\beta)1_{(0,\infty))}(x - r(s,a)) + \alpha\beta\sum_{s'\in S} p(s,a,s')\Phi_g(s', x - r(s,a))\right] \\
&= r(s,a) - k(1-\alpha\beta)1_{(0,\infty)}(x - r(s,a)) \\
&\quad +\alpha\beta\left[\sum_{s'\in S} p(s,a,s')\left[V_g(s', x - r(s,a)) - k\cdot\Phi_g(s, x - r(s,a))\right]\right] \\
&= r(s,a) - k(1-\alpha\beta)1_{(0,\infty)}(x - r(s,a)) \\
&\quad +\alpha\beta\sum_{s'\in S} p(s,a,s')\mathcal{V}_g(s', x - r(s,a))
\end{aligned}$$

□

Now we are in a position to state the optimality equation, the corresponding policies, the convergence of the method of successive approximations and properties of the value function.

Theorem 9.

(i) $\mathcal{V}(s,x)$ *is the unique solution in* $\mathfrak{V}$ *for the optimality equation* $\mathcal{V} = \mathcal{U}\mathcal{V}$*, i.e. we have for all* $s \in S, x \in X$

$$\begin{aligned}
\mathcal{V}(s,x) &= \max_{a\in D(s)}\Big\{r(s,a) - k(1-\alpha\beta)1_{(0,\infty)}(x - r(s,a)) \\
&\quad +\alpha\beta\sum_{s'\in S} p(s,a,s')\mathcal{V}(s, x - r(s,a))\Big\}.
\end{aligned}$$

(ii) Each decision rule $g^* \in G$ *formed by actions* $g^*(s,x)$*, minimizing the right hand side of (xx) (i.e., for which* $\mathcal{U}\mathcal{V} = \mathcal{U}_{g^*}\mathcal{V}$ *holds) is optimal.*

(iii) $\mathcal{V} = \lim_{n\to\infty}\mathcal{U}^n\nu, \nu \in \mathfrak{V}$.

(iv) $\mathcal{V}(s,\cdot), s \in S$, *is decreasing and left continuous in* x.

(v) For $s \in S$, *the smallest (largest) minimizer* $g^* \in G$ *of (xx) is left continuous in* x.

Proof. The proof is in analogy to the proof contained in Theorem 6. □

4.2.1 Discretization of the target space

For numerical calculations it is necessary to approximate the target space X by a finite set X_Δ. The discretization is based on the former chapter and adopted to the new context. Due to the mixed-criterion approach, we have to adapt the upper and lower bound of the value function. Based on $r^\pm, h^\pm$ and $\rho_x^\pm : \mathbb{N}_0 \to \mathbb{R}, x \in \mathbb{R}$, defined by $\rho_x^\pm(t) := x - (t+1)r^\mp - h^\mp, t \in \mathbb{N}_0$, introduce $\kappa^\pm : X \to \mathbb{R}$,

$$\kappa^\pm(x) := \frac{r^\pm}{(1-\beta)} - k(1-\beta)\sum_{t=0}^{\infty}\beta^t 1_{(0,\infty)}(\rho_x^\mp(t)), x \in X.$$

It is easily verified that $\kappa^+(x)$ and $\kappa^-(x)$ are upper and lower bounds to $\mathcal{V}_\delta(\cdot,x), \delta \in G^\infty$. In fact, for $(s,x) \in S \times X$,

$$\begin{aligned}
\mathcal{V}_\delta(s,x) &\leq E_\delta\left[\sum_{t=0}^{\tau-1}\alpha^t r^+ | \zeta_0 = s, \xi_0 = x\right] - kP_\delta\left(\sum_{t=0}^{\tau-1}\alpha^t r^- + \alpha^\tau h^- < x | \zeta_0 = s, \xi_0 = x\right) \\
&= \frac{r^+}{1-\beta} - k\sum_{\nu=1}^{\infty}(1-\beta)\beta^{\nu-1}1_{(0,\infty)}(\rho_x^+(\nu-1)) \\
&= \kappa^+(x).
\end{aligned}$$

Since the bounds are independent of δ, the same holds for V, that is

$$\kappa^-(x) \leq \mathcal{V}(s,x) \leq \kappa^+(x), (s,x) \in S \times X. \tag{4.1}$$

Proposition 6. *It holds:*

(i) If $r^- \geq 0$, *then* $\frac{r^-}{1-\beta} - k(1-\beta^{t_1^+(x)}) \leq \mathcal{V}(\cdot,x) \leq \frac{r^+}{1-\beta} - k(1-\beta^{t_1^-(x)})$ *for* $x > r^- + h^-$ *and* $\frac{r^-}{1-\beta} \leq \mathcal{V}(\cdot,x) \leq \frac{r^+}{1-\beta}$, *otherwise.*

(ii) *If* $r^+ \leq 0$, *then* $\frac{r^-}{1-\beta} - k\beta^{t_0^+(x)} \leq \mathcal{V}(\cdot, x) \leq \frac{r^+}{1-\beta} - k\beta^{t_0^-(x)}$ *for* $x \leq r^+ + h^+$ *and* $\mathcal{V}(\cdot, x) = E(s, x)$, *otherwise.*

(iii) *If* $r^- < 0 < r^+$, *then* $1 - \beta^{t_1^-(x)} \leq \mathcal{V}(\cdot, x) \leq \beta^{t_0^+(x)}$ *for* $x \in X$.

Additionally we have $\lim_{x\to-\infty} \mathcal{V}(s, x) = E(s, x)$ *and* $\lim_{x\to\infty} \mathcal{V}(s, x) = E(s, x) - k$ *for* $s \in S$.

Proof. The proof is a direct consequence of the former chapter. □

4.2.2 Decomposition of the discretized MDP

In order to improve the efficiency of the optimization procedure, we present a decomposition scheme for the discretized MDP that is based on the decomposition scheme presented in the former chapter.

The discretization of the target space leads to a partition $J_1 \cup J_2 \ldots \cup J_m$ of the state space $S \times X_\Delta$ consisting of $m \in \mathbb{N}$ closed subsets $J_1, \ldots, J_m$. The partition allows us to determine $V_\Delta^\pm$, separately by restricting the corresponding optimality equations to the closed subset under consideration. Some of these subsets are also absorbing for which $V_\Delta^\pm$ is already known in advance.

