Mathematical Methods of Operations Research (2024) 99:141–178
https://doi.org/10.1007/s00186-024-00857-0

ORIGINAL ARTICLE

Markov decision processes with risk-sensitive criteria: an overview

Nicole Bäuerle[1] · **Anna Jaśkiewicz**[2]

Received: 9 November 2023 / Revised: 27 February 2024 / Accepted: 1 March 2024 /
Published online: 1 April 2024

Abstract
The paper provides an overview of the theory and applications of risk-sensitive Markov decision processes. The term 'risk-sensitive' refers here to the use of the Optimized Certainty Equivalent as a means to measure expectation and risk. This comprises the well-known entropic risk measure and Conditional Value-at-Risk. We restrict our considerations to stationary problems with an infinite time horizon. Conditions are given under which optimal policies exist and solution procedures are explained. We present both the theory when the Optimized Certainty Equivalent is applied recursively as well as the case where it is applied to the cumulated reward. Discounted as well as non-discounted models are reviewed.

Keywords Markov decision process · Risk-sensitive decision · Optimized certainty equivalent · Optimal policy

1 Introduction

The theory of Markov decision processes (MDPs) deals with stochastic, dynamic optimization problems. In the classical situation, the aim is to maximize an expected cumulative or averaged reward of a system. Since the first formulations by Richard Bellman in the 1950s, the theory has developed tremendously. In particular, one branch of literature is devoted to extending this theory beyond the simple expectation, since there is an evidence from various fields that the expectation should be replaced by some

✉ Nicole Bäuerle
nicole.baeuerle@kit.edu

Anna Jaśkiewicz
anna.jaskiewicz@pwr.edu.pl

[1] Department of Mathematics, Karlsruhe Institute of Technology (KIT), Karlsruhe 76131, Germany

[2] Faculty of Pure and Applied Mathematics, Wrocław University of Science and Technology, Wrocław, Poland

criterion which allows to model risk-sensitivity of the decision maker. This evidence comes from disciplines like psychology, economics and biology. For instance, Braun et al. (2011) reviewed evidence for risk-sensitivity in motor control tasks.

From a mathematical point of view, the decision problem gets of course more complicated when risk-sensitivity is taken into account. Loosely speaking, risk-sensitivity weights the possible fluctuations around the mean. A simple way to deal with this is to consider a weighted criterion of the expectation and the variance of a random income, i.e. to include the second moment into the decision. This has for example been propagated in Markowitz (1952). Naturally, one can generalize this idea to higher moments. One of the ways is to use an exponential function which plays a prominent role in risk-sensitive MDPs. Then, all moments of a random payoff are taken into account if we consider the expectation of an exponential function of this random payoff. This fact can be seen via the Taylor series expansion of the exponential function around 0. To be more precise let us consider for example the following expression

$$J(x,\pi) = -\frac{1}{\gamma} \ln \mathbb{E}_x^\pi \left[\exp\left(-\gamma \sum_{k=0}^{\infty} \beta^k r(X_k, A_k) \right) \right]$$

where $(X_k, A_k)_k$ is a controlled state-action process, r is a one-stage reward function, β is a discount factor, $\gamma \neq 0$ is a risk-sensitivity parameter and the transition law is determined by a policy π. The initial state is $X_0 = x$. A target function like this has first been studied in Howard and Matheson (1972). Indeed, for small γ this is approximately equal to

$$J(x,\pi) \approx \mathbb{E}_x^\pi \left[\sum_{k=0}^{\infty} \beta^k r(X_k, A_k) \right] - \frac{\gamma}{2} Var_x^\pi \left(\sum_{k=0}^{\infty} \beta^k r(X_k, A_k) \right).$$

However, from a mathematical point of view it is more tractable than the variance. From the approximation it is also obvious that $\gamma > 0$ models a risk-averse decision maker (since then the variance is subtracted), whereas $\gamma < 0$ corresponds to a risk-loving decision maker. The preceding target function is a special case of the situation we consider here in this paper. It can also be interpreted as a Certainty Equivalent of the exponential utility function. This point of view can then be generalized to Optimized Certainty Equivalents which we consider in this survey.

