MORE RISK-SENSITIVE MARKOV DECISION PROCESSES

NICOLE BÄUERLE* AND ULRICH RIEDER‡

ABSTRACT. We investigate the problem of minimizing a certainty equivalent of the total or discounted cost over a finite and an infinite horizon which is generated by a Markov Decision Process (MDP). The certainty equivalent is defined by $U^{-1}(\mathbb{E}\,U(Y))$ where U is an increasing function. In contrast to a risk-neutral decision maker this optimization criterion takes the variability of the cost into account. It contains as a special case the classical risk-sensitive optimization criterion with an exponential utility. We show that this optimization problem can be solved by an ordinary MDP with extended state space and give conditions under which an optimal policy exists. In the case of an infinite time horizon we show that the minimal discounted cost can be obtained by value iteration and can be characterized as the unique solution of a fixed point equation using a 'sandwich' argument. Interestingly, it turns out that in case of a power utility, the problem simplifies and is of similar complexity than the exponential utility case, however has not been treated in the literature so far. We also establish the validity (and convergence) of the policy improvement method. A simple numerical example, namely the classical repeated casino game is considered to illustrate the influence of the certainty equivalent and its parameters. Finally also the average cost problem is investigated. Surprisingly it turns out that under suitable recurrence conditions on the MDP for convex power utility U, the minimal average cost does not depend on U and is equal to the risk neutral average cost. This is in contrast to the classical risk sensitive criterion with exponential utility.

KEY WORDS: Markov Decision Problem, Certainty Equivalent, Positive Homogeneous Utility, Exponential Utility, Value Iteration, Policy Improvement, Risk-sensitive Average Cost.

AMS SUBJECT CLASSIFICATIONS: 90C40, 91B06.

1. INTRODUCTION

Since the seminal paper by Howard & Matheson (1972) the notion *risk-sensitive* Markov Decision Process (MDP) seems to be reserved for the criterion $\frac{1}{\gamma}\log\mathbb{E}[e^{\gamma Y}]$ where Y is some cumulated cost and γ represents the degree of risk aversion or risk attraction. However in the recent decade a lot of alternative ways of measuring performance with a certain emphasis on risk arose. Among them *risk measures* and the well-known *certainty equivalents*. Certainty equivalents have a long tradition and its use can be traced back to the 1930ies (for a historic review see Muliere & Parmigiani (1993)). They are defined by $U^{-1}(\mathbb{E}\,U(Y))$ where U is an increasing function. We consider here a discrete-time MDP evolving on a Borel state space which accumulates cost over a finite or an infinite time horizon. The one-stage cost are bounded. The aim is to minimize the certainty equivalent of this accumulated cost. In case of an infinite time horizon, cost have to be discounted. We will also consider an average cost criterion.

The problems we treat here are generalizations of the classical 'risk-sensitive' case which is obtained when we set $U(y) = \frac{1}{\gamma}e^{\gamma y}$. On the other hand they are more specialized than the problems with expected utility considered for example in Kreps (1977a,b). There the author considers a countable state, finite action MDP where the utility which has to be maximized may depend on the complete history of the process. In such a setting it is already hard to obtain some kind of stationarity. It is achieved by introducing what the author calls a *summary space* - an idea which is natural and which will also appear in our analysis. A forward recursive utility and its summary (which is there called forward recursive accumulation) are investigated for a finite horizon problem in Iwamoto (2004).

Somehow related studies can be found in Jaquette (1973, 1976) where finite state, finite action MDPs are considered and moment optimality and the exponential utility of the infinite horizon discounted reward is investigated. General optimization criteria can also be found in Chung & Sobel (1987). There the authors first consider fixed point theorems for the complete distribution of the infinite horizon discounted reward in a finite MDP and later also consider the exponential utility. In Collins & McNamara (1998) the authors deal with a finite horizon problem and maximize a strictly concave functional of the distribution of the terminal state. Another non-standard optimality criterion is the *target level criterion* where the aim is to maximize the probability that the total discounted reward exceeds a given target value. This is e.g. investigated in Wu & Lin (1999); Boda et al. (2004). Other probabilistic criteria, mostly in combination with long-run performance measures, can be found in the survey of White (1988).

Only recently some papers appeared where risk measures have been used for optimization of MDPs. In Ruszczyński (2010) general space MDPs are considered in a finite horizon model as well as in a discounted infinite horizon model. A dynamic Markov risk measure is used as optimality criterion. In both cases value iteration procedures are established and for the infinite horizon model the validity and convergence of the policy improvement method is shown. The concrete risk measure Average-Value-at-Risk has been used in Bäuerle & Ott (2011) for minimizing the discounted cost over a finite and an infinite horizon for a general state MDP. Value iteration methods have been established and the optimality of Markov policies depending on a certain 'summary' has been shown. Some numerical examples have also been given, illustrating the influence of the 'risk aversion parameter'. In Bäuerle & Mundt (2009) a Mean-Average-Value-at-Risk problem has been solved for an investor in a binomial financial market.

Classical risk-sensitive MDPs have been intensively studied since Howard & Matheson (1972). In particular the average cost criterion has attracted a lot of researchers since it behaves considerable different to the classical risk neutral average cost problem (see e.g. Cavazos-Cadena & Hernández-Hernández (2011); Cavazos-Cadena & Fernández-Gaucherand (2000); Jaśkiewicz (2007); Di Masi & Stettner (1999)). The infinite horizon discounted classical risk-sensitive MDP and its relation to the average cost problem is considered in Di Masi & Stettner (1999). As far as applications are concerned, risk-sensitive problems can e.g. be found in Bielecki et al. (1999) where portfolio management is considered, in Denardo et al. (2011, 2007) where multi-armed bandits are investigated and in Barz & Waldmann (2007) where revenue problems are treated.

In this paper we investigate the problem of minimizing the certainty equivalent of the total and discounted cost over a finite and an infinite horizon which is generated by a Markov Decision Process. We consider both the risk averse and the risk seeking case. We show that these problems can be solved by ordinary MDPs with extended state space and give continuity and compactness conditions under which optimal policies exist. In the case of discounting we have to enlarge the state space by another component since the discount factor implies some kind of non-stationarity. The enlargement of the state space is partly dispensable for exponential or power utility functions. Interestingly, the problem with power utility shows a similar complexity than the classical exponential case, but to the best of our knowledge has not been considered in the MDP literature so far. In the case of an infinite horizon we show that the minimal value can be obtained by value iteration and can be characterized as the unique solution of a fixed point equation using a 'sandwich' argument. We also establish the validity (and convergence) of the policy improvement method. A simple numerical example, namely a classical repeated casino game, is considered to illustrate the influence of the function U and its parameters. Finally also the average cost problem is investigated. Surprisingly it turns out that under suitable recurrence conditions on the MDP for $U(y) = \frac{1}{\gamma}y^\gamma$ with $\gamma \geq 1$, the minimal average cost does not depend on γ and are equal to the risk neutral average cost. This is in contrast to the classical risk sensitive criterion. The average cost case with $\gamma < 1$ remains an open problem.

The paper is organized as follows: In Section 2 we introduce the MDP model, our continuity and compactness assumptions and the admissible policies. In Section 3 we solve the finite horizon problem with certainty equivalent criterion. We consider the total cost problem as well as the

discounted cost problem. In the latter case we have to further extend the state space of the MDP. One subsection deals with the repeated casino game which is solved explicitly. Next, in Section 4 we consider and solve the infinite horizon problem. We show that the minimal discounted cost can be obtained by value iteration and can be characterized as the unique solution of a fixed point equation. We also establish the validity (and convergence) of the policy improvement method. Finally in Section 5 we investigate the average cost problem for power utility.

2. General Risk-Sensitive Markov Decision Processes

We suppose that a controlled Markov state process (X_n) in discrete time is given with values in a Borel set E. More precisely it is specified by:

- The Borel *state space* E, endowed with a Borel σ-algebra $\mathcal{E}$.
- The Borel *action space* A, endowed with a Borel σ-algebra $\mathcal{A}$.
- The set $D \subset E \times A$, a nonempty Borel set and subsets $D(x) := \{a \in A : (x,a) \in D\}$ of *admissible actions in state* x.
- A regular conditional distribution Q from D to E, the *transition law.*
- A measurable *cost function* $c : D \to [\underline{c}, \bar{c}]$ with $0 < \underline{c} < \bar{c}$.

Note that we assume here for simplicity that the cost are positive and bounded. Next we introduce the sets of histories for $k \in \mathbb{N}$ by:

$$H_0 := E, \quad H_n := D^n \times E$$

where $h_n = (x_0, a_0, x_1, \ldots, a_{n-1}, x_n) \in H_n$ gives a history up to time n. A *history-dependent policy* $\sigma = (g_n)_{n\in\mathbb{N}_0}$ is given by a sequence of measurable mappings $g_n : H_n \to A$ such that $g_n(h_n) \in D(x_n)$. We denote the set of all such policies by Π. Each policy $\sigma \in \Pi$ induces together with the initial state x a probability measure $\mathbb{P}_x^\sigma$ and a stochastic process (X_n, A_n) on H_∞ such that X_n is the random state at time n and A_n is the action at time n. (For details see e.g. Bäuerle & Rieder (2011), Section 2).