Example 5. Suppose we are interested in calculating $\mathcal{V}_\Delta^\pm$ on the basis of X_Δ introduced in the proposal. Let $\beta > 0.5$. We consider two cases:

(1) $0 \leq r^- < r^+$.
Recall that $X_\Delta = \{-\infty, x_0, x_1, \ldots, x_{k-1}, \infty\}$. The determination of j^* and j^{**} from the former chapter leads to the following partition of the state space. The sets $S \times \{x_{k-j^{**}}\}, \ldots, S \times \{x_{k-1}\}, S \times \{\infty\}$ are $\lceil\cdot\rceil_\Delta$-1-absorbing. Consequently, $\mathcal{V}(s, x_j) = E(s, x_j) - k$ within this set. The set $\{-\infty, x_0\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing. Consequently, $\mathcal{V}(s, x_j) = E(s, x_j)$ within this set. These observations allow us to decompose the optimization problem: First solve $\mathcal{V}_\Delta^+ = \mathcal{U}\mathcal{V}_\Delta^+$ on $S \times \{x_0, \ldots, x_{k-j^*-1}\}$ using that $S \times \{x_0\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing and $\mathcal{V}^+ = E(s, x_0)$. Then, for $j^{**} < j \leq j^*$, solve $\mathcal{V}_\Delta^+ = \mathcal{U}^+\mathcal{V}_\Delta^+$ on $S \times \{x_{k-j}\}$. On the remaining subset $S \times \{x_{k-j^{**}}, \ldots, \infty\}$, $\mathcal{V}_\Delta^+$ is $\lceil\cdot\rceil_\Delta$-1-absorbing and thus $\mathcal{V}^+ = E(s, x_0) - k$.

(2) $r^- < 0 < r^+$.

Recall that $X_\Delta^+ = \{-\infty, x_{-k+1}, \dots, x_{-1}, 0, x_1, \dots, x_{k-1}, \infty\}$ with representatives x_j. Define j^* and j^{**} as in (1). Then $S \times \{x_{k-j^{**}}, \dots, \infty\}$ is $\lceil\cdot\rceil_\Delta$-1-absorbing, and $S \times \{x_{k-j^*}, \infty, x_{k-1}, \infty\}$ is $\lceil\cdot\rceil_\Delta$-closed.

Since we round up, an analogous line of argumentation is not possible for negative values of X_Δ. However, for small enough j, that is for $1 \leq j^{***} \leq k$ such that $r^- \ln(j/(j+1)) < r^+ \ln \beta$, the sets $S \times \{x_{-k+j^{***}}, \dots, x_{k-1}, \infty\}$ is $\lceil\cdot\rceil_\Delta$-closed. Finally, set $S \times \{-\infty\}$ is $\lceil\cdot\rceil_\Delta$-0-absorbing.

These observations allow us to decompose the optimization problem: First solve $\mathcal{V}_\Delta^+ = \mathcal{U}\mathcal{V}_\Delta^+$ on the subset $S \times \{x_{k-j^*}, \dots, \infty\}$ using that $S \times \{x_{k-j^{**}}, \dots, \infty\}$ is $\lceil\cdot\rceil_\Delta$-1-absorbing. Then solve $\mathcal{V}_\Delta^+ = \mathcal{U}\mathcal{V}_\Delta^+$ on the subset $S \times \{x_{-k+j^{***}}, \dots, \infty\}$ using that Φ_Δ^+ is already known on $S \times \{x_{k-j^*}, \dots, \infty\}$ using that $\mathcal{V}_\Delta^+$ is already known on $S \times \{x_{j-j^*}, \dots, \infty\}$. Finally, for $j < j^{***}$, based on the knowledge of Φ_Δ^+ on $S \times \{x_{-k+j+1}, \dots, \infty\}$ solve $\mathcal{V}_\Delta^+ = \mathcal{U}\mathcal{V}_\Delta^+$ on $S \times \{x_{-k+j}, \dots, \infty\}$.

4.2.3 Extrapolation

The convergence of the value iteration algorithm is usually slow. Combining the value iteration (ν_n) with an extrapolation giving monotone upper and lower bounds

$$\begin{aligned}
\omega_n^+(s,x) &= \nu_n(s,x) + \frac{\alpha\beta}{1-\alpha\beta} \sup_{(s',x)\in S\times X_\Delta} \{\nu_n(s',x) - \nu_{n-1}(s',x)\} \\
\omega_n^-(s,x) &= \nu_n(s,x) + \frac{\alpha\beta}{1-\alpha\beta} \inf_{(s',x)\in S\times X_\Delta} \{\nu_n(s',x) - \nu_{n-1}(s',x)\}, (s,x) \in S \times X_\Delta
\end{aligned}$$

Theorem 10. *For all $n \in \mathbb{N}$ and all $(s,x) \in S \times X$ it holds that*

(i) $\omega_n(s,x) \leq \omega_{n+1}^-(s,x) \leq \mathcal{V}(s,x) \leq \omega_{n+1}^+(s,x) \leq \omega_n^+(s,x)$,

(ii) $\lim_{n\to\infty} \omega_n^-(s,x)) = \lim_{n\to\infty} \omega_n^+(s,x) = \mathcal{V}(s,x)$.

(iii) Let $g_n \in G$ with $\nu_n = \mathcal{U}\nu_{n-1}$. Then it holds that $\mathcal{V}_{g_n} \geq \omega_n^-$.

Proof. The proof is in analogy to the proof contained in Waldmann and Stocker (2013). □

Algorithm 8 Value Iteration with Extrapolation, Penalty Approach

Input: $n = 0, \nu_0 \in \mathfrak{V}, \varepsilon > 0$
repeat
$n = n + 1$
for all $(s, x) \in S \times X_\Delta$ **do**

$$\begin{aligned}\nu_n(s,x) &= \max_{a \in D(s)} \big\{ r(s,a) - k(1-\alpha\beta)1_{(0,\infty)} \\ &\qquad + \alpha\beta \sum_{s' \in S} p(s,a,s')\nu_{n-1}(s', x - r(s,a)) \big\}\end{aligned}$$

$g_n(s,x) = \arg\max \{\nu_n(s,x)\}$
end for
for all $(s, x) \in S \times X_\Delta$ **do**
$\omega_n^-(s) = \nu_n(s) + \dfrac{\alpha\beta}{1-\alpha\beta} \inf\limits_{(s',x) \in S} \{v_n(s) - v_{n-1}(s)\}$
$\omega_n^+(s) = \nu_n(s) + \dfrac{\alpha\beta}{1-\alpha\beta} \sup\limits_{(s',x) \in S} \{v_n(s) - v_{n-1}(s)\}$
end for
until $\|\omega_n^+ - \omega_n^-\| < 2\epsilon$
$\mathcal{V} = (\omega_n^- + \omega_n^+)/2$
$f^* = f_n$
Output: approximation of the value function V^τ and a r-optimal decision rule f^*

4.2.4 Limit behavior

Reformulating $V(s,x)$ as $V^{(k)}(s,x)$ in order to explicitly express the dependence on k, we are in a position to obtain the models and as special cases for $k \to 0$ the convergence towards the risk-neutral model and for $k \to \infty$ the convergence towards the negative Target Value model, respectively. The set F_k and F_k^* denotes the set of decision rules and optimal decision rules depending on k.