The aim of this paper is to provide an overview of the ideas, concepts and literature in this area. We will also discuss the situation where the Optimized Certainty Equivalent is applied to the single-stage rewards in a recursive way. However, we will stay within the setting where optimal policies are stationary in a certain sense and can be computed from optimality equations, thus naturally avoiding time-inconsistency issues. Our point of view is mainly from the economics and operations research perspective. We do not consider problems with a finite time horizon, nor do we treat problems in continuous time. These issues were described in the recent survey of Biswas and Borkar (2023).

The outline of our survey is the following. In the next section we explain and discuss our main building block for the target function: the Optimized Certainty Equivalent.

The Optimized Certainty Equivalents have been introduced by Ben-Tal and Teboulle (2007) and provide a useful generalization of Certainty Equivalents. They comprise important cases like the entropic risk measure and the Conditional Value-at-Risk and are still tractable from a mathematical point of view. In Sect. 3 we introduce the theory of Markov decision processes. We restrict our attention to stationary problems (i.e. the model data do not depend on the time point) with an infinite time horizon. Conditions are given under which optimal policies exist and a solution procedure is explained. Section 4 presents the theory when the Optimized Certainty Equivalent is applied recursively. Some generalizations and related problems are discussed at the end. Section 5, on the other hand, treats the situation when the Optimized Certainty Equivalent is applied to the cumulative reward. Here the presented solution technique is via an extension of the state space. Finally in Sect. 6 we provide an overview on the risk-sensitive average cost case. Section 7 summarizes some typical applications of the presented theory. The appendix contains two proofs.

Notation. As usual, the symbol $\mathbb{N}$ denotes the set of positive integers and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. By $\mathbb{R}$ ($\mathbb{R}_+$) we denote the set of all (non-negative) real numbers. For $x \in \mathbb{R}$ we denote by x^+ the positive part of x. The Dirac measure is given by $\delta_x(B)$ and is equal to one if $x \in B$ and zero otherwise. We use the following abbreviations: *w.r.t.* means *with respect to*, *r.h.s* means *right-hand side* and *l.h.s* means *left-hand side*.

2 Certainty equivalents and optimized certainty equivalents

Decision makers are often risk averse when faced with decisions,[1] in particular when monetary rewards or costs have to be optimized. Consider for example the following two lotteries:

- Lottery 1: receive a reward of 1000 with probability 0.05 and 0 else.
- Lottery 2: receive a reward of 50 with probability 1.

Both lotteries have an expected value of 50. However when confronted with this choice in reality, most people prefer lottery 2, since they are risk averse and consider the probability of 0.05 to be very low. Thus, it is reasonable to model risk aversion in decision making. This can be done for example by using risk measures.

In what follows let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. All random variables which appear here are defined on this space. We will consider Certainty Equivalents and *Optimized Certainty Equivalents*. Let $u : \mathbb{R} \to [-\infty, \infty)$ be a strictly increasing, strictly concave utility function. The main purpose of the utility function is to provide a systematic way to rank alternatives that captures the principle of risk aversion, see Von Neumann and Morgenstern (2007). This is accomplished whenever the utility function is concave. The degree of risk aversion exhibited by the utility function corresponds to the magnitude of the bend in the function, i.e. the stronger the bend the greater the risk aversion. The degree of risk aversion is formally defined by the

[1] The St. Petersburg Paradox which is due to Daniel Bernoulli in 1738 is often mentioned as the first discussion of this topic. For an English translation of the original paper in Latin see Bernoulli (1954).

Arrow-Pratt absolute risk aversion coefficient (Arrow 1971; Pratt 1964):

$$\gamma(x) := -\frac{u''(x)}{u'(x)}.$$

Basically, the parameter shows how risk aversion changes with the wealth level. Although the actual value of the expected utility of a random outcome is meaningless except with comparison with other alternatives, there is a derived measure with units that has intuitive meaning. The Certainty Equivalent of a bounded random income $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ is defined as

$$CE(X) = u^{-1}\mathbb{E}u(X) \tag{1}$$

where $\mathbb{E}$ is the expectation operator with respect to the probability measure $\mathbb{P}$. $CE(X)$ is the sure amount which yields the same utility as the random outcome. The Optimized Certainty Equivalent is defined as follows (Ben-Tal and Teboulle 2007):

Definition 1 Let $u : \mathbb{R} \to [-\infty, \infty)$ be a proper, closed, concave and non-decreasing utility function with $u(0) = 0$ and $u'_+(0) \le 1 \le u'_-(0)$ where u'_+ and u'_- are the right and left derivatives of u.[2] Further let $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ be a bounded random variable. The *Optimized Certainty Equivalent* (OCE) for X is a map $S_u : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ with

$$S_u(X) = \sup_{\eta \in \mathbb{R}} \{\eta + \mathbb{E}u(X - \eta)\}$$

which is assumed to be a proper function, which means that the domain $domS_u := \{X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P}) : \ S_u(X) > -\infty\}$ is not empty and S_u is finite on this domain.