There is a discount factor $\beta \in (0,1]$ and we will either consider a finite planning horizon $N \in \mathbb{N}_0$ or an infinite planning horizon. Thus we will either consider the cost

$$C_\beta^N := \sum_{k=0}^{N-1} \beta^k c(X_k, A_k) \quad \text{or} \quad C_\beta^\infty := \sum_{k=0}^{\infty} \beta^k c(X_k, A_k).$$

If $\beta = 1$ we will shortly write C^N instead of C_1^N. In the last section we will also consider average cost problems. Instead of minimizing the expected cost we will now treat a general non-standard risk-sensitive criterion. To this end let U be a continuous and strictly increasing function such that the inverse U^{-1} exists. The aim now is to solve:

$$\inf_{\sigma\in\Pi} U^{-1}\Big(\mathbb{E}_x^\sigma\left[U(C_\beta^N)\right]\Big), \quad x \in E, \tag{2.1}$$

$$\inf_{\sigma\in\Pi} U^{-1}\Big(\mathbb{E}_x^\sigma\left[U(C_\beta^\infty)\right]\Big), \quad x \in E \tag{2.2}$$

where $\mathbb{E}_x^\sigma$ is the expectation w.r.t. $\mathbb{P}_x^\sigma$. Note that the problems in (2.1) and (2.2) are well-defined. If U is in addition strictly convex, then the quantity $U^{-1}\Big(\mathbb{E}\left[U(Y)\right]\Big)$ can be interpreted as the *mean-value premium* of the risk Y as is done in actuarial sciences (see e.g. Kaas et al. (2009)). If U is strictly concave, then U is a utility function and the quantity represents a *certainty equivalent* also known as a *quasi-linear mean.* It may be written (assuming enough regularity of U) using the Taylor rule as

$$U^{-1}\Big(\mathbb{E}\left[U(Y)\right]\Big) \approx \mathbb{E}\,Y - \frac{1}{2} l_U(\mathbb{E}\,Y) Var[Y]$$

where

$$l_U(y) = -\frac{U''(y)}{U'(y)}$$

is the *Arrow-Pratt* function of absolute risk aversion. Hence the second term accounts for the variability of X (for a discussion see Bielecki & Pliska (2003)). If U is concave, the variance is subtracted and hence the decision maker is risk seeking in case cost are minimized, if U is convex, then the variance is added and the decision maker is risk averse. A prominent special case is the choice

$$U(y) = \frac{1}{\gamma} e^{\gamma y}, \quad \gamma \neq 0$$

in which case $l_U(y) = -\gamma$. When we speak of minimizing cost, the case $\gamma > 0$ corresponds to a risk averse decision maker and the case $\gamma < 0$ to a risk-seeking decision maker. Note that this interpretation changes when we maximize reward. The limiting case $\gamma \to 0$ coincides with the classical risk-neutral criterion.

Other interesting choices are $U(y) = \frac{1}{\gamma} y^\gamma$ with $\gamma > 0$. For $\gamma < 1$ the function U is strictly concave and $l_U(y) = \frac{1-\gamma}{y}$. This is the risk-seeking case for the cost problem. If $\gamma \geq 1$ we can also write

$$U^{-1}\Big(\mathbb{E}[U(Y)]\Big) = \Big(\mathbb{E}\, Y^\gamma\Big)^{\frac{1}{\gamma}} = \|Y\|_\gamma$$

where $\|\cdot\|_\gamma$ is the usual L^γ-norm. Of course $\gamma = 1$ is again the risk neutral case.

In this paper we impose the following continuity and compactness assumptions (CC) on the data of the problem:

(i) $U : [0, \infty) \to \mathbb{R}$ is continuous and strictly increasing,
(ii) $D(x)$ is compact for all $x \in E$,
(iii) $x \mapsto D(x)$ is upper semicontinuous, i.e. for all $x \in E$ it holds: If $x_n \to x$ and $a_n \in D(x_n)$ for all $n \in \mathbb{N}$, then (a_n) has an accumulation point in $D(x)$,
(iv) $(x, a) \mapsto c(x, a)$ is lower semicontinuous,
(v) Q is weakly continuous, i.e. for a all $v : E \to \mathbb{R}$ bounded and continuous

$$(x, a) \mapsto \int v(x')Q(dx'|x, a)$$

is again continuous.

Note that assumptions (CC) will later imply the existence of optimal policies and the validity of the value iteration. It is also possible to show these statements under other assumptions, in particular under so-called structure assumptions. For a discussion see e.g. Bäuerle & Rieder (2011), Section 2.4.

3. Finite Horizon Problems

3.1. **Total Cost Problems.** We start investigating the case of a finite time horizon N and $\beta = 1$. Since U is strictly increasing, so is U^{-1} and we can obviously skip it from the optimization problem. In what follows we denote by

$$J_N(x) := \inf_{\sigma \in \Pi} \mathbb{E}_x^\sigma \Big[U\Big(\sum_{k=0}^{N-1} c(X_k, A_k)\Big)\Big] = \inf_{\sigma \in \Pi} \mathbb{E}_x^\sigma [U(C^N)], \quad x \in E. \tag{3.1}$$

Though this problem is not directly separable, we will show that it can be solved by a bivariate MDP as follows. For this purpose let us define for $n = 0, 1, \ldots, N$

$$\begin{aligned} V_{n\sigma}(x, y) &:= \mathbb{E}_x^\sigma \big[U\big(C^n + y\big)\big], \quad x \in E, y \in \mathbb{R}_+, \sigma \in \Pi, \\ V_n(x, y) &:= \inf_{\sigma \in \Pi} V_{n\sigma}(x, y), \quad x \in E, y \in \mathbb{R}_+. \end{aligned} \tag{3.2}$$

Obviously $V_N(x, 0) = J_N(x)$. The idea is that y summarizes the cost which has been accumulated so far. This idea can already be found in Kreps (1977a,b). We consider now a Markov Decision Model which is defined on the state space $\tilde{E} := E \times \mathbb{R}_+$ with action space A and admissible

actions given by the set D. The one-stage cost are zero and the terminal cost function is $V_0(x,y) := U(y)$. The transition law is given by $\tilde{Q}(\cdot|x,y,a)$ defined by

$$\int v(x',y')\tilde{Q}(d(x',y')|x,y,a) = \int v(x',c(x,a)+y)Q(dx'|x,a).$$

Decision rules are here given by measurable mappings $f : \tilde{E} \to A$ such that $f(x,y) \in D(x)$. We denote by F the set of decision rules and by Π^M the set of Markov policies $\pi = (f_0, f_1, \ldots)$ with $f_n \in F$. Note that 'Markov' refers to the fact that the decision at time n depends only on x and y. Obviously in (3.2) only the first n decision rules of σ are relevant. Note that we have $\Pi^M \subset \Pi$ in the following sense: For every $\pi = (f_0, f_1, \ldots) \in \Pi^M$ we find a $\sigma = (g_0, g_1, \ldots) \in \Pi$ such that

$$\begin{aligned} g_0(x_0) &:= f_0(x_0, 0), \\ g_n(x_0, a_0, x_1, \ldots, x_n) &:= f_n\big(x_n, \sum_{k=0}^{n-1} c(x_k, a_k)\big),\ n \in \mathbb{N}. \end{aligned}$$

With this interpretation $V_{n\pi}$ is also defined for $\pi \in \Pi^M$. For convenience we introduce the set

$$\begin{aligned} \mathcal{C}(\tilde{E}) := \Big\{ & v : \tilde{E} \to \mathbb{R} : v \text{ is lower semicontinuous}, v(x,\cdot) \text{is continuous} \\ & \text{and increasing for } x \in E \text{ and } v(x,y) \geq U(y) \Big\}. \end{aligned}$$

Note that $v \in \mathcal{C}(\tilde{E})$ is bounded from below. For $v \in \mathcal{C}(\tilde{E})$ and $f \in F$ we denote the operator

$$(T_f v)(x,y) := \int v\big(x', c(x, f(x,y)) + y\big) Q\big(dx'|x, f(x,y)\big), \quad (x,y) \in \tilde{E}.$$

The minimal cost operator of this Markov Decision Model is given by

$$(Tv)(x,y) = \inf_{a \in D(x)} \int v\big(x', c(x,a) + y\big) Q\big(dx'|x,a\big), \quad (x,y) \in \tilde{E}. \tag{3.3}$$

If a decision rule $f \in F$ is such that $T_f v = Tv$, then f is called a *minimizer* of v. In what follows we will always assume that the empty sum is zero. Then we obtain:

Theorem 3.1. *It holds that*

a) *For a policy $\pi = (f_0, f_1, f_2, \ldots) \in \Pi^M$ we have the following cost iteration: $V_{n\pi} = T_{f_0} \ldots T_{f_{n-1}} U$ for $n = 1, \ldots, N$.*
b) *$V_0(x,y) := U(y)$ and $V_n = TV_{n-1}$, for $n = 1, \ldots, N$ i.e.*

$$V_n(x,y) = \inf_{a \in D(x)} \int V_{n-1}\big(x', c(x,a) + y\big) Q\big(dx'|x,a\big).$$

Moreover, $V_n \in \mathcal{C}(\tilde{E})$.
c) *For every $n = 1, \ldots, N$ there exists a minimizer $f_n^* \in F$ of V_{n-1} and $(g_0^*, \ldots, g_{N-1}^*)$ with*

$$g_n^*(h_n) := f_{N-n}^*\big(x_n, \sum_{k=0}^{n-1} c(x_k, a_k)\big), \quad n = 0, \ldots, N-1$$

is an optimal policy for problem (3.1). Note that the optimal policy consists of decision rules which depend on the current state and the accumulated cost so far.

Proof. We will first prove part a) by induction. By definition $V_{0\pi}(x,y) = U(y)$ and

$$V_{1\pi}(x,y) = U\big(c(x, f_0(x,y)) + y\big) = (T_{f_0} U)(x,y).$$

Now suppose the statement holds for $V_{n-1\pi}$ and consider $V_{n\pi}$. In order to ease notation we denote for a policy $\pi = (f_0, f_1, f_2, \ldots) \in \Pi^M$ by $\vec{\pi} = (f_1, f_2, \ldots)$ the shifted policy. Hence

$$\begin{aligned}(T_{f_0} \ldots T_{f_{n-1}} U)(x,y) &= \int V_{n-1\vec{\pi}}\big(x', c(x, f_0(x,y)) + y\big) Q(dx'|x, f_0(x,y)) \\ &= \int \mathbb{E}^{\vec{\pi}}_{x'} \Big[U\big(\sum_{k=0}^{n-2} c(X_k, A_k) + c(x, f_0(x,a)) + y \big) \Big] Q(dx'|x, f_0(x,a)) \\ &= V_{n\pi}(x,y).\end{aligned}$$

Next we prove part b) and c) together. From part a) it follows that for $\pi \in \Pi^M$, the value functions in problem (3.2) indeed coincide with the value functions of the previously defined MDP. From MDP theory it follows in particular that it is enough to consider Markov policies Π^M, i.e. $V_n = \inf_{\sigma \in \Pi} V_{n\sigma} = \inf_{\pi \in \Pi^M} V_{n\pi}$ (see e.g. Hinderer (1970) Theorem 18.4). Next consider functions $v \in \mathcal{C}(\tilde{E})$. We show that $Tv \in \mathcal{C}(\tilde{E})$ and that there exists a minimizer for v. Statements b) and c) then follow from Theorem 2.3.8 in Bäuerle & Rieder (2011).