Theorem 11. *For all $(s,x) \in S \times X_\Delta$ it holds that*

(i) $\lim\limits_{k \to 0} V^{(k)}(s,x) = V^{(0)}(s,x) = \hat{V}(s)$

(ii) $\lim\limits_{k \to \infty} \dfrac{V^{(k)}(s,x)}{k} = -\Phi(s,x).$

(iii) There exists some $k_0 > 0$ such that $F_k^* \subset F^*$ for all $k < k_0$.

(iv) There exists some $k_0 > 0$ such that $F_k^* \subset G^*$ for all $k > k_0$.

Proof. (i) For $k \geq 0$, introduce $U^{(k)}$ such that $V^{(k)} = U^{(k)}V^{(k)}$. Note that

$$U^{(0)}v - k(1-\beta) \leq U^{(k)}v \leq U^{(k)}v \leq U^{(0)}v, \text{ for all } v \in S \times X_\Delta.$$

Together wit the monotonicity of $U^{(k)}$ it then follows that

$$V^{(0)}(s,x) - k \leq V^{(k)}(s,x) \leq V^{(0)}(s,x).$$

Moreover, $V^{(0)}(s,x)$ is easily seen to be independent of x. Thus (i) holds.

(ii) For $n \in \mathbb{N}$, set $v_n^{(k)} := (U^{(k)})^n 0$ and let $f_n \in F$ such that $\Phi_n = U_{f_n}\Phi_{n-1}$. Then $v_n^{(k)} \geq U_{f_n}^{(k)} v_{n-1}^{(k)}$ and it easily follows by induction on n that

$$\lim_{k\to\infty} \frac{v_n^{(k)}}{k} \geq -\Phi_n, n \in \mathbb{N}.$$

Now let $\epsilon > 0$ arbitrary. Then there exists $k_0(\epsilon)$ and $n_0(\epsilon)$ such that

$$\frac{\left\| v_n^{(k)} - V^k \right\|}{k} \leq \epsilon, n \geq n_0(\epsilon), k \geq k_0(\epsilon).$$

We finally get

$$\lim_{k\to\infty} \frac{V^{(k)}(s,x)}{k} \geq \lim_{k\to\infty} \frac{v_n^{(k)}(s,x)}{k} - \epsilon \geq -\Phi(s,x) - \epsilon.$$

To show the reverse inequality, first observe that, for $\epsilon > 0$, there exists $k_0(\epsilon)$ such that

$$\left\| \frac{V^{(k)}}{k} - \frac{V^{(k')}}{k} \right\| \leq \epsilon, k, k' \geq k(\epsilon).$$

Next, fix $V(s,x)$ and let $a_k = f_n^{(k)}(s,x)$ a maximizing action state (s,x) of iteration n. Then there exists a convergent subsequence (a_{k_m}) which is constant

for k_m sufficiently large since $D(s)$ is finite. Thus

$$\lim_{m\to\infty} \frac{v_n^{k_m}(s,x)}{k_m} \leq -\Phi(s,x).$$

Finally,

$$\lim_{k\to\infty} \frac{V^{(k)}}{k} \leq \Phi,$$

which competes the proof of (ii).

(iii) Suppose that $F_k^* \subset F^*$ does not hold for all $k < k_0$ and some $k_0 > 0$. Then there exists a sequence $(k_m) \to 0$ and a sequence $f_{n_1}^*$ of k_m-optimal decision rules fulfilling $f_{n_1}^* \notin F^*$. Since the set F_k is finite there is some subsequence (k_{m_k}) with constant $f_{m_k}^* = g$, say. Now, applied to g,we have for some $(s,x) \in S \times X_\Delta$,

$$\begin{aligned} \lim_{k_{m_k}\to 0} V^{k_{m_k}}(s,x) &= \lim_{k_{m_k}\to 0} V_g^{k_{m_k}}(s,x) \\ &= \hat{V}_g(s,x) \\ &< \hat{V}(s,x). \end{aligned}$$

On the other hand, for all $f^* \in F^*$,

$$\begin{aligned} \lim_{k_{m_k}\to 0} V^{k_{m_k}}(s,x) &\geq \lim_{k_{m_k}\to 0} V_{f^*}^{k_{m_k}}(s,x) \\ &= \hat{V}_{f^*}(s,x) \\ &= \hat{V}(s,x), \end{aligned}$$

which is the desired contradiction. Hence $F_k^* \subset F^*$ holds for all $k < k_0$.

(iv) The proof is in analogy to the proof contained in (iii).

□

4.3 Parametric penalty cost

In this section we consider the penalty cost as a real parameter. We solve the problem for the whole parameter range. We start with an initial policy and a combination of

two right hand sides risulting from the risk-neutral and the Target Value approach. In a second step, the policy improvement phase is deployed. For each pair of actions, the intersection point k is calculated. As k depends on the chosen actions and k is bounded from below we have a minimum value for the intersection points. The minimum value of the intersection points is the upper end of the stability interval of the actual policy. That minimum value is the upper end of the stability interval for the actual policy.

$$W_g(s,x) = V_g(s,x) - k \cdot \Phi_g(s,x).$$

Algorithm 9 Policy Iteration, Parametric Approach

Input: MDP, initial decision rule $f_o \in G$
$n = 0$
(Policy evaluation)
Calculate V_{f_n} as a solution of the linear equation system
$V_{f_n}(s,x) = \overline{r}(s, f_n(s,x))) + \alpha\beta \sum_{s' \in S} p(s, f_n(s,x), s') V_{f_n}(s,x), (s,x) \in S \times X.$
Calculate $\Phi_{f_n}(s,x)$ as a solution of the linear equation system
$\Phi_{f_n}(s,x) = c(s,x,f_n(s,x)) + \alpha\beta \sum_{s' \in S} p(s, f_n(s,x), s') \Phi_{f_n}(s', x - r(s,a)).$
(Policy improvement and stability)
for all $a \in D(s)$ **do**
  for all $(s,x) \in S \times X$ **do**
    $V_a(s,x) = \overline{r}(s,a) + \alpha\beta \sum_{s' \in S} p(s,a,s') V_{f_n}(s',x)$
    $\Phi_a(s,x) = c(s,x,a) + \alpha\beta \sum_{s' \in S} p(s,a,s') \Phi_{f_n}(s', x - r(s,a))$
    $\tilde{V}_a(s,x) = V_a(s,x) - k\Phi_a(s,x)$
  end for
end for
for all $a \in D(s)$ **do**
  Find minimum intersection point $k_{\min}$.
end for

4.4 Numerical Examples

Example 6. This example shows the limiting behavior. For a penalty cost of 0 the risk neutral solution is gained. With increased penalty cost the penalty solution approaches the solution of the target value. Figure shows the limit behavior of the value function. The lower and upper bound are plotted for penalties $k_1 = 0, k_2 = 100, k_3 = 100$.