The interpretation here is that the decision maker may consume the amount η today and obtain the present value $\eta + \mathbb{E}u(X - \eta)$ as a result. Optimizing over the consumption then yields the present value of X. Among others, $S_u(X)$ has the following properties for $X, Y \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ (see Ben-Tal and Teboulle 2007):

(P1) monotonicity: $X \le Y \Rightarrow S_u(X) \le S_u(Y)$;
(P2) shift additivity : $S_u(X + c) = S_u(X) + c$, for any $c \in \mathbb{R}$;
(P3) Jensen inequality: $S_u(X) \le \mathbb{E}X$;
(P4) consistency: $S_u(c) = c$ for any $c \in \mathbb{R}$.

Indeed it can be shown that $-S_u$ is a convex risk measure in the sense of Föllmer and Schied (2010). A random variable X is now preferred over Y if $S_u(X) \ge S_u(Y)$. Thus (P3) and (P4) imply that this preference order models *risk aversion* since $S_u(X) \le \mathbb{E}X = S_u(\mathbb{E}X)$, i.e. the sure amount $\mathbb{E}X$ is preferred over a random amount X with same expectation. Moreover, it holds that

$$\lim_{\delta \to 0} \frac{1}{\delta} S_u(\delta X) = \mathbb{E}X$$

[2] Note that $u(x) \ge 0$ for all $x \ge 0$ and $u(x) \le x$ for all $x \in \mathbb{R}$.

which means that the *risk-neutral* setting is achieved in the limit. Note that for simplicity we restrict here to bounded random variables. This is sufficient for the applications given in further sections where we consider bounded reward functions.

A further representation of S_u is due to Ben-Tal and Teboulle (2007) given by

$$S_u(X) = \inf_{\mathbb{Q}\in\mathcal{Q}} \{I_\varphi(\mathbb{Q}, \mathbb{P}) + \mathbb{E}_{\mathbb{Q}} X\}$$

where $\mathcal{Q}$ is the set of all probability measures $\mathbb{Q}$ absolutely continuous w.r.t. $\mathbb{P}$ such that $\frac{d\mathbb{Q}}{d\mathbb{P}} \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and I_φ is the usual φ-divergence defined by

$$I_\varphi(\mathbb{Q}, \mathbb{P}) = \begin{cases} \int \varphi\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right) d\mathbb{P}, & \text{if } \mathbb{Q} \ll \mathbb{P} \\ \infty, & \text{else.} \end{cases}$$

Here $\varphi : \mathbb{R} \to [0, +\infty]$ is a proper closed convex function with closed interval (containing 1) as domain and $\varphi(1) = 0$. This representation can be exploited in the analysis of risk-sensitive problems and in order to construct a connection to robust decision making, see Dai Pra et al. (1996), Bäuerle and Glauner (2022a). Indeed this representation consists of the risk-neutral part $\mathbb{E}_{\mathbb{Q}} X$ where however the infimum over a set $\mathcal{Q}$ of probability measures is taken. The φ-divergence term penalizes the distance of $\mathbb{Q}$ to $\mathbb{P}$. Therefore, it resembles a robust approach. The following examples list important special cases of the Optimized Certainty Equivalent.