Now suppose $v \in \mathcal{C}(\tilde{E})$. Taking into account our standing assumptions (CC) (i),(iv) at the end of section 2 it obviously follows that $(x, y, a, x') \mapsto v\big(x', c(x,a) + y\big)$ is lower semicontinuous. Moreover $y \mapsto v\big(x', c(x,a) + y\big)$ is increasing and continuous. We can now apply Theorem 17.11 in Hinderer (1970) to obtain that $(x, y, a) \mapsto \int v(x, y, a, x') Q(dx'|x, a)$ is lower semicontinuous. By Proposition 2.4.3 in Bäuerle & Rieder (2011) it follows that $(x, y) \mapsto (Tv)(x, y)$ is lower semicontinuous and there exists a minimizer of v.

Further it is clear that $y \mapsto \int v\big(x', c(x,a) + y\big) Q(dx'|x, a)$ is increasing and continuous (by monotone convergence), i.e. in particular upper semicontinuous. Now since the infimum of an arbitrary number of upper semicontinuous functions is upper semicontinuous, we obtain $y \mapsto (Tv)(x, y)$ is continuous and also increasing. The inequality $(Tv)(x, y) \geq U(y)$ follows directly. □

The last theorem shows that the optimal policy of (3.1) can be found in the smaller set Π^M which makes the problem computationally tractable.

In the special case $U(y) = \frac{1}{\gamma} e^{\gamma y}$ with $\gamma \neq 0$ the iteration simplifies and the second component can be skipped.

Corollary 3.2 (Exponential Utility)**.** *In case $U(y) = \frac{1}{\gamma} e^{\gamma y}$ with $\gamma \neq 0$, we obtain*

a) $V_n(x, y) = e^{\gamma y} h_n(x), \quad n = 0, \ldots, N$ *and* $J_N(x) = h_N(x)$.
b) *The functions h_n from part a) are given by $h_0 = \frac{1}{\gamma}$ and*

$$h_n(x) = \inf_{a \in D(x)} \big\{ e^{\gamma c(x,a)} \int h_{n-1}(x') Q(dx'|x, a) \big\}.$$

Proof. We prove the statements a) and b) by induction. For $n = 0$ we obtain $V_0(x, y) = \frac{1}{\gamma} e^{\gamma y} = e^{\gamma y} \cdot \frac{1}{\gamma}$, hence $h_0 \equiv \frac{1}{\gamma}$. Now suppose part a) is true for $n - 1$. From the Bellman equation (Theorem 3.1 b)) we obtain:

$$\begin{aligned}V_n(x,y) &= \inf_{a \in D(x)} \int V_{n-1}\big(x', c(x,a) + y\big) Q\big(dx'|x, a\big) \\ &= \inf_{a \in D(x)} \int e^{\gamma(y + c(x,a))} h_{n-1}(x') Q\big(dx'|x, a\big) \\ &= e^{\gamma y} \inf_{a \in D(x)} \big\{ e^{\gamma c(x,a)} \int h_{n-1}(x') Q\big(dx'|x, a\big) \big\}.\end{aligned}$$

Hence the statement follows by setting $h_n(x) = \inf_{a \in D(x)} \big\{ e^{\gamma c(x,a)} \int h_{n-1}(x') Q(dx'|x, a) \big\}$. □

Remark 3.3. Taking the logarithm in the equation of part b) we obtain the maybe more familiar form

$$\log h_n(x) = \inf_{a\in D(x)} \Big\{\gamma c(x,a) + \log \int h_{n-1}(x')Q(dx'|x,a)\Big\}$$

see e.g. Bielecki et al. (1999).

Remark 3.4. Of course instead of minimizing cost one could also consider the problem of maximizing reward. Suppose that $r : D \to [\underline{r}, \bar{r}]$ (with $0 < \underline{r} < \bar{r}$) is a one-stage reward function and the problem is

$$J_N(x) := \sup_{\sigma\in\Pi} \mathbb{E}_x^\sigma \Big[U\Big(\sum_{k=0}^{N-1} r(X_k, A_k)\Big)\Big], \quad x \in E. \tag{3.4}$$

It is possible to treat this problem in exactly the same way. The value iteration is given by $V_0(x, y) := U(y)$ and

$$V_n(x,y) = \sup_{a\in D(x)} \int V_{n-1}\big(x', r(x,a)+y\big)Q\big(dx'|x,a\big).$$

Remark 3.5. It is possible to state similar results for models where the cost does also depend on the next state, i.e. $c = c(X_k, A_k, X_{k+1})$. In particular, the value iteration reads here

$$V_n(x,y) = \inf_{a\in D(x)} \int V_{n-1}\big(x', y + c(x,a,x')\big)Q\big(dx'|x,a\big).$$

Note however, that for exponential utility we have to modify the iteration in Corollary 3.2 accordingly.

3.2. **Application: Casino Game.** In this section, we are going to illustrate the results of the previous section and the influence of the choice of the function U by means of a simple numerical example. For the given horizon $N \in \mathbb{N}$, we consider N independent identically distributed games. The probability of winning one game is given by $p \in (0, 1)$. We assume that the gambler starts with initial capital $x_0 > 0$. Further, let X_{k-1}, $k = 1, \dots, N$, be the capital of the gambler right before the k-th game. The final capital is denoted by X_N. Before each game, the gambler has to decide how much capital she wants to bet in the following game in order to maximize her risk-adjusted profit. The aim is to find

$$J_N(x_0) := \sup_{\sigma\in\Pi} \mathbb{E}_x^\sigma \big[U(X_N)\big], \quad x_0 > 0. \tag{3.5}$$

This is obviously a reward maximization problem, but can be treated by the same means (see Remark 3.4). As one-stage reward we choose $r(X_k, A_k, X_{k+1}) = X_{k+1} - X_k$. Note that here the reward depends on the outcome of the next state (see Remark 3.5). In what follows we will distinguish between two cases: In the first one we choose $U(y) = y^\gamma$ for $\gamma > 0$ and in the second one $U(y) = \frac{1}{\gamma}e^{\gamma y}$ for $\gamma \neq 0$. Let us denote by $Z_1, \dots, Z_N$ independent and identically distributed random variables which describe the outcome of the games. More precisely, $Z_k = 1$ if the k-th game is won and $Z_k = -1$ if the k-th game is lost. Let us denote by Q^Z the distribution of Z.

Case 1: Let $U(y) = y^\gamma$ with $\gamma > 0$. Since the games are independent it is not difficult to see that in this case we do not need the artificial state variable y or can identify x and y when we choose x to be the current capital (=accumulated reward). Moreover, it is reasonable to describe the action in terms of the fraction of money that the gambler bets. Hence $E := \mathbb{R}_+$ and $A = [0, 1]$ where $D(x) = A$. We obtain $X_{k+1} = X_k + X_k A_k Z_{k+1}$ and hence $r(X_k, A_k, X_{k+1}) = X_{k+1} - X_k = X_k A_k Z_{k+1}$. The value iteration is given by

$$V_n(x) = \sup_{a\in[0,1]} \int V_{n-1}\big(x + xaz\big)Q^Z(dz).$$

We have to start the iteration with $V_0(x) := x^\gamma$ and are interested in obtaining $V_N(x_0)$. It is easy to see by induction that $V_n(x) = x^\gamma d_n$ for some constants d_n and all one-stage optimization problems reduce to

$$\sup_{a\in[0,1]} \int (1+az)^\gamma Q^Z(dz) = \sup_{a\in[0,1]} \big\{p(1+a)^\gamma + (1-p)(1-a)^\gamma\big\}. \tag{3.6}$$

Hence the optimal fraction to bet does not depend on the time horizon nor on the current capital. Depending on γ the optimal policy can be discussed explicitly.

Case $\gamma = 1$: : This is the *risk neutral* case. The function in (3.6) reduces to the linear function $1 + a(2p-1)$. Obviously the optimal policy is $f_n^*(x) = 1$ if $p > \frac{1}{2}$ and $f_n^*(x) = 0$ if $p \le \frac{1}{2}$. If $p = \frac{1}{2}$ all policies are optimal.

Case $\gamma > 1$: : This is the *risk-seeking* case. The function which has to be maximized in (3.6) is convex on $[0,1]$ hence the maximum points are on the boundary of the interval. We obtain $f^*(x) = 1$ if $p > \frac{1}{2^\gamma}$ and $f^*(x) = 0$ if $p \le \frac{1}{2^\gamma}$.

Case $\gamma < 1$: : This is the *risk-averse* case. We can find the maximum point of the function in (3.6) by inspecting its derivative. We obtain (let us denote $\rho = \frac{1-p}{p}$)

$$f^*(x) = \frac{\rho^{\frac{1}{\gamma-1}} - 1}{1 + \rho^{\frac{1}{\gamma-1}}}$$

if $p > \frac{1}{2}$ and $f^*(x) = 0$ if $p \le \frac{1}{2}$.