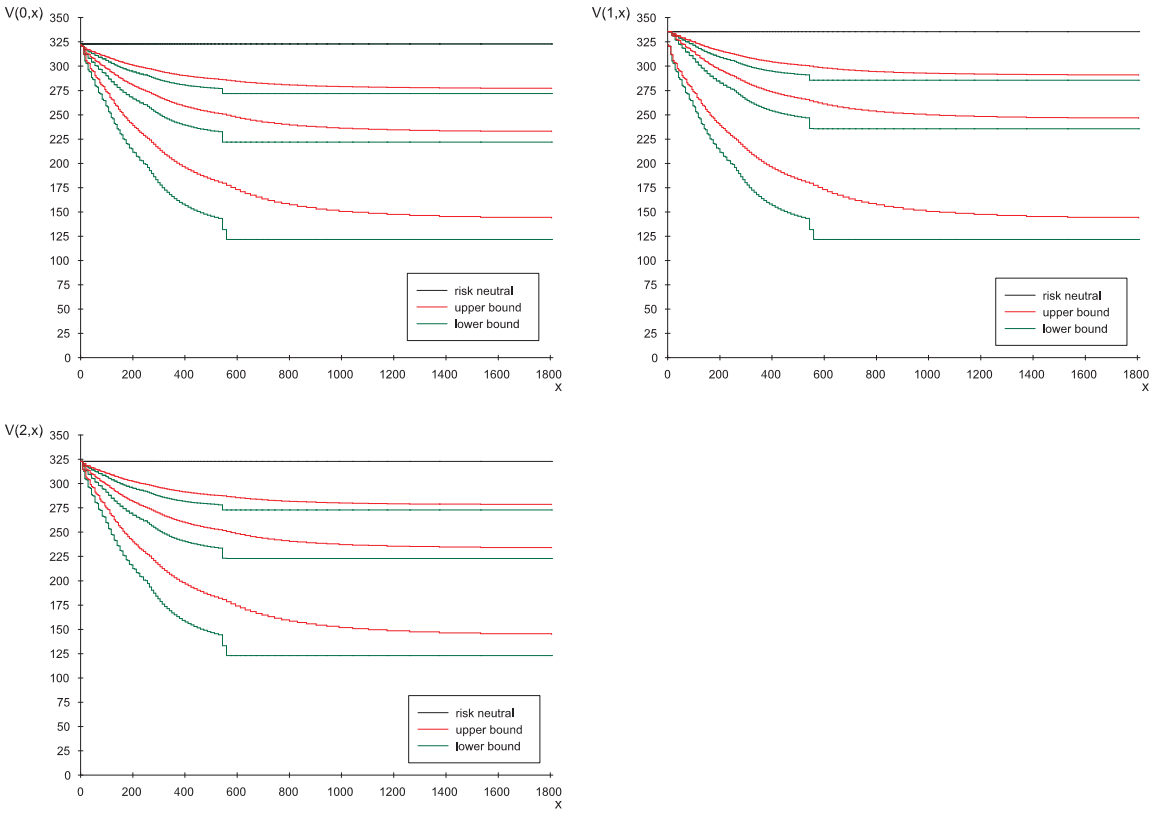

Figure 4.1: Example 6, limit behavior for Δ=0.0125

Figure 4.2 shows the limiting behavior towards the target value criterion.

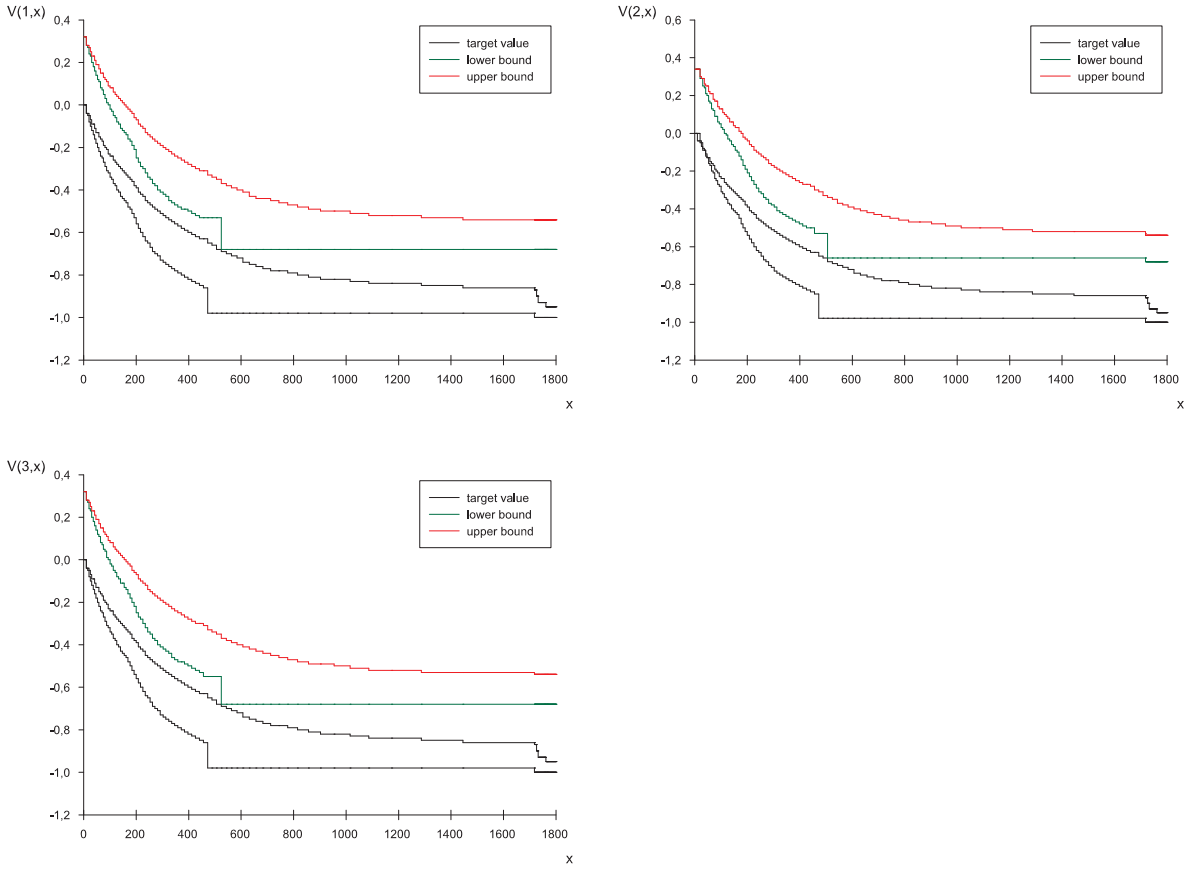

Figure 4.2: Example 6, limit behavior for Δ=0.0125

Table 4.1 shows the reduction of the computation effort that results from the utilization of the extrapolation method.