Example 1 a) When we choose $u(t) = \frac{1}{\gamma}(1 - e^{-\gamma t})$ for $\gamma > 0$ we obtain

$$S_u(X) = -\frac{1}{\gamma} \ln \mathbb{E} e^{-\gamma X}. \tag{2}$$

The quantity $-S_u(X)$ is known as the *entropic risk measure*, see p. 184 in Föllmer and Schied (2010). However, we shall further also refer to (2) as the entropic risk measure. It is easy to see that in this case

$$S_u(X) = u^{-1}\mathbb{E}u(X)$$

coincides with the Certainty Equivalent of X w.r.t. u. A Taylor series expansion yields

$$S_u(X) \approx \mathbb{E}X - \frac{\gamma}{2} Var(X)$$

which connects the entropic risk to the mean-variance criterion. The entropic risk measure is the most widely used functional which is applied in risk-sensitive dynamic decision making. This is mainly because it is still mathematically tractable. Indeed, the paper of Howard and Matheson (1972) which is considered to be the first work in this field, coined the name *risk-sensitive Markov decision process*. Since then the adjective 'risk-sensitive' is often used as a synonym for applying the entropic risk measure.

b) If for $\alpha \in (0, 1)$ we choose

$$u(t) = -\frac{1}{\alpha}(-t)^+$$

then $S_u(X) = -CVaR_\alpha(X)$ where the risk measure *Conditional Value-at-Risk* (CVaR) is defined as

$$CVaR_\alpha(X) = \inf_{\eta \in \mathbb{R}} \left\{ \frac{1}{\alpha}\mathbb{E}(\eta - X)^+ - \eta \right\}.$$

The Conditional Value-at-Risk is sometimes also called *Average Value-at-Risk* or *Expected Shortfall.* It can be represented as

$$CVaR_\alpha(X) = \frac{1}{\alpha}\int_0^\alpha VaR_\gamma(X)d\gamma$$

where $VaR_\alpha(X) = \inf\{c \in \mathbb{R} : \mathbb{P}(X + c < 0) \le \alpha\}$ is the Value-at-Risk. In case of a continuous random variable X we also have

$$CVaR_\alpha(X) = \mathbb{E}[-X| - X \ge VaR_\alpha(X)].$$

The Conditional Value-at-Risk is not only a convex, but also a coherent risk measure. It is the smallest convex risk measure which dominates Value-at-Risk, see Remark 4.56 in Föllmer and Schied (2010).

c) If we choose

$$u(t) = \begin{cases} t - \frac{1}{2}t^2, & t < 1 \\ \frac{1}{2}, & t \ge 1 \end{cases}$$

then for random variables with $X \le 1 + \mathbb{E}X$ we obtain that $S_u(X)$ is the *mean-variance criterion*

$$S_u(X) = \mathbb{E}X - \frac{1}{2}Var(X).$$

The mean-variance criterion is a popular decision criterion in finance since its first appearance in Markowitz (1952). However the interpretation is here restricted to random variables with bounded support.

Remark 1 In what follows we consider optimization problems with rewards. Thus, we maximize S_u. In case we want to minimize cost, we have to define the criterion in a different way. In this case let $\ell : \mathbb{R} \to (-\infty, \infty]$ be a proper, closed, convex and non-decreasing function bounded from below with $\ell(0) = 0$ and $\ell'_+(0) \ge 1 \ge \ell'_-(0)$. For $X \in L^\infty(\Omega, \mathcal{F}, \mathbb{P})$ the *Optimized Certainty Equivalent* is then $S_\ell : L^\infty(\Omega, \mathcal{F}, \mathbb{P}) \to \mathbb{R}$ with

$$S_\ell(X) = \inf_{\eta \in \mathbb{R}} \{\eta + \mathbb{E}\ell(X - \eta)\}$$

which is assumed to be a proper function. For X being a cost, this criterion has to be minimized.

3 Markov decision processes

3.1 The model

By a Borel space Y we mean a non-empty Borel subset of a Polish space. We assume that Y is equipped with the Borel σ-algebra $\mathcal{B}(Y)$. For dynamic decision making we consider the following controlled Markov process in discrete time (a comprehensive treatment of this theory is e.g. given in Puterman 2014; Hernández-Lerma and Lasserre 1996; Bäuerle and Rieder 2011).

(a) The state space E is a Borel space.
(b) The action space A is a Borel space.
(c) $D \subset E \times A$ is the set of admissible state-action combinations. D contains the graph of a measurable mapping $f : E \to A$. The sets $D(x) = \{a \in A : (x, a) \in D\}$ of admissible actions in state x are assumed to be compact.
(d) q is a regular conditional distribution from D to E.
(e) The one-stage reward $r : D \to \mathbb{R}_+$ is a bounded Borel measurable function, i.e. $r(x, a) \leq d$ for all $(x, a) \in D$ for some constant $d > 0$.