An illustration of the optimal policy can be seen in figure 1. There, the optimal fraction of the wealth which should be bet is plotted for different parameters γ. The red line is $\gamma = \frac{1}{2}$ and belongs to the risk neutral gambler. The green line belongs to $\gamma = 2$ and represents a risk seeking gambler. She will bet all her capital as soon as $p > \frac{1}{4}$. The other three non-linear curves belong to risk averse gamblers with $\gamma = \frac{2}{3}, \frac{1}{2}, \frac{1}{3}$ respectively. The smaller γ, the lower the fraction which will be bet. The limiting case $\gamma \to 0$ corresponds to the logarithmic utility $U(y) = \log(y)$. In this case we have to maximize

$$\sup_{a\in[0,1]} \big\{p\log(1+a) + (1-p)\log(1-a)\big\}$$

and the optimal policy is given by $f_n^*(x) = 0$ if $p \le \frac{1}{2}$ and $f_n^*(x) = 2p-1$ for $p > \frac{1}{2}$.

Case 2: Let $U(y) = \frac{1}{\gamma}e^{\gamma y}$ with $\gamma \neq 0$. Here we describe the action in terms of the amount of money that the gambler bets. Hence $E := \mathbb{R}_+$ and $A = \mathbb{R}_+$ where $D(x) = [0,x]$. We obtain $X_{k+1} = X_k + A_k Z_{k+1}$. The value iteration is given by

$$V_n(x) = \sup_{a\in[0,x]} \int V_{n-1}\big(x + az\big) Q^Z(dz).$$

We have to start the iteration with $V_0(x) := \frac{1}{\gamma}e^{\gamma x}$ and are interested in obtaining $V_N(x_0)$. The solution is more complicated in this case. We distinguish between $\gamma < 0$ and $\gamma > 0$:

Case $\gamma > 0$: : This is the *risk-seeking* case. It follows from Proposition 2.4.21 in Bäuerle & Rieder (2011) that the value functions are convex and hence a *bang-bang* policy is optimal. We compute the optimal stake $f_1^*(x)$ for one game. It is given by

$$f_1^*(x) = \begin{cases} 0 & \text{if } p \le \frac{1-e^{-\gamma x}}{e^{\gamma x} - e^{-\gamma x}} =: p(x,\gamma), \\ x & \text{else.} \end{cases}$$

Note that the critical level $p(x,\gamma)$ has the following properties:

$$0 \le p(x,\gamma) \le \frac{1}{2}$$

$$\lim_{x\to\infty} p(x,\gamma) = 0, \quad \text{and} \quad \lim_{x\to 0} p(x,\gamma) = \frac{1}{2}$$

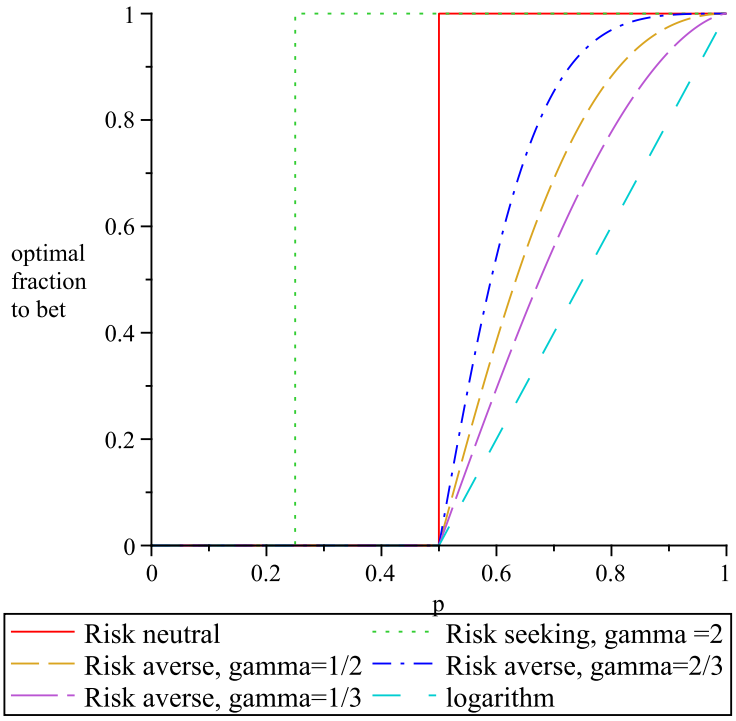

FIGURE 1. Optimal fractions of wealth to bet in the case of power utility with different γ.

which means that a gambler with low capital will behave approximately as a risk neutral gambler and someone with a large capital will stake the complete capital even when the probability of winning is quite small. Similarly

$$\lim_{\gamma\to\infty} p(x,\gamma) = 0, \quad \text{and} \quad \lim_{\gamma\to 0} p(x,\gamma) = \frac{1}{2}$$

i.e. if the gambler is more risk-seeking (γ large), she will stake her whole capital even for small success probabilities. The limiting case $\gamma = 0$ corresponds to the risk neutral gambler. In figure 2 the areas below the lines show the combinations of success probability and capital where it is optimal to bet nothing, depending on different values of γ. It can be seen that this area gets smaller for larger γ, i.e. when the gambler is more risk-seeking.

Case $\gamma < 0$: : This is the *risk-averse* case. In order to obtain a simple solution we allow the gambler to take a credit, i.e. $E = \mathbb{R}$ and $A = \mathbb{R}_+$, but the stake must be non-negative. In this setting we obtain an optimal policy where decisions are independent of the time horizon and given by

$$f_n^*(x) = \begin{cases} 0 & \text{if } p \le \frac{1}{2}, \\ -\frac{1}{2\gamma}\log\big(p/(1-p)\big) & \text{else.} \end{cases}$$

The optimal amount to bet for different γ can be seen in figure 3. The smaller γ, the larger the risk aversion and the smaller the amount to bet.

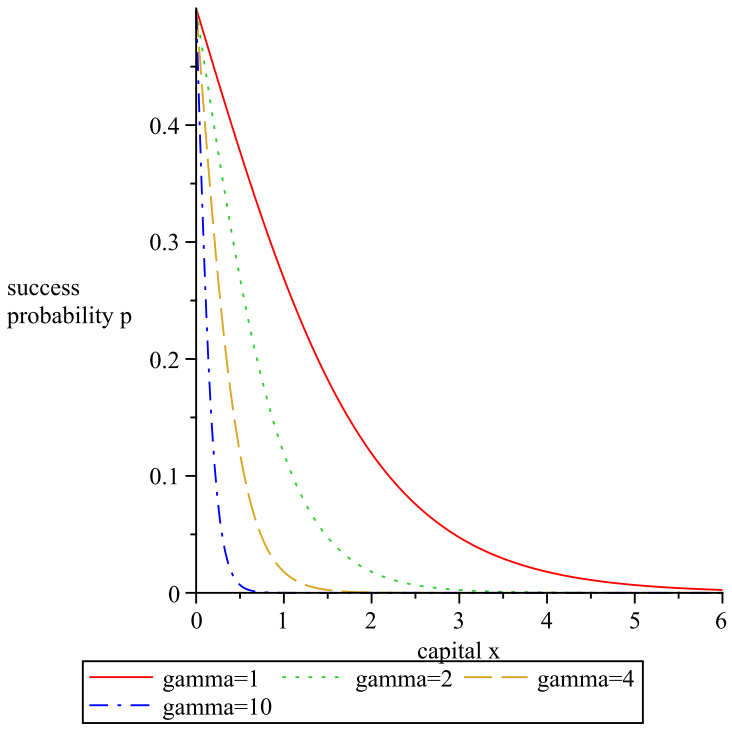

FIGURE 2. Function $p(x,\gamma)$ for different γ and exponential function.

3.3. **Strictly Discounted Problems.** Here we consider a finite time horizon and C_β^N with $\beta \in (0,1)$, i.e.

$$J_N(x) := \inf_{\sigma\in\Pi} \mathbb{E}_x^\sigma \left[U\Big(\sum_{k=0}^{N-1} \beta^k c(X_k, A_k)\Big)\right] = \inf_{\sigma\in\Pi} \mathbb{E}_x^\sigma \left[U(C_\beta^N)\right], \quad x \in E. \tag{3.7}$$

The discount factor implies some kind of non-stationarity which makes the problem more difficult. In what follow we have to introduce another state variable $z \in (0,1]$ which keeps track of the discounting. We denote now by $\hat{E} := E \times \mathbb{R}_+ \times (0,1]$ the new state space. Decision rules f are now measurable mappings from $\hat{E}$ to A respecting $f(x,y,z) \in D(x)$. Policies are defined in an obvious way. Let us denote for $n = 0, 1, \ldots, N$

$$\begin{aligned} V_{n\sigma}(x,y,z) &:= \mathbb{E}_x^\sigma \Big[U\Big(zC_\beta^n + y\Big)\Big], \quad (x,y,z) \in \hat{E}, \sigma \in \Pi, \\ V_n(x,y,z) &:= \inf_{\sigma\in\Pi} V_{n\sigma}(x,y,z), \quad (x,y,z) \in \hat{E}. \end{aligned} \tag{3.8}$$

Obviously we are interested in obtaining $V_N(x,0,1) = J_N(x)$. Let

$$\begin{aligned} \mathcal{C}(\hat{E}) := \Big\{ & v : \hat{E} \to \mathbb{R} : v \text{ is lower semicontinuous}, v(x,\cdot,\cdot) \text{ is continuous} \\ & \text{and increasing for } x \in E \text{ and } v(x,y,z) \geq U(y) \Big\}. \end{aligned}$$

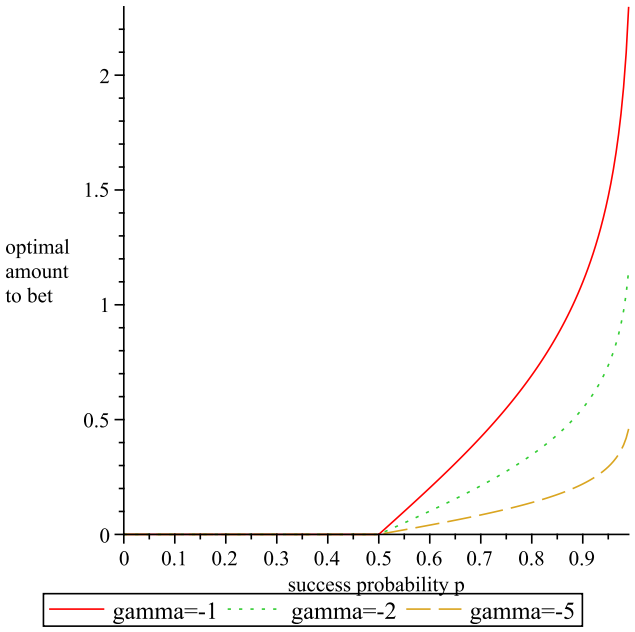

FIGURE 3. Optimal amount to bet in the case of exponential utility for different γ.