method	iteration lower bound	iteration upper bound
without extrapolation	92	92
with extrapolation	64	68

Table 4.1: Example 6, comparisons of value iteration with value iteration with extrapolation

Chapter 5

Case Study

5.1 Inventory Management

We consider a single-product inventory model. At discrete points in time $n \in N_0$, the inventory position is reviewed. In dependence on the actual stock s_n an additional amount $b_n \geq 0$ is ordered. The replenishment occurs instantly. $a_n = s_n + b_n$ determines the inventory position at time n right after the replenishment order.

The demand z_n between the times n and $n+1$ is a realization of a discrete random variable Z_n with values in $\{0, \ldots, m\}$. The random variables $Z_0, Z_1, \ldots$ are independent identical distributed with the distribution function $P(Z = z) = q(z), z \in \{0, \ldots, m\}$ with expected value μ.

In dependence on s_n, a_n and z_n the inventory position s_{n+1} at $n+1$ is $s_{n+1} = a_n - z_n$. The unsatisfied demand is backlogged. We assume an inventory with limited capacity. It follows that $s_n \leq a_n \leq M \in \mathbb{N}$. A negative stock s_n corresponds to a reservation, that is satisfied by the next order, that has to take place immediately. It follows that $a_n \geq 0$.

The placement of an order triggers cost of $c \times b_n$. Moreover there are stock and shortage costs $l(a_n - z_n)$ in dependence of the stock and shortage position $a_n - z_n$ at

the end of the order period. There is

$$l(s) = \begin{cases} l_1 \cdot s & \text{for } s \geq 0 \\ -l_2 \cdot s & \text{for } s < 0 \end{cases}$$

with $l_2 > l_1 \geq 0$. We formulate a model under the following set of assumptions.

(i) The decision to order additional stock is made at the beginning of each month and delivery occurs instantaneously.

(ii) Demand for the product arrives throughout the month but all orders are filled on the last day of the month.

(iii) If demand exceeds inventory, the demand is backlogged.

(iv) The revenues, costs, and the demand distribution do not vary from month to month.

(v) The product is sold only in whole items.

(vi) The warehouse has capacity of M units.

The MDP formulation is stated below.

(i) $S = \{-m, \ldots, -1, 0, 1, \ldots, M\}$. State s_n denotes the inventory at time n before the placement of an order.

(ii) $A = \{0, 1, \ldots, M\}$ and $D(s) = \{\max\{0, s\}, \ldots, M\}$ for $s \in S$. Action a_n with $a_n \geq \max\{0, s_n\}$ denotes the inventory at time n after the placement of an order.

(iii) $p(s, a, s') = q(a - j)$ for $a - j \in \{0, \ldots, m\}$ and 0 other times.

(iv) the one-stage rewards

$$r(s,a) = -c(a-s) - \alpha \sum_{z=0}^{m} q(z) l(a-z).$$

The negative sign results from the presentation of costs as negative rewards.

The optimality equation reads

$$V(s) = \max_{a \in D(s)} \left\{ -c(a-s) - \alpha \sum_{z=0}^{m} q(z) l(a-z) + \alpha \sum_{z=0} q(z) V(a-z) \right\}$$

The policy minimizing the probability of failing a target value is not unique. From a user's point of view, it is desirable that the policy is as smooth as possible, since frequent changes of the inventory stock level are not desirable. Therefore, we suggest a smoothing technique within the optimization algorithm. Since there is a large equivalence class of policies, we can choose a suitable structure. That means if several actions are equivalent, we choose the actions that has the smallest difference to the discretization point in the neighborhood.

5.1.1 Numerical Example

We use the following problem data. The stock capacity is $M = 4$, order cost are $c = 3$, $l_1 = 1$, $l_2 = 10$, $\beta = 0.96$, $\alpha = 1$. The probability distribution of the demand is contained in the following table.

z	0	1	2	3	4
$P(Z = z)$	0.1	0.1	0.4	0.2	0.2

Table 5.1: Probability distribution inventory management

The following figure shows the resulting policies. It is obvious that S^* is not smooth. However, it is desirable to get a smooth upper bound like in the risk neutral case.