We define the set of histories of the process. At time $k = 0$ we have $H_0 = E$. For $k \geq 1$ the set of histories are given by $H_k = D^k \times E$ and $H_\infty = D \times D \times \ldots$. A policy $\pi = (\pi_k)_{k\in\mathbb{N}_0}$ is a sequence of decision rules (Borel measurable mappings) from H_k to A such that $\pi_k(h_k) \in D(x_k)$ where $h_k = (x_0, a_0, \ldots, x_k) \in H_k$. The set of all policies is denoted by Π. Let F be the set of all measurable mappings $f : E \to A$ such that $f(x) \in D(x)$ for every $x \in E$. By our assumption $F \neq \emptyset$. A Markovian policy is a sequence $(f_k)_{k\in\mathbb{N}_0}$ where each $f_k \in F$. The class of Markovian policies is denoted by Π^M. A Markovian policy $(f_k)_{k\in\mathbb{N}_0}$ is stationary if there is some $f \in F$ such that $f_k = f$ for every $k \in \mathbb{N}_0$, i.e. the same decision rule f is used throughout the time. We identify a stationary policy with the element of the sequence. Therefore, the set of all stationary policies will be denoted by F. We have

$$F \subset \Pi^M \subset \Pi.$$

Let $(\Omega, \mathcal{F})$ be a measurable space consisting of the sample space $\Omega = (E \times A)^\infty$ with the corresponding product σ-algebra $\mathcal{F}$ on Ω. The elements of Ω are the sequences $\omega = (x_0, a_0, x_1, a_1, \ldots) \in H_\infty$ with $x_n \in E$ and $a_n \in A$ for $n \in \mathbb{N}_0$. The random variables $X_0, A_0, X_1, A_1, \ldots$ are defined by

$$X_k(\omega) = x_k, \quad A_k(\omega) = a_k, \qquad k \in \mathbb{N}_0$$

and represent the state and action process, respectively. Let $\pi \in \Pi$ and the initial state $x \in E$ be fixed. Then according to the Ionescu-Tulcea theorem there exists a unique probability measure $\mathbb{P}_x^\pi$ on $(\Omega, \mathcal{F})$ which is supported on H_∞, i.e. $\mathbb{P}_x^\pi(H_\infty) = 1$. Moreover, for $k \in \mathbb{N}_0$:

(a) $\mathbb{P}_x^\pi(X_0 \in B) = \delta_x(B)$ for all $B \in \mathcal{B}(E)$,
(b) $\mathbb{P}_x^\pi(A_k \in C|h_k) = \delta_{\pi_k(h_k)}(C)$ for all $h_k \in H_k$ and $C \in \mathcal{B}(A)$,
(c) $\mathbb{P}_x^\pi(X_{k+1} \in B|x, a_0, \ldots, x_k, a_k) = q(B|x_k, a_k)$ for all $B \in \mathcal{B}(E)$.

Note that the theory is also established for unbounded reward functions in which case it is common to work with the concept of so-called 'bounding functions'.

3.2 Risk-Neutral decision maker

One of the standard optimization problems for Markov decision processes is to find the maximal expected discounted reward:

$$J_\beta^*(x) = \sup_{\pi\in\Pi} J_\beta(x,\pi) \quad \text{with} \quad J_\beta(x,\pi) = \mathbb{E}_x^\pi\left[\sum_{k=0}^{\infty} \beta^k r(X_k, A_k)\right] \tag{3}$$

where $\beta \in [0, 1)$ is a discount coefficient and, if possible, an optimal policy $\pi^* \in \Pi$ such that $J_\beta^*(x) = J_\beta(x, \pi^*)$. Under some continuity and compactness assumptions, the maximal value J_β^* and an optimal policy can be characterized via the Bellman equation. In order to establish this equation we may use one of two different sets of conditions which are common in the literature, see Schäl (1975), Schäl (1983):

Condition (S):

(a) The sets $D(x)$, $x \in E$, are compact.
(b) For each $x \in E$ and every Borel set $C \subset E$ the function $q(C|x, \cdot)$ is continuous on $D(x)$.
(c) The reward $r(x, \cdot)$ is upper semicontinuous on $D(x)$ for each $x \in E$.