We define for $v \in \mathcal{C}(\hat{E})$ and decision rule $f \in F$ the operators

$$\begin{aligned}(T_f v)(x,y,z) &= \int v\Big(x', zc\big(x, f(x,y,z)\big) + y, z\beta\Big) Q\big(dx'|x, f(x,y,z)\big), \quad (x,y,z) \in \hat{E},\\ (Tv)(x,y,z) &= \inf_{a \in D(x)} \int v\Big(x', zc(x,a) + y, z\beta\Big) Q\big(dx'|x,a\big), \quad (x,y,z) \in \hat{E}. \end{aligned} \tag{3.9}$$

Note that both operators are increasing in the sense that if $v, w \in \mathcal{C}(\hat{E})$ with $v \le w$, then $T_f v \le T_f w$ and $Tv \le Tw$.

Then we obtain the main result for discounted problems:

Theorem 3.6. *It holds that*

a) *For a policy $\pi = (f_0, f_1, f_2, \ldots) \in \Pi^M$ we have the following cost iteration: $V_{n\pi} = T_{f_0} \ldots T_{f_{n-1}} U$ for $n = 1, \ldots, N$.*

b) *$V_0(x,y,z) := U(y)$ and $V_n = TV_{n-1}$, i.e.*

$$V_n(x,y,z) = \inf_{a \in D(x)} \int V_{n-1}\Big(x', zc(x,a) + y, z\beta\Big) Q\big(dx'|x,a\big), \quad n = 1, \ldots, N.$$

Moreover, $V_n \in \mathcal{C}(\hat{E})$.

c) *For every $n = 1, \ldots, N$ there exists a minimizer $f_n^* \in F$ of V_{n-1} and $(g_0^*, \ldots, g_{N-1}^*)$ with*

$$\begin{aligned} g_0^*(x_0) &:= f_N^*(x_0, 0, 1),\\ g_n^*(h_n) &:= f_{N-n}^*\Big(x_n, \sum_{k=0}^{n-1} \beta^k c(x_k, a_k), \beta^n\Big) \end{aligned}$$

is an optimal policy for (3.7).

Proof. We prove part a) by induction on n.

Note that $V_{0\pi}(x,y,z) = U(y)$ and let $\pi = (f_0, f_1, f_2, \ldots) \in \Pi^M$. We have

$$V_{1\pi}(x,y,z) = U(zc(x, f_0(x,y,z)) + y) = (T_{f_0}U)(x,y,z).$$

Now suppose the statement holds for $V_{n-1\pi}$ and consider $V_{n\pi}$.

$$\begin{aligned}
(T_{f_0}\ldots T_{f_{n-1}}U)(x,y,z) &= \int V_{n-1\vec{\pi}}\Big(x', zc(x,f_0(x,y,z)) + y, z\beta\Big)Q(dx'|x, f_0(x,y,z)) \\
&= \int \mathbb{E}^{\vec{\pi}}_{x'}\Big[U\Big(z\beta\sum_{k=0}^{n-2}\beta^k c(X_k,A_k) + zc(x,f_0(x,y,z)) + y\Big)\Big]Q(dx'|x,f_0(x,y,z)) \\
&= \int \mathbb{E}^{\vec{\pi}}_{x'}\Big[U\Big(z\sum_{k=0}^{n-2}\beta^{k+1} c(X_k,A_k) + zc(x,f_0(x,y,z)) + y\Big)\Big]Q(dx'|x,f_0(x,y,z)) \\
&= \mathbb{E}^{\pi}_{x}\Big[U\Big(z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k) + y\Big)\Big] \\
&= V_{n\pi}(x,y,z).
\end{aligned}$$

The remaining statements follow similarly to the proof of Theorem 3.1. We show that whenever $v \in \mathcal{C}(\hat{E})$ then $Tv \in \mathcal{C}(\hat{E})$ and there exists a minimizer for v. The proof is along the same lines as in Theorem 3.1. For the inequality note that we obtain directly $U(y) \le (Tv)(x,y,z)$ and the statements follows. □

The next corollary can be shown by induction. It states that the value iteration not only simplifies in the case of an exponential utility, but also in the case of a power or logarithmic utility. Note that part b) and part a) with $\gamma < 0$ do not follow directly from the previous theorem since $U(0+)$ is not finite. However because $\underline{c} > 0$ we can use similar arguments to prove the statements.

Corollary 3.7. a) *In case $U(y) = \frac{1}{\gamma}y^\gamma$ with $\gamma \neq 0$, we obtain $V_n(x,y,z) = z^\gamma d_n(x, \frac{y}{z})$ and $J_N(x) = d_N(x,0)$. The iteration for the $d_n(\cdot)$ simplifies to $d_0(x,y) = U(y)$ and*

$$d_n(x,y) = \beta^\gamma \inf_{a\in D(x)} \int d_{n-1}\Big(x', \frac{c(x,a)+y}{\beta}\Big)Q(dx'|x,a).$$

b) *In case $U(y) = \log(y)$, we obtain $V_n(x,y,z) = \log(z) + d_n(x, \frac{y}{z})$ and $J_N(x) = d_N(x,0)$. The iteration for the $d_n(\cdot)$ simplifies to $d_0(x,y) = U(y)$ and*

$$d_n(x,y) = \log(\beta) + \inf_{a\in D(x)} \int d_{n-1}\Big(x', \frac{c(x,a)+y}{\beta}\Big)Q(dx'|x,a).$$

c) *In case $U(y) = \frac{1}{\gamma}e^{\gamma y}$ with $\gamma \neq 0$, we obtain $V_n(x,y,z) = e^{\gamma y}h_n(x,z)$ and $J_N(x) = h_N(x,1)$. The iteration for the $h_n(\cdot)$ simplifies to $h_0(x,z) = \frac{1}{\gamma}$ and*

$$h_n(x,z) = \inf_{a\in D(x)} e^{z\gamma c(x,a)} \int h_{n-1}(x', z\beta)Q\big(dx'|x,a\big). \tag{3.10}$$

Remark 3.8. Note that the iteration in (3.10) already appears in Di Masi & Stettner (1999) p.68. However, there the authors do not consider a finite horizon problem.

4. Infinite Horizon Discounted Problems

Here we consider an infinite time horizon and C^∞_β with $\beta \in (0,1)$, i.e. we are interested in

$$J_\infty(x) := \inf_{\sigma\in\Pi} \mathbb{E}^\sigma_x\Big[U\Big(\sum_{k=0}^{\infty}\beta^k c(X_k,A_k)\Big)\Big] = \inf_{\sigma\in\Pi}\mathbb{E}^\sigma_x\big[U(C^\infty_\beta)\big], \quad x \in E. \tag{4.1}$$

We will consider concave and convex utility functions separately.

4.1. **Concave Utility Function.** We first investigate the case of a concave utility function $U : \mathbb{R}_+ \to \mathbb{R}$. This situation represents a risk seeking decision maker.

In this subsection we use the following notations:

$$\begin{aligned} V_{\infty\sigma}(x,y,z) &:= \mathbb{E}_x^\sigma \left[U(zC_\beta^\infty + y)\right], \\ V_\infty(x,y,z) &:= \inf_{\sigma\in\Pi} V_{\infty\sigma}(x,y,z), \qquad (x,y,z) \in \hat{E}. \end{aligned} \tag{4.2}$$

We are interested in obtaining $V_\infty(x,0,1) = J_\infty(x)$. For a stationary policy $\pi = (f,f,\ldots) \in \Pi^M$ we write $V_{\infty\pi} = V_f$ and denote $\bar{b}(y,z) := U(z\bar{c}/(1-\beta)+y)$ and $\underline{b}(y,z) := U(z\underline{c}/(1-\beta)+y)$.

Theorem 4.1. *The following statements hold true:*

a) *V_∞ is the unique solution of $v = Tv$ in $\mathcal{C}(\hat{E})$ with $\underline{b}(y,z) \le v(x,y,z) \le \bar{b}(y,z)$ for T defined in (3.9). Moreover, $T^nU \uparrow V_\infty$ and $T^n\bar{b} \downarrow V_\infty$ for $n \to \infty$.*

b) *There exists a minimizer f^* of V_∞ and $(g_0^*, g_1^*, \ldots)$ with*

$$g_n^*(h_n) = f^*\Big(x_n, \sum_{k=0}^{n-1} \beta^k c(x_k,a_k), \beta^n\Big)$$

is an optimal policy for (4.1).

Proof. a) We first show that $V_n = T^nU \uparrow V_\infty$ for $n \to \infty$. To this end note that for $U : \mathbb{R}_+ \to \mathbb{R}$ increasing and concave we obtain the inequality

$$U(y_1+y_2) \le U(y_1) + U'_-(y_1)y_2, \quad y_1,y_2 \ge 0$$

where U'_- is the left-hand side derivative of U which exists since U in concave. Moreover, $U'_-(y) \ge 0$ and U' is decreasing. For $(x,y,z) \in \hat{E}$ and $\sigma \in \Pi$ it holds

$$\begin{aligned} V_n(x,y,z) &\le V_{n\sigma}(x,y,z) \le V_{\infty\sigma}(x,y,z) = \mathbb{E}_x^\sigma[U(zC_\beta^\infty + y)] \\ &= \mathbb{E}_x^\sigma\Big[U\Big(zC_\beta^n + y + \beta^n z\sum_{k=n}^\infty \beta^{k-n}c(X_k,A_k)\Big)\Big] \\ &\le \mathbb{E}_x^\sigma[U(zC_\beta^n+y)] + \mathbb{E}_x^\sigma[U'_-(zC_\beta^n+y)]\beta^n\frac{z\bar{c}}{1-\beta} \\ &\le V_{n\sigma}(x,y,z) + U'_-(z\underline{c}+y)\beta^n\frac{z\bar{c}}{1-\beta} = V_{n\sigma}(x,y,z) + \varepsilon_n(y,z), \end{aligned}$$

where $\varepsilon_n(y,z) := U'_-(z\underline{c}+y)\beta^n\frac{z\bar{c}}{1-\beta}$.