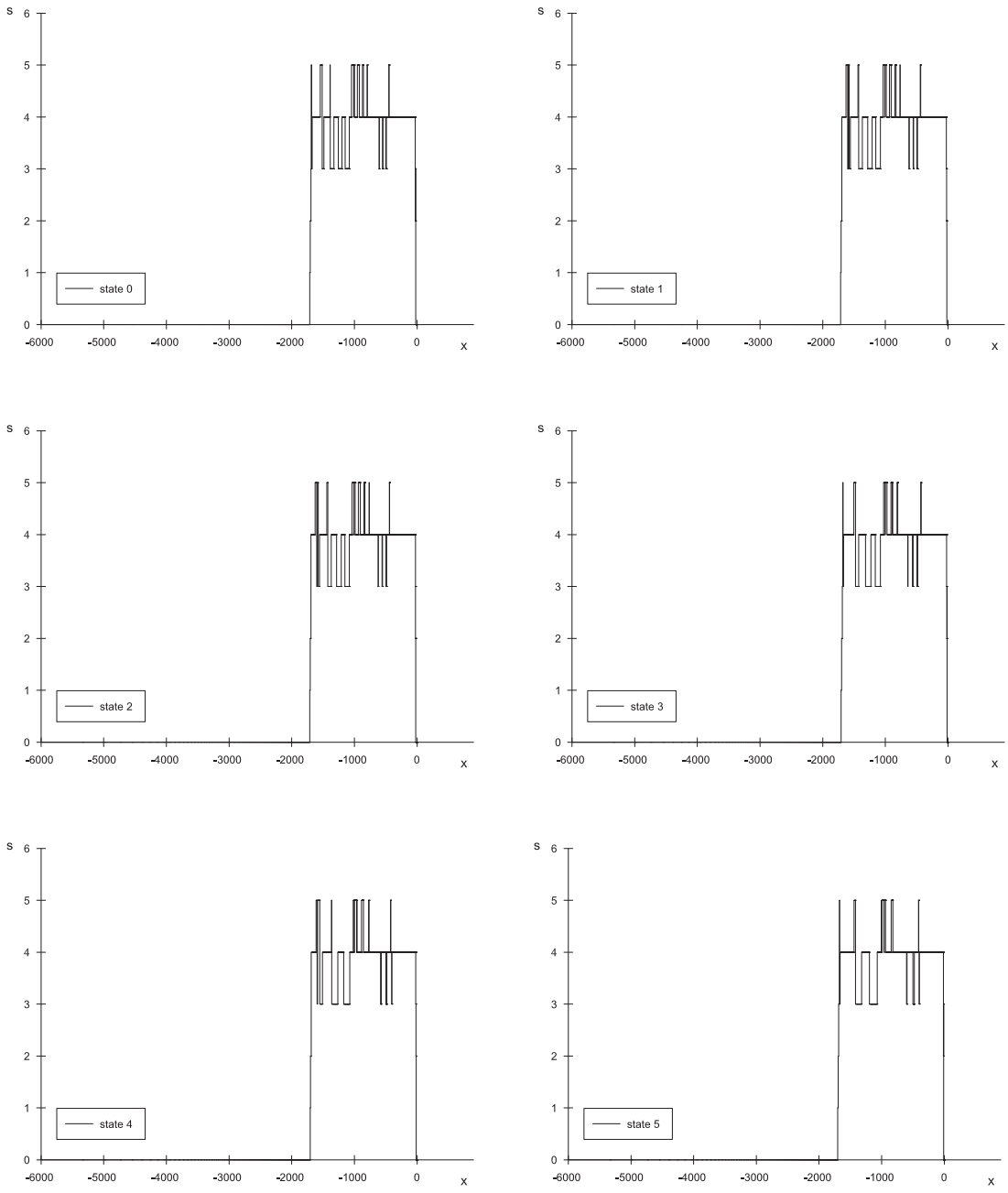

Figure 5.1: Case Study, policies according to initial inventory level

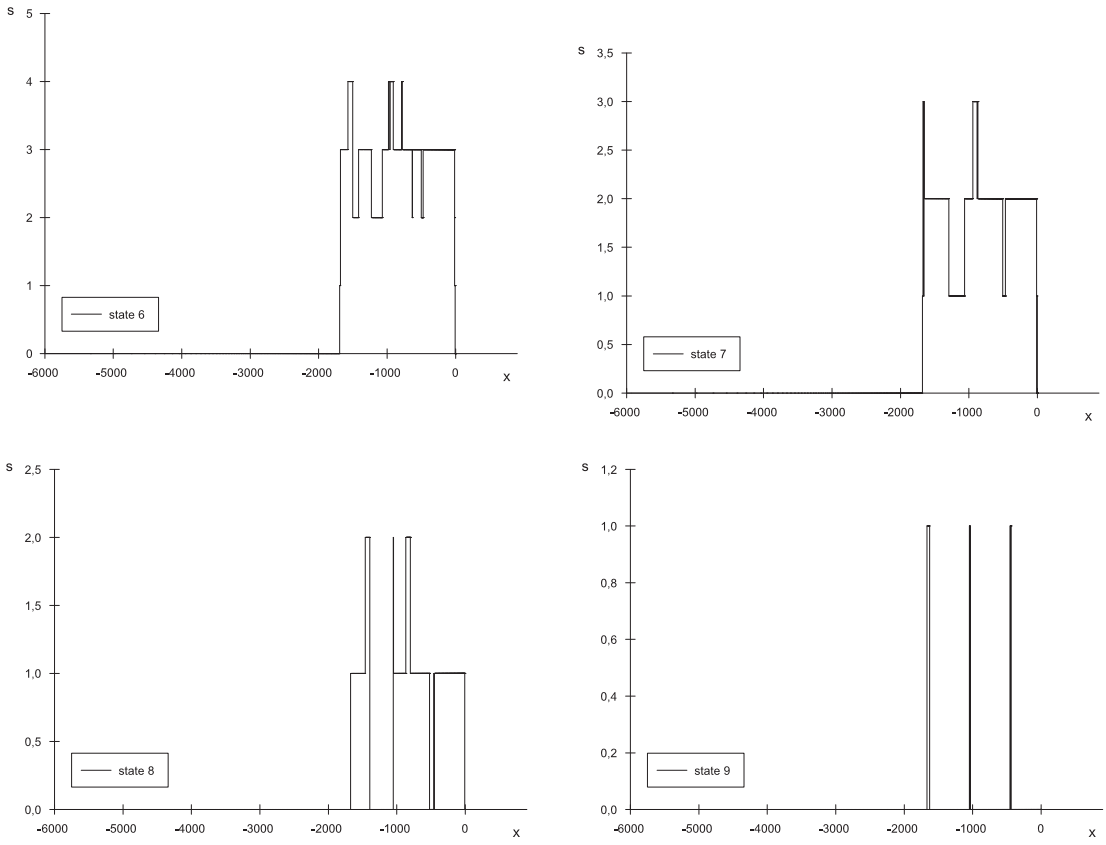

Figure 5.2: Case Study, policies according to initial inventory level continued

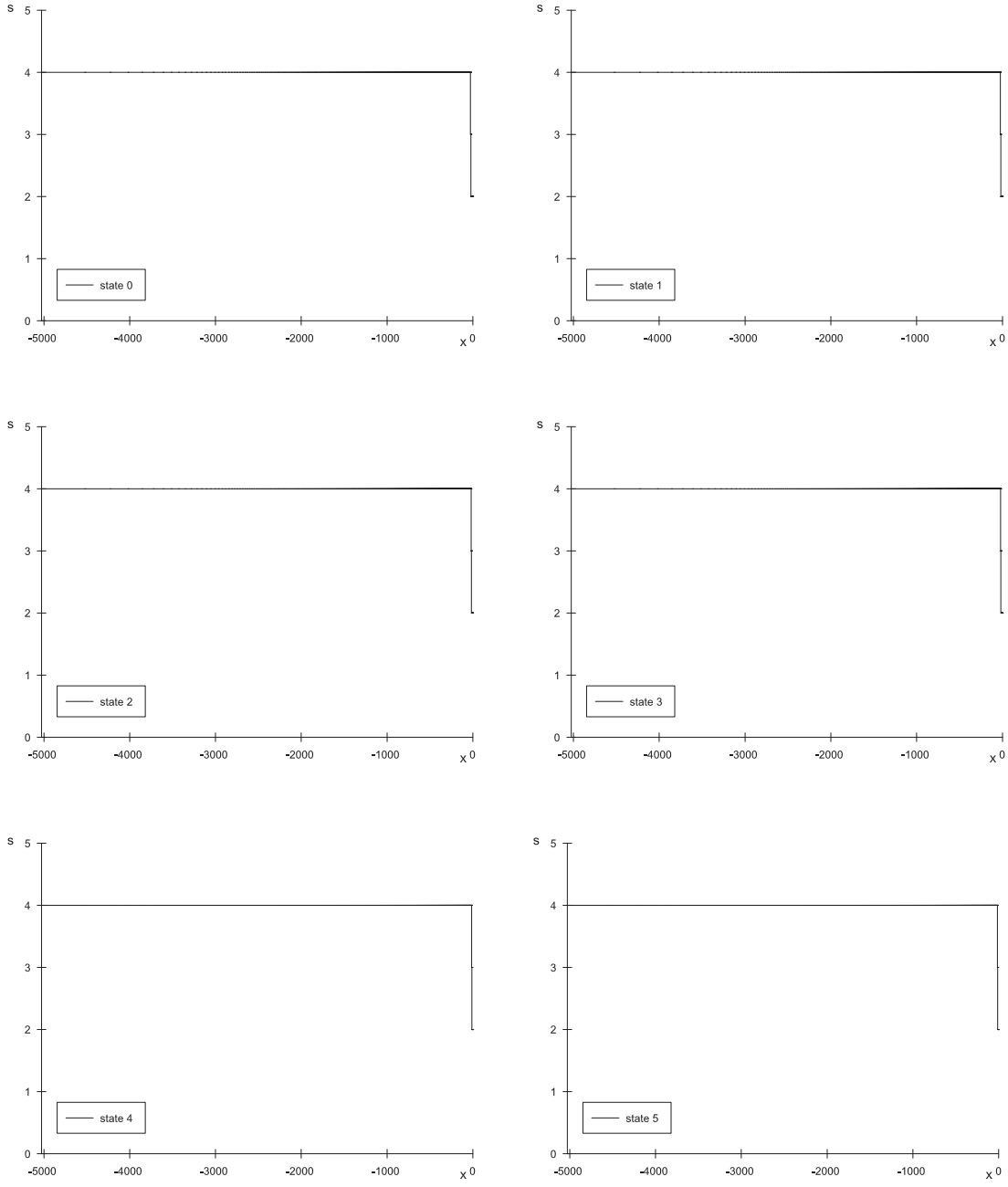

Figure 5.3: Case Study, smoothed policies according to initial inventory level

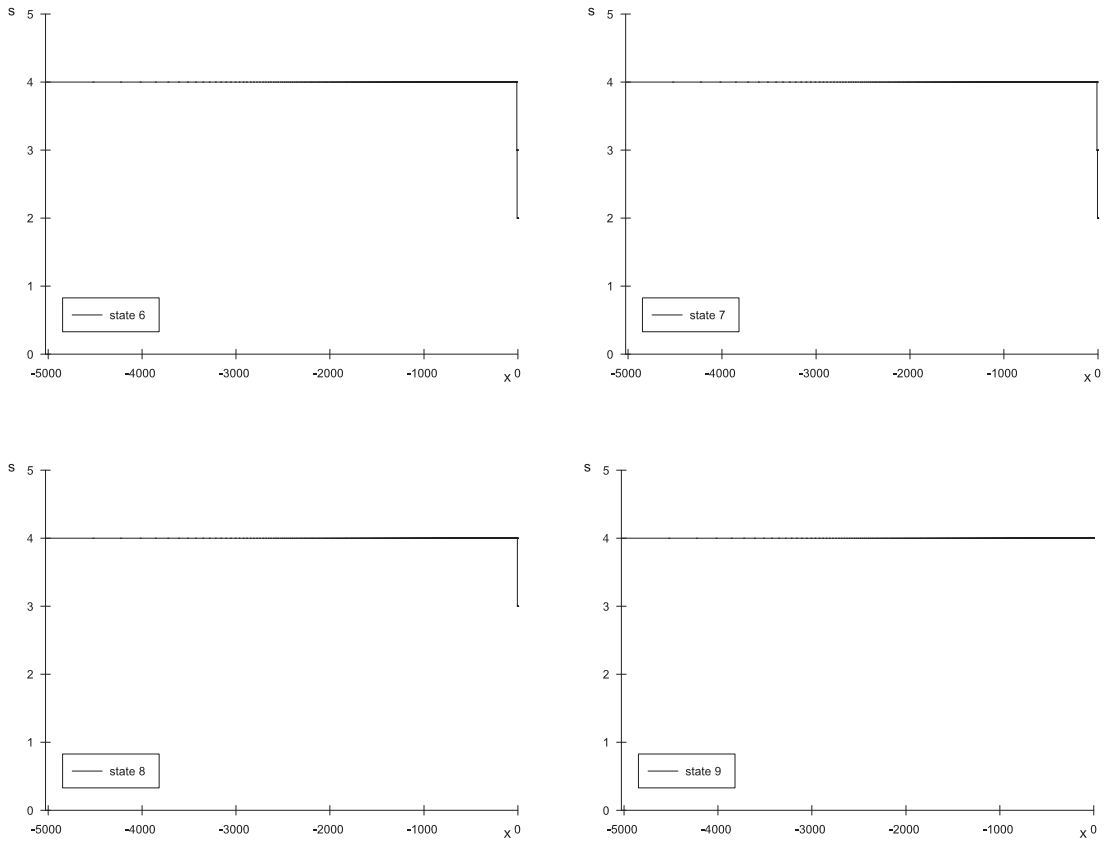

Figure 5.4: Case Study, smoothed policies according to initial inventory level continued

Chapter 6

Conclusion

6.1 Summary

We provided an overview about risk-sensitivity in Markov Decision Processes and showed that the research focused on using utility functions or mean-variance trade offs to express risk-sensitivity. We introduced Markov Decision Processes with a geometrically distributed planning horizon. A motivating example showed that the utilization of a risk-neutral policy leads to a high probability that the calculated expected total reward will not be achieved. This could be problematic for a risk-sensitive decision-maker.