Condition (W):

(a) The sets $D(x)$, $x \in E$, are compact and the mapping $x \to D(x)$ is upper semicontinuous.
(b) The transition law q is weakly continuous on D, i.e. the function

$$(x, a) \to \int h(y)q(dy|x, a)$$

is continuous for each continuous bounded function h.
(c) The reward r is upper semicontinuous on D.

In what follows let $U(E)$ be the set of all bounded, non-negative upper semicontinuous functions on E and $B(E)$ the set of all bounded, non-negative Borel measurable functions on E. We equip these spaces with the supremum norm $\|\cdot\|$.

Theorem 1 *Assume (W) or (S). Then*

a) *There exist a unique function $V_\beta \in U(E)$ in case (W) holds and $V_\beta \in B(E)$ in case (S) holds and a decision rule $f^* \in F$ such that for all $x \in E$:*

$$V_\beta(x) = \sup_{a\in D(x)} \left\{r(x,a) + \beta \int V_\beta(y)q(dy|x,a)\right\} \tag{4}$$

$$= r(x, f^*(x)) + \beta \int V_\beta(y) q(dy|x, f^*(x)).$$

b) *Moreover,* $V_\beta(x) = J^*_\beta(x) = J_\beta(x, f^*)$ *for all* $x \in E$, *i.e.* $f^* \in F$ *is an optimal stationary policy.*

Theorem 1 can be used to establish the link between the expected discounted reward and the long-run average reward defined as:

$$\mathcal{J}(x, \pi) = \liminf_{n\to\infty} \frac{1}{n} \mathbb{E}^\pi_x \left[\sum_{k=0}^{n-1} r(X_k, A_k) \right]$$

for any initial state $x \in E$ and $\pi \in \Pi$. The aim is to find a policy $\pi^* \in \Pi$ such that $\mathcal{J}(x) := \sup_{\pi\in\Pi} \mathcal{J}(x, \pi) = \mathcal{J}(x, \pi^*)$ for every $x \in E$. A first relation between discounted reward and long-run average reward is provided by the Tauberian theorem of Hardy-Littlewood. Variants of this result and historical comments may be found in Appendix H in Filar and Koos (1997) or pages 417 and 432 in Puterman (2014) and references cited therein. Basically, it claims that for bounded sequences of real numbers $(R_k)_{k\in\mathbb{N}_0}$ it holds

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} R_k \le \liminf_{\beta\to 1} (1-\beta) \sum_{k=0}^{\infty} \beta^k R_k.$$

When we set $R_k := \mathbb{E}^\pi_x r(X_k, A_k)$ then we immediately obtain

$$\mathcal{J}(x, \pi) \le \liminf_{\beta\to 1} (1-\beta) J_\beta(x, \pi), \quad x \in E, \ \pi \in \Pi,$$

and consequently

$$\sup_{\pi\in\Pi} \mathcal{J}(x, \pi) \le \liminf_{\beta\to 1} (1-\beta) J^*_\beta(x), \quad x \in E.$$

A second relation is given via (4). Let $z \in E$ be a fixed state and put $h_\beta(x) := V_\beta(x) - V_\beta(z)$. Then simple rearrangements in (4) yield

$$(1-\beta) V_\beta(z) + h_\beta(x) = \sup_{a\in D(x)} \left\{ r(x, a) + \beta \int h_\beta(y) q(dy|x, a) \right\}, \quad x \in E.$$

Under certain set of conditions and letting $\beta \to 1$, the pair $((1-\beta)V_\beta(z), h_\beta(\cdot))$ would converge to a pair $(\xi, h(\cdot))$ that satisfies the average reward optimality equation

$$\xi + h(x) = \sup_{a\in D(x)} \left\{ r(x, a) + \int h(y) q(dy|x, a) \right\}, \quad x \in E. \tag{5}$$

If a set of "reasonably mild" assumptions is imposed on the family of functions $\{h_\beta(\cdot)\}$ then a pair $(\xi, h(\cdot))$ meets the average reward optimality inequality

$$\xi + h(x) \leq \sup_{a \in D(x)} \left\{ r(x,a) + \int h(y) q(dy|x,a) \right\}, \quad x \in E. \tag{6}$$

If the maximizer, say $f_* \in F$, of the r.h.s. in (5) or (6) exists, it constitutes an optimal stationary policy, i.e. $\mathcal{J}(x, f_*) = \sup_{\pi \in \Pi} \mathcal{J}(x, \pi)$ for every $x \in E$ and moreover, the optimal average reward is independent of the initial state and $\xi = \mathcal{J}(x, f_*)$. This approach is well-described in the literature. The reader is referred to Hernández-Lerma and Lasserre (1996), Piunovskiy (2013), Puterman (2014) where also other methods are presented with comments and illustrative examples.