Obviously $\lim_{n\to\infty}\varepsilon_n(y,z) = 0$. Taking the infimum over all policies in the preceding inequality yields:

$$V_n(x,y,z) \le V_\infty(x,y,z) \le V_n(x,y,z) + \varepsilon_n(y,z).$$

Letting $n \to \infty$ yields $V_n = T^nU \uparrow V_\infty$ for $n \to \infty$.

Obviously $\underline{b} \le V_\infty \le \bar{b}$. We next show that $V_\infty = TV_\infty$. Note that $V_n \le V_\infty$ for all n. Since T is increasing we have $V_{n+1} = TV_n \le TV_\infty$ for all n. Letting $n \to \infty$ implies $V_\infty \le TV_\infty$. For the reverse inequality recall $V_n + \varepsilon_n \ge V_\infty$. Applying the T-operator yields $V_{n+1} + \varepsilon_{n+1} \ge T(V_n+\varepsilon_n) \ge TV_\infty$ and letting $n \to \infty$ we obtain $V_\infty \ge TV_\infty$. Hence it follows $V_\infty = TV_\infty$.

Next, we obtain

$$\begin{aligned} T\bar{b}(y,z) &= \inf_{a\in D(x)} U\Big(\frac{z\beta\bar{c}}{1-\beta} + zc(x,a) + y\Big) \le U\Big(z\big(\frac{\beta\bar{c}}{1-\beta} + \bar{c}\big) + y\Big) \\ &= U\Big(\frac{z\bar{c}}{1-\beta} + y\Big) = \bar{b}(y,z). \end{aligned}$$

Analogously $T\underline{b} \geq \underline{b}$. Thus we get that $T^n\bar{b} \downarrow$ and $T^n\underline{b} \uparrow$ and the limits exist. Moreover, we obtain by iteration:

$$\begin{aligned}(T^nU)(x,y,z) &= \inf_{\pi\in\Pi^M} \mathbb{E}_x^\pi \Big[U\Big(z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k)+y\Big)\Big]\\ (T^n\bar{b})(x,y,z) &= \inf_{\pi\in\Pi^M} \mathbb{E}_x^\pi \Big[U\Big(\frac{z\bar{c}\beta^n}{1-\beta}+z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k)+y\Big)\Big]\end{aligned}$$

Using $U(y_1+y_2)-U(y_1) \leq U'_-(y_1)y_2$ we obtain:

$$\begin{aligned}0 &\leq (T^n\bar{b})(y,z,x)-(T^n\underline{b})(x,y,z) \leq (T^n\bar{b})(y,z,x)-(T^nU)(x,y,z)\\ &\leq \sup_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[U\Big(\frac{z\bar{c}\beta^n}{1-\beta}+z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k)+y\Big)-U\Big(z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k)+y\Big)\Big]\\ &\leq \varepsilon_n(y,z)\end{aligned}$$

and the right-hand side converges to zero for $n\to\infty$. As a result $T^n\bar{b}\downarrow V_\infty$ and $T^n\underline{b}\uparrow V_\infty$ for $n\to\infty$.

Since V_n is lower semicontinuous, this yields immediately that V_∞ is again lower semicontinuous. Moreover, $(y,z)\mapsto (T^n\bar{b})(x,y,z)$ is upper semicontinuous which yields together with $T^n\bar{b}\downarrow V_\infty$ that $(y,z)\mapsto V_\infty(x,y,z)$ is upper semicontinuous. Altogether $V_\infty \in \mathcal{C}(\hat{E})$.

For the uniqueness suppose that $v\in\mathcal{C}(\hat{E})$ is another solution of $v=Tv$ with $\underline{b}\leq v\leq \bar{b}$. Then $T^n\underline{b}\leq v\leq T^n\bar{b}$ for all $n\in\mathbb{N}$ and since the limit $n\to\infty$ of the right and left-hand side are equal to V_∞ the statement follows.

b) The existence of a minimizer follows from (CC) as in the proof of Theorem 3.1. From our assumption and the fact that $V_\infty(x,y,z)\geq U(y)$ we obtain

$$V_\infty = \lim_{n\to\infty} T_{f^*}^n V_\infty \geq \lim_{n\to\infty} T_{f^*}^n U = \lim_{n\to\infty} V_{n(f^*,f^*,\ldots)} = V_{f^*} \geq V_\infty$$

where the last equation follows with dominated convergence. Hence $(g_0^*,g_1^*,\ldots)$ is optimal for (4.1).

□

Obviously it can be shown that for a policy $\pi=(f_0,f_1,f_2,\ldots)\in\Pi^M$ we have the following cost iteration: $V_{\infty\pi}(x,y,z)=\lim_{n\to\infty}(T_{f_0}\ldots T_{f_n}U)(x,y,z)$. For a stationary policy $(f,f,\ldots)\in\Pi^M$ the cost iteration reads $V_f=T_fV_f$.

Remark 4.2. Consider now the reward maximization problem of Remark 3.4 with discounting and an infinite time horizon, i.e.

$$J_\infty(x) := \sup_{\sigma\in\Pi}\mathbb{E}_x^\sigma\Big[U\Big(\sum_{k=0}^{\infty}\beta^k r(X_k,A_k)\Big)\Big],\quad x\in E. \tag{4.3}$$

Define

$$V_\infty(x,y,z) := \sup_{\sigma\in\Pi}\mathbb{E}_x^\sigma\Big[U\Big(z\sum_{k=0}^{\infty}\beta^k r(X_k,A_k)+y\Big)\Big],\qquad (x,y,z)\in\hat{E}.$$

Using again the fact that V_n is increasing and bounded we obtain that $\lim_{n\to\infty}V_n$ exists. Moreover, we obtain for all $\sigma\in\Pi$ that $V_{\infty\sigma}\leq V_{n\sigma}+\varepsilon_n$ with the same ε_n as in Theorem 4.1. This implies $V_n\leq V_\infty\leq V_n+\varepsilon_n$ which in turn yields $\lim_{n\to\infty}V_n=V_\infty$. The fact that $V_\infty=TV_\infty$ can be shown as in Theorem 4.1. Also it holds that if f^* is a maximizer of V_∞, then $(g_0^*,g_1^*,\ldots)$ defined in Theorem 4.1, is an optimal policy. This follows since

$$V_\infty = \lim_{n\to\infty}T_{f^*}^nV_\infty \leq \lim_{n\to\infty}T_{f^*}^n(U+\varepsilon_0)\leq \lim_{n\to\infty}(T_{f^*}^nU+\varepsilon_n)=V_{f^*}$$

which implies the result.

For computational reasons it is interesting to know that the optimal policy can be found among stationary policies in Π^M and that the value of the infinite horizon problem can be approximated arbitrarily close by the 'sandwich method' $T^n U \leq V_\infty \leq T^n b$. Moreover, also the policy improvement works in this setting. This is formulated in the next theorem. For a decision rule $f \in F$ and $(x,y,z) \in \hat{E}$ denote $D(x,y,z,f) := \{a \in D(x) : LV_f(x,y,z,a) < V_f(x,y,z)\}$.

Theorem 4.3 (Policy improvement)**.** *Suppose $f \in F$ is an arbitrary decision rule.*

a) *Define a decision rule $h \in F$ by $h(\cdot) \in D(\cdot, f)$ if the set $D(\cdot, f)$ is not empty and by $h = f$ else. Then $V_h \leq V_f$ and the improvement is strict in states with $D(\cdot, f) \neq \emptyset$.*
b) *If $D(\cdot, f) = \emptyset$ for all states, then $V_f = V_\infty$ and f defines an optimal policy as in Theorem 4.1.*
c) *Suppose f_{k+1} is a minimizer of V_{f_k} for $k \in \mathbb{N}_0$ where $f_0 = f$. Then $V_{f_{k+1}} \leq V_{f_k}$ and $\lim_{k\to\infty} V_{f_k} = V_\infty$.*

Proof. a) By definition of h we obtain $T_h V_f(x,y,z) < V_f(x,y,z)$ in those states where $D(x,y,z,f) \neq \emptyset$, else we have $T_h V_f(x,y,z) = V_f(x,y,z)$. Thus, by induction we obtain

$$V_f \geq T_h V_f \geq T_h^n V_f \geq T_h^n U.$$

Since the right hand side converges to V_h, the statement follows. Note that the first inequality is strict for states with $D(x,y,z,f) \neq \emptyset$.

b) Our assumption implies that $TV_f \geq V_f$. Since we always have $TV_f \leq T_f V_f = V_f$ we obtain $TV_f = V_f$. Moreover $V_\infty \leq V_f \leq b$ which implies that $V_f = V_\infty$ since $T^n b \downarrow V_\infty$ for $n \to \infty$.

c) Since by construction the sequence (V_{f_k}) is decreasing we obtain $\lim_{k\to\infty} V_{f_k} =: \underline{V}$ exists and $\underline{V} \geq V_\infty$. We show now that $\lim_{k\to\infty} TV_{f_k} = T\underline{V}$. Since $V_{f_k} \geq \underline{V}$ it follows immediately that $\lim_{k\to\infty} TV_{f_k} \geq T\underline{V}$. Now for the reverse inequality note that $TV_{f_k} \leq LV_{f_k}(\cdot, a)$ for all admissible actions a. Taking the limit $k \to \infty$ on both sides yields with monotone convergence that $\lim_{k\to\infty} TV_{f_k} \leq L\underline{V}(\cdot, a)$ for all admissible actions a. Taking the infimum over all admissible a yields $\lim_{k\to\infty} TV_{f_k} \leq T\underline{V}$. Next by construction of the sequence (f_k) we obtain

$$V_{f_{k+1}} = T_{f_{k+1}} V_{f_{k+1}} \leq T_{f_{k+1}} V_{f_k} = TV_{f_k} \leq V_{f_k}.$$

Taking the limit $k \to \infty$ on both sides and applying our previous findings yields $\underline{V} = T\underline{V}$. Since $\underline{V} \leq b$ we obtain $\underline{V} \leq T^n b$ and with $n \to \infty$: $\underline{V} \leq V_\infty$. Altogether we have $\underline{V} = V_\infty$ and the statement is shown.