Applying the Target Value Criterion means to minimize the probability that the total reward is below a predetermined target. Target that are set below the expected value can be interpreted as risk averse behavior and targets set above the expected value can be interpreted as risk loving behavior. We derived an optimality equation and proved the existence of an optimal stationary policy in a generalized state space, where the target space incorporates the realized one-stage rewards. The structure of the value function, i.e. monotonicity and asymptotic behavior was exploited to approximate the target space by a finite subset. Based on these structural results, upper and lower bounds were derived for the value function as well as nearly optimal policies. Since the value function is left continuous, an error integral is introduced to study the area between the value function of the discretized and the original

problem. The discretization allows a decomposition of the problem, which was used to recursively determine its solution.

As an extension we combined the total reward criterion and the Target Value Criterion in a penalty approach. The structure of the value function and the optimality of a stationary policy was proven. Moreover, the dependence on the optimal stationary policy on the penalty factor was examined.

The thesis closes with a case study regarding an exemplary application.

6.2 Future Research

Structure of the optimal decision rule

One of the main success stories in the application of Markov Decision Process is the utilization of the structure of optimal decision rules. Famous examples are (s, S)-policies in inventory management or protection level structures in capacitated revenue management problems. It could be investigated how the utilization of the structure of optimal decision rule can be used to reduce the computational effort.

Connection to utility functions

Another possible topic is the connection between probability distributions and utility functions. Utility functions can be normalized between zero and one, so that they have the same mathematical properties as a distribution function. Reversing the roles of the transition probabilities and the utility function provides a kind of dual problem. In the dual world, an aspiration-equivalent can be calculated by replacing the utility function with an equivalent step utility function, which has the same expected utility as the original utility function.

Bibliography

Alagoz, O. et al. (2010). Markov decision processes: a tool for sequential decision making under uncertainty. *Medical Decision Making 30*(4), 474–483.

Altman, E. (2001). *Applications of Markov Decision Processes in Communication Networks: a Survey*. Markov Decision Processes, Models, Methods, Directions, and Open Problems. Kluwer.

Barz, C. and K.-H. Waldmann (2007). Risk-sensitive capacity control in revenue management. *Mathematical Methods of Operations Research 65*, 365–579.

Bouakiz, M. and M. J. Sobel (1992). Inventory control with an exponential utility criterion. *Operations Research 40*(3), pp. 603–608.

Bouakiz, M. Kebir, Y. (1995, July). Target-level criterion in markov decision processes. *Journal of Optimization Theory and Applications 86*(1), 1–15.

Charnes, A. and W. Cooper (1959, Oct.). Chance-constrained programming. *Management Science 6*(1), 73–79.

d'Epenoux, F. (1960). Sur un probleme de production et de stockage dans laléatoire. *Revue Francaise Recherche Opratione lle 14*, 3–16.

Feinberg, E. A. (2002). *Handbook of Markov decision processes : methods and applications*. International series in operations research & management science;40. Boston: Springer.

Grävenstein, J. H. (2008). *Die Optimalitt strukturierter Entscheidungsfunktionen bei der Steuerung eines Reservoirs*. Göttingen: Sierke.

Hinderer, K. (1970). *Foundations of non-stationary dynamic programming with discrete time parameter*. Berlin: Springer.

Hinderer, K. and K.-H. Waldmann (2003). The critical discount factor for finite marmarkov decision processes with an absorbing set. *Mathematical Methods of Operations Research 57*, 1–19.

Hinderer, K. and K.-H. Waldmann (2005). Algorithms for countable state markov decision models with an absorbing set. *SIAM J. Control Optim. 43*(6), 2109–2131.

Howard, R. A. and J. E. Matheson (1972, Mar.). Risk-sensitive markov decision processes. *Management Science 18*(7), 356–369.

Koenig, M. and J. Meissner (2009). Risk minimizing strategies for revenue management problems with target values. Working paper, Lancaster University Management School.

Law, A. M. and W. D. Kelton (2000). *Simulation modeling and analysis* (3. ed. ed.). McGraw-Hill series in industrial engineering and management science. Boston: McGraw-Hill.

Manne, A. S. (1960). Linear programming and sequential decisions. *Management Science 6*, 259–267.

Ohtsubo, Y. (2003). Value iteration methods in risk minimizing stopping problems. *Journal of Compuational and Applied Mathematics 152*, 427–439.

Ohtsubo, Y. and K. Toyonaga (2002). Ovalues policy for minimizing tisk models in markov decision processes. *Journal of Mathematical Analysis and Applications 271*, 66–81.

Powell, W. B. (2011). *Approximate dynamic programming : solving the curses of dimensionality* (2. ed. ed.). Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Puterman, M. L. (2005). *Markov Decision Processes : Discrete Stochastic Dynamic Programming*. Wiley series in probability and statistics Eiley-interscience paperback series. Hoboken, NJ: Wiley-Interscience.

Puterman, M. L. and S. L. Brumelle (1979, Feb.). On the convergence of policy iteration in stationary dynamic programming. *Mathematics of Operations Research 4*(1), 60–69.

Ross, S. M. (1995). *Introduction to stochastic dynamic programming* (6. [print.] ed.). Probability and mathematical statistics. San Diego: Acad. Pr.

Ross, S. M. (2013). *Simulation*. Amsterdam: Academic Press.

Schaefer, A. J. et al. (2005). Modeling medical treatment using markov decision processes. *Operations Research and Health Care*, 593–612.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99–118.

Simon, H. A. (1957). *Models of man; social and rational.* Wiley.

Sladkỳ, K. and M. Sitavar (2004). Optimal solutions for undiscounted variance penalized markov decision chains. In *Dynamic Stochastic Optimization*, pp. 43–66. Springer.

Van Dijk, N. M., K. Sladkỳ, et al. (2006). On the total reward variance for continuous-time markov reward chains. *Journal of applied probability 43*(4), 1044–1052.

Waldmann, K.-H. (2006). On markov decision models with an absorbing set. *Decision Theory and Multi-Agent Planning*, 145–163.

Waldmann, K.-H. and U. M. Stocker (2013). *Stochastische Modelle : eine anwendungsorientierte Einfhrung*. Berlin: Springer.

White, D. (1988a, Sep.-Oct.). Further real applications of markov decision processes. *Interfaces 18*(5), 55–61.

White, D. (1988b). Mean, variance, and probabilistic criteria in finite markov decision processes: a review. *Journal of Optimization Theory and Applications 56*(1), 1–29.

White, D. (1993a). *Markov Decision Processes*. Chichester: Wiley.

White, D. (1993b). Minimising a treshold probability in discounted markov decision processes. *Journal of Mathematical Analysis and Applications 173*, 636–646.

White, D. J. (1985, Nov.-Dec.). Real applications of markov decision processes. *Interfaces 15*(6), 73–83.

Wu, C. and Y. Lin (1999). Minimizing risk models in markov decision processes with policies depending on target values. *Journal of Mathematical Analysis and Application*, 47–67.