There are a number of established computational approaches which can, often after modifications, also be applied to the risk-sensitive cases which we discuss later. For example if we consider the setting of Theorem 1, the operator

$$T_\beta v(x) := \sup_{a \in D(x)} \left\{ r(x,a) + \beta \int v(y) q(dy|x,a) \right\} \tag{7}$$

is a contraction on a suitable function space into the same function space. Then applying Banach's fixed point theorem, the value function and the optimal policy can be approximated by iterating the T_β-operator. Alternatively, one can start with an arbitrary stationary policy, given by a decision rule $f \in F$, compute the corresponding value $V_f(x) = J_\beta(x, f)$ (see (3)) and improve it by computing the maximum points on the r.h.s. of (7) with v replaced by V_f. Under mild assumptions this procedure converges to the optimal solution. For computational purposes it is often more convenient to consider the so-called Q-function, which is defined as follows

$$Q(x,a) := r(x,a) + \beta \int V_\beta(y) q(dy|x,a).$$

Note that we have $V_\beta(x) = \sup_{a \in D(x)} Q(x,a)$ and

$$Q(x,a) = r(x,a) + \beta \int \sup_{a' \in D(y)} Q(y,a') q(dy|x,a).$$

This representation has the advantage that the maximization can be done before the integration. The algorithms discussed so far are only applicable when the state and action spaces are of low dimension and all data of the model are known. Modern approximate solution techniques are summarized under the name *Reinforcement Learning* (RL). The aim of these methods is to find an optimal strategy while simultaneously learn the right model. A popular approach is Q-learning, where the learned action-value function $Q^{(t)}$ directly approximates the Q-function. First we initialize $Q^{(0)}$ arbitrarily. Then we repeat the following steps:

1. Choose an admissible pair (x, a) at random and observe the next state y (or generate $y \sim q(\cdot|x,a)$).

2. Update at (x, a) :

$$Q^{(t+1)}(x,a) := (1-\alpha_t)Q^{(t)}(x,a) + \alpha_t\Big(r(x,a) + \beta \sup_{a' \in D(y)} Q^{(t)}(y,a').\Big)$$

where the learning rates (α_t) have to be chosen appropriately.

Under mild assumptions this method is known to converge to the Q-function. Other methods parametrize the class of policies and thus, the value function and estimate the optimal parameters. Though not being optimal, in this situation it is more convenient to work with randomized policies. In order to find the best parameters in this setting, often the gradient is computed and parameters are updated by a gradient ascent rule. For computational issues consult among others with Sutton and Barto (2018), Powell (2022), Hambly et al. (2023).

As discussed in the previous section this criterion does not account for deviations around the mean or in other words the risk of the decision maker. Thus, in what follows we consider risk-sensitive optimization criteria.

4 Markov decision processes with recursive risk-sensitive preferences

Measuring risk in a stochastic dynamic process is much more complicated than in a single-step situation. For instance, if the decision maker is pessimistic, he may assume that everything that can go wrong will go wrong. Then, he tries to minimize the losses under this assumption. This leads to minmax optimization and is sometimes useful, but, most often, the resulting policies are overly cautious. Involving risk in Markov decision processes is generally difficult. First of all, optimizing many risk sensitive objectives is often NP-hard and computationally intractable. For instance, Mannor and Tsitsiklis (2011) illustrate this fact for mean-variance optimization. Moreover, they also show that the payoff criteria based on expectation and variance can lead to counter-intuitive policies. Another example is provided in Moldovan and Abbeel (2012). Therefore, in our approach we work with the Optimized Certainty Equivalents that enjoys useful properties given in Sect. 2. These properties allow to incorporate risk into a decision process and the problems can be solved efficiently by the dynamic programming techniques. General principles for the specification of utility functions to a potential decision maker are given in Luenberger (2014). Basically, risk might be taken into account at every stage and then the payoff is aggregated or risk might be applied to the aggregated discounted reward. Whereas the second approach is easy to apply also to continuous-time decision problems, the first one is more challenging and there are only a few approaches like the stochastic differential utility, introduced in Duffie and Epstein (1992) to aggregate risk in continuous-time. For further discussions see the end of this section.