□

4.2. **Convex Utility Function.** Here we consider the problem with convex utility U. This situation represents a risk averse decision maker. The value functions $V_{n\sigma}, V_n, V_{\infty\sigma}, V_\infty$ are defined as in the previous section.

Theorem 4.4. *Theorem 4.1 also holds for convex U.*

Proof. The proof follows along the same lines as in Theorem 4.1. The only difference is that we have to use another inequality: Note that for $U : \mathbb{R}_+ \to \mathbb{R}$ increasing and convex we obtain the inequality

$$U(y_1 + y_2) \leq U(y_1) + U'_+(y_1 + y_2)y_2, \quad y_1, y_2 \geq 0$$

where U'_+ is the right-hand side derivative of U which exists since U in convex. Moreover, $U'_+(y) \geq 0$ and U' is increasing. Thus, we obtain for $(x,y,z) \in \hat{E}$ and $\sigma \in \Pi$:

$$\begin{aligned} V_n(x,y,z) &\leq V_{n\sigma}(x,y,z) \leq V_{\infty\sigma}(x,y,z) = \mathbb{E}_x^\sigma[U(zC_\beta^\infty + y)] \\ &= \mathbb{E}_x^\sigma\Big[U\Big(zC_\beta^n + y + z\sum_{k=n}^{\infty}\beta^k c(X_k,A_k)\Big)\Big] \\ &\leq \mathbb{E}_x^\sigma[U(zC_\beta^n + y)] + \mathbb{E}_x^\sigma\Big[U'_+\big(zC_\beta^\infty + y\big)z\sum_{k=n}^{\infty}\beta^k c(X_k,A_k)\Big] \\ &\leq \mathbb{E}_x^\sigma[U(zC_\beta^n + y)] + U'_+\Big(\frac{z\bar{c}}{1-\beta} + y\Big)\frac{z\bar{c}\beta^n}{1-\beta} \end{aligned}$$

Note that the last inequality follows from the fact that c is bounded from above by $\bar{c}$. Now denote $\delta_n(y,z) := U'_+\Big(\frac{z\bar{c}}{1-\beta} + y\Big)\frac{z\bar{c}\beta^n}{1-\beta}$. Obviously $\lim_{n\to\infty}\delta_n(y,z) = 0$. Taking the infimum over all policies in the above inequality yields:

$$V_n(x,y,z) \leq V_\infty(x,y,z) \leq V_n(x,y,z) + \delta_n(y,z).$$

Letting $n \to \infty$ yields $T^n U \to V_\infty$.

Further we have to use the inequality

$$\begin{aligned} 0 &\leq (T^n\bar{b})(y,z,x) - (T^n\underline{b})(x,y,z) \leq (T^n\bar{b})(x,y,z) - (T^nU)(x,y,z) \\ &\leq \sup_{\pi\in\Pi}\mathbb{E}_x^\pi\Big[U\Big(\frac{z\bar{c}\beta^n}{1-\beta} + z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k) + y\Big) - U\Big(z\sum_{k=0}^{n-1}\beta^k c(X_k,A_k) + y\Big)\Big] \\ &\leq U'_+\Big(\frac{z\bar{c}}{1-\beta} + y\Big)\frac{z\bar{c}\beta^n}{1-\beta} = \delta_n(y,z) \end{aligned}$$

and the right-hand side converges to zero for $n \to \infty$. □

The policy improvement for convex utility functions works in exactly the same way as for the concave case and we do not repeat it here.

From Theorem 4.1 and Theorem 4.4 we obtain (again part b) and part a) with $\gamma < 0$ can be shown by similar arguments):

Corollary 4.5. a) *In case $U(y) = \frac{1}{\gamma}y^\gamma$ with $\gamma \neq 0$, we obtain $V_\infty(x,y,z) = z^\gamma d_\infty(x,\frac{y}{z})$ and $J_\infty(x) = d_\infty(x,0)$. The function $d_\infty(\cdot)$ is the unique fixed point of*

$$d_\infty(x,y) = \beta^\gamma \inf_{a\in D(x)}\int d_\infty\Big(x', \frac{c(x,a)+y}{\beta}\Big)Q(dx'|x,a)$$

with $U(\frac{\underline{c}}{1-\beta} + y) \leq d_\infty(x,y) \leq U(\frac{\bar{c}}{1-\beta} + y)$.

b) *In case $U(y) = \log(y)$, we obtain $V_\infty(x,y,z) = \log(z) + d_\infty(x,\frac{y}{z})$ and $J_\infty(x) = d_\infty(x,0)$. The function $d_\infty(\cdot)$ is the unique fixed point of*

$$d_\infty(x,y) = \log(\beta) + \inf_{a\in D(x)}\int d_\infty\Big(x', \frac{c(x,a)+y}{\beta}\Big)Q(dx'|x,a)$$

with $U(\frac{\underline{c}}{1-\beta} + y) \leq d_\infty(x,y) \leq U(\frac{\bar{c}}{1-\beta} + y)$.

c) *In case $U(y) = \frac{1}{\gamma}e^{\gamma y}$ with $\gamma \neq 0$, we obtain $V_\infty(x,y,z) = e^{\gamma y}h_\infty(x,z)$ and $J_\infty(x) = h_\infty(x,1)$. The function $h_\infty(\cdot)$ is the unique fixed point of*

$$h_\infty(x,z) = \inf_{a\in D(x)} e^{z\gamma c(x,a)}\int h_\infty(x', z\beta)Q\big(dx'|x,a\big)$$

with $U(\frac{z\underline{c}}{1-\beta}) \leq h_\infty(x,z) \leq U(\frac{z\bar{c}}{1-\beta})$.

5. Risk-sensitive Average Cost

Let us now consider the case of average cost, i.e. for $\sigma \in \Pi$ consider

$$\begin{aligned} J_\sigma(x) &:= \limsup_{n\to\infty} \frac{1}{n} U^{-1}\Big(\mathbb{E}_x^\sigma \Big[U\big(\sum_{k=0}^{n-1} c(X_k, A_k)\big)\Big]\Big), \quad x \in E \\ J(x) &= \inf_{\sigma\in\Pi} J_\sigma(x), \quad x \in E. \end{aligned} \tag{5.1}$$

Note that we have $J_\pi(x) \in [\underline{c}, \bar{c}]$ for all $x \in E$.

5.1. **Power Utility Function.** In the case of a positive homogeneous utility function $U(y) = y^\gamma$ with $\gamma > 0$ we obtain:

$$J_\sigma(x) = \limsup_{n\to\infty} \frac{1}{n} U^{-1}\Big(\mathbb{E}_x^\sigma \Big[U(C^n)\Big]\Big) = \limsup_{n\to\infty} U^{-1}\Big(\mathbb{E}_x^\sigma \Big[U\big(\frac{C^n}{n}\big)\Big]\Big).$$

Hence we obtain the following result:

Theorem 5.1. *Suppose that $\pi = (f, f, \ldots) \in \Pi^M$ is a stationary policy such that the corresponding controlled Markov chain (X_n) is positive Harris recurrent. Then $J_\pi(x)$ exists and is independent of $x \in E$ and γ. In particular, it coincides with the average cost of a risk neutral decision maker.*

Proof. Theorem 17.0.1 in Meyn & Tweedie (2009) implies that

$$\lim_{n\to\infty} \frac{C^n}{n} = \int c(x, f(x))\mu_f(dx) \quad \mathbb{P}^\pi - a.s.$$

where μ_f is the invariant distribution of (X_n) under $\mathbb{P}^\pi$. By dominated convergence and since U is increasing, we obtain

$$U^{-1}\Big(\mathbb{E}_x^\pi \Big[U\big(\lim_{n\to\infty} \frac{C^n}{n}\big)\Big]\Big) = \lim_{n\to\infty} U^{-1}\Big(\mathbb{E}_x^\pi \Big[U\big(\frac{C^n}{n}\big)\Big]\Big) = J_\pi(x).$$

Note in particular, the limit is a real number and we can skip the expectation on the left hand side which yields the result. □

In the following theorem we assume that the MDP is *positive Harris recurrent*, i.e. for every stationary policy the corresponding state process is positive Harris recurrent.

Theorem 5.2. *Let $\gamma \geq 1$ and suppose that the MDP is positive Harris recurrent. Let $\pi^* = (f^*, f^*; \ldots)$ be an optimal stationary policy for the risk neutral average cost problem. Then π^* is optimal for problem* (5.1)*. Note that the optimal policy does not depend on γ.*

Proof. Suppose $\pi^* = (f^*, f^*; \ldots)$ is an optimal policy for the risk neutral expected average cost problem and let

$$g := \lim_{n\to\infty} \mathbb{E}_x^{\pi^*} \Big[\frac{C^n}{n} \Big] = \int c(x, f^*(x))\mu^*(dx)$$

where μ^* is the invariant distribution of (X_n) under $\mathbb{P}^{\pi^*}$. For an arbitrary policy $\sigma \in \Pi$ we obtain with the Jensen inequality and the convexity of U:

$$J_{\pi^*}(x) = g \leq \limsup_{n\to\infty} \mathbb{E}_x^\sigma \Big[\frac{C^n}{n} \Big] \leq \limsup_{n\to\infty} U^{-1}\Big(\mathbb{E}_x^\sigma \Big[U\big(\frac{C^n}{n}\big)\Big]\Big) = J_\sigma(x)$$

which implies the statement. □

For the following corollary we assume that state and action space are finite and that the MDP is unichain, i.e. for every stationary policy the corresponding state process consists of exactly one class of recurrent states and additionally of a class of transient states which could be empty.

Corollary 5.3. *Let $\gamma \geq 1$ and E and A be finite and suppose the MDP is unichain. Then there exists an optimal stationary policy for* (5.1) *which is independent of γ.*

Proof. Note that under these conditions it is a classical result (see e.g. Sennott (1999), Chapter 6.2) that there exists an optimal stationary policy for the risk-neutral decision maker. □

Finally we obtain the following connection to the discounted problem.

Theorem 5.4. *Let $\pi \in \Pi^M$ and suppose that $g_\pi := \lim_{n\to\infty} \frac{C^n}{n}$ exists $\mathbb{P}^\pi$-a.s. Then it holds*

$$g_\pi \;=\; J_\pi(x) = \lim_{n\to\infty} \frac{1}{n} U^{-1}\Big(\mathbb{E}_x^\pi\Big[U(C^n)\Big]\Big) = \lim_{\beta\uparrow 1}(1-\beta)U^{-1}\Big(\mathbb{E}_x^\pi\, U\big(C_\beta^\infty\big)\Big)$$

Proof. From a well-known Tauberian theorem (see e.g. Sennott (1999) Theorem A.4.2) we obtain

$$\lim_{n\to\infty}\frac{1}{n}C^n \le \liminf_{\beta\uparrow 1}(1-\beta)C_\beta^\infty \le \limsup_{\beta\uparrow 1}(1-\beta)C_\beta^\infty \le \lim_{n\to\infty}\frac{1}{n}C^n$$

$\mathbb{P}^\pi$-a.s. and hence $g_\pi = \lim_{\beta\uparrow 1}(1-\beta)C_\beta^\infty$ $\mathbb{P}^\pi$-a.s. Because of the fact that U is increasing and continuous we obtain

$$\lim_{n\to\infty} U\Big(\frac{C^n}{n}\Big) \le \liminf_{\beta\uparrow 1}(1-\beta)^\gamma U\Big(C_\beta^\infty\Big) \le \limsup_{\beta\uparrow 1}(1-\beta)^\gamma U\Big(C_\beta^\infty\Big) \le \lim_{n\to\infty} U\Big(\frac{C^n}{n}\Big).$$

Dominated convergence and the Lemma of Fatou yields:

$$\lim_{n\to\infty}\mathbb{E}_x^\pi U\Big(\frac{C^n}{n}\Big) \le \mathbb{E}_x^\pi\Big[\liminf_{\beta\uparrow 1}(1-\beta)^\gamma U\big(C_\beta^\infty\big)\Big] \le \liminf_{\beta\uparrow 1}(1-\beta)^\gamma\,\mathbb{E}_x^\pi\, U\big(C_\beta^\infty\big)$$

$$\le \limsup_{\beta\uparrow 1}(1-\beta)^\gamma\,\mathbb{E}_x^\pi\, U\big(C_\beta^\infty\big) \le \mathbb{E}_x^\pi\Big[\limsup_{\beta\uparrow 1}(1-\beta)^\gamma U\big(C_\beta^\infty\big)\Big] \le \lim_{n\to\infty}\mathbb{E}_x^\pi U\Big(\frac{C^n}{n}\Big)$$

which implies the statement. □

Obviously Theorem 5.4 shows that the so-called *vanishing discount approach* works in this setting in contrast to the classical risk sensitive case.

5.2. **Relation to risk measures.** Another reasonable optimization problem would be to consider $\limsup_{n\to\infty}\frac{1}{n}\rho(C^n)$ for a risk measure ρ. In case ρ is homogeneous, i.e. $\rho(\alpha X) = \alpha\rho(X)$ for all $\alpha \ge 0$ and continuous, i.e. $\lim_{n\to\infty}\rho(X_n) = \rho(X)$ for all bounded sequences $X_n \to X$ we obtain in the case of Harris recurrent Markov chain (X_n) under a stationary policy $\pi = (f, f, \ldots) \in \Pi^M$ that

$$\lim_{n\to\infty}\frac{1}{n}\rho\Big(\sum_{k=0}^{n-1} c(X_k, f(X_k))\Big) \;=\; \lim_{n\to\infty}\rho\big(\frac{1}{n}C_f^n\big) = \rho\big(\lim_{n\to\infty}\frac{1}{n}C_f^n\big) = \rho\Big(\int c(x,f(x))\mu_f(dx)\Big)$$

$$= \;\rho(1)\int c(x,f(x))\mu_f(dx).$$

Thus, when we minimize over all stationary policies, the minimal average cost does not depend on the precise risk measure and coincides with the value in the risk neutral case (if $\rho(1) = 1$).

Note that a certainty equivalent is in general not a convex risk measure (see e.g. Müller (2007), Ben-Tal & Teboulle (2007)), the only exception is the classical risk sensitive case with $U(y) = \frac{1}{\gamma}\exp(\gamma y)$, however both share a certain kind of representation. The problem of minimizing the Average-Value-at-Risk of the average cost has been investigated in Ott (2010).

References

Barz, C. & Waldmann, K.-H. (2007). Risk-sensitive capacity control in revenue management. *Math. Methods Oper. Res.* **65**, 565–579.

Bäuerle, N. & Mundt, A. (2009). Dynamic mean-risk optimization in a binomial model. *Math. Methods Oper. Res.* **70**, 219–239.

Bäuerle, N. & Ott, J. (2011). Markov decision processes with average-value-at-risk criteria. *Math. Methods Oper. Res.* **74**, 361–379.

Bäuerle, N. & Rieder, U. (2011). *Markov Decision Processes with applications to finance*. Springer.

Ben-Tal, A. & Teboulle, M. (2007). An old-new concept of convex risk measures: the optimized certainty equivalent. *Math. Finance* **17**, 449–476.

Bielecki, T., Hernández-Hernández, D. & Pliska, S. R. (1999). Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. *Math. Methods Oper. Res.* **50**, 167–188. Financial optimization.

Bielecki, T. & Pliska, S. R. (2003). Economic properties of the risk sensitive criterion for portfolio management. *Rev. Account. Fin.* **2**, 3–17.

Boda, K., Filar, J. A., Lin, Y. & Spanjers, L. (2004). Stochastic target hitting time and the problem of early retirement. *IEEE Trans. Automat. Control* **49**, 409–419.

Cavazos-Cadena, R. & Fernández-Gaucherand, E. (2000). The vanishing discount approach in Markov chains with risk-sensitive criteria. *IEEE Trans. Automat. Control* **45**, 1800–1816.

Cavazos-Cadena, R. & Hernández-Hernández, D. (2011). Discounted approximations for risk-sensitive average criteria in Markov decision chains with finite state space. *Math. Oper. Res.* **36**, 133–146.

Chung, K. & Sobel, M. (1987). Discounted MDP's: Distribution functions and exponential utility maximization. *SIAM J Contr. Optim.* **25**, 49–62.

Collins, E. & McNamara, J. (1998). Finite-horizon dynamic optimisation when the terminal reward is a concave functional of the distribution of the final state. *Advances in Applied Probability* **30**, 122–136.

Denardo, E., Feinberg, E. & Rothblum, U. (2011). The multi-armed bandit, with constraints. *Preprint.* 1–25.

Denardo, E. V., Park, H. & Rothblum, U. G. (2007). Risk-sensitive and risk-neutral multiarmed bandits. *Math. Oper. Res.* **32**, 374–394.

Di Masi, G. B. & Stettner, L. (1999). Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM J. Control Optim.* **38**, 61–78 ,.

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

Howard, R. & Matheson, J. (1972). Risk-sensitive Markov Decision Processes. *Management Science* **18**, 356–369.

Iwamoto, S. (2004). Stochastic optimization of forward recursive functions. *J. Math. Anal. Appl.* **292**, 73–83.

Jaquette, S. (1973). Markov Decision Processes with a new optimality criterion: discrete time. *Ann. Statist.* **1**, 496–505.

Jaquette, S. (1976). A utility criterion for Markov Decision Processes. *Managem. Sci.* **23**, 43–49.

Jaśkiewicz, A. (2007). Average optimality for risk-sensitive control with general state space. *Ann. Appl. Probab.* **17**, 654–675.

Kaas, R., Goovaerts, M., Dhaene, J. & Denuit, M. (2009). *Modern Actuarial Risk Theory*. Springer-Verlag, Berlin.

Kreps, D. M. (1977a). Decision problems with expected utility criteria. I. Upper and lower convergent utility. *Math. Oper. Res.* **2**, 45–53.

Kreps, D. M. (1977b). Decision problems with expected utility criteria. II. Stationarity. *Math. Oper. Res.* **2**, 266–274.

Meyn, S. & Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Second ed., Cambridge University Press, Cambridge.

Muliere, P. & Parmigiani, G. (1993). Utility and means in the 1930s. *Statist. Sci.* **8**, 421–432.

Müller, A. (2007). Certainty equivalents as risk measures. *Braz. J. Probab. Stat.* **21**, 1–12.

Ott, J. (2010). *A Markov decision model for a surveillance application and risk-sensistive Markov decision processes*. Ph.D. thesis, Karlsruhe Institute of Technology, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000020835.

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Math. Program.* **125**, 235–261.

Sennott, L. (1999). *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley& Sons, New York.

White, D. J. (1988). Mean, variance, and probabilistic criteria in finite Markov Decision Processes: a review. *J. Optim. Theory Appl.* **56**, 1–29.

Wu, C. & Lin, Y. (1999). Minimizing risk models in Markov Decision Processes with policies depending on target values. *J. Math. Anal. Appl.* **231**, 47–67.

(N. Bäuerle) Institute for Stochastics, Karlsruhe Institute of Technology, D-76128 Karlsruhe, Germany

E-mail address: nicole.baeuerle@kit.edu

(U. Rieder) University of Ulm, D-89069 Germany

E-mail address: ulrich.rieder@uni-ulm.de