In what follows we concentrate on the underlying controlled stochastic dynamic process to be Markovian (like in the previous section) and that the Optimized Certainty Equivalents are applied recursively. This setting guarantees that the optimality principle holds and optimal policies are stationary.

For $k \in \mathbb{N}_0$ let

$$B(H_k) := \{v : H_k \to \mathbb{R}_+ : v \text{ is measurable, bounded}\}$$

be equipped with the supremum norm $\|\cdot\|$. Let $\pi = (\pi_k)_{k\in\mathbb{N}_0} \in \Pi$ be an arbitrary policy. For $v_{k+1} \in B(H_{k+1})$ and $h_k \in H_k$ we define a conditional Optimized Certainty Equivalent

$$\begin{aligned} &S_u^{(x_k,\pi_k(h_k))}\big(v_{k+1}(h_k, \pi_k(h_k), X_{k+1})\big) \\ &\quad := \sup_{\eta\in\mathbb{R}} \Big\{\eta + \int u(v_{k+1}(h_k, \pi_k(h_k), y) - \eta) q(dy|x_k, \pi_k(h_k))\Big\} \end{aligned}$$

where the random variable X_{k+1} has the distribution $q(\cdot|x_k, \pi_k(h_k))$. Then we define the operator L_{π_k} as follows:

$$\begin{aligned} (L_{\pi_k} v_{k+1})(h_k) &= L_{\pi_k} v_{k+1}(h_k) \\ &:= r(x_k, \pi_k(h_k)) + \beta S_u^{(x_k,\pi_k(h_k))}\big(v_{k+1}(h_k, \pi_k(h_k), X_{k+1})\big) \end{aligned}$$

where $\beta \in [0, 1)$ is a discount factor. The operator L_{π_k} is monotone by (P1), i.e.

$$v_{k+1} \le w_{k+1} \Rightarrow L_{\pi_k} v_{k+1} \le L_{\pi_k} w_{k+1}$$

for $v_{k+1}, w_{k+1} \in B(H_{k+1})$. By (P1) and (P4) it holds

$$0 \le L_{\pi_k} v_{k+1}(h_k) \le d + \beta\|v_{k+1}\| \quad \text{for any} \quad h_k \in H_k \text{ and } k \in \mathbb{N}_0. \tag{8}$$

Let now $N \in \mathbb{N}$. For the N-stage decision model we apply these operators recursively. Thus, for an initial state $x \in E$, the total discounted recursive risk-sensitive reward under policy π is given by

$$J_N(x, \pi) = (L_{\pi_0} \circ \ldots \circ L_{\pi_{N-1}})\mathbf{0}(x)$$

where $\mathbf{0}$ is the function $\mathbf{0}(h_k) \equiv 0$ for all $h_k \in H_k, k \in \mathbb{N}_0$. For $N = 2$ this equation reads

$$\begin{aligned} J_2(x, \pi) &= (L_{\pi_0} \circ L_{\pi_1})\mathbf{0}(x) = L_{\pi_0}(L_{\pi_1}\mathbf{0})(x) \\ &= r(x, \pi_0(x)) + \beta S_u^{(x,\pi_0(x))}\big(r(X_1, \pi_1(x, \pi_0(x), X_1))\big). \end{aligned}$$

Aggregation over time is still additive in this approach. By our assumptions and (P1), the sequence $(J_N(x, \pi))_{N\in\mathbb{N}}$ is non-decreasing and bounded from below by 0 for all $x \in E$ and $\pi \in \Pi$. Moreover, by (8) we obtain

$$J_N(x, \pi) \le \frac{d}{1-\beta}, \quad x \in E,\ \pi \in \Pi,\ N \in \mathbb{N}.$$

Hence the limit $\lim_{N\to\infty} J_N(x, \pi)$ exists for every $x \in E$ and $\pi \in \Pi$.

Problem 1 *For an initial wealth $x \in E$ and a policy $\pi \in \Pi$ we define the total discounted recursive risk-sensitive reward by*

$$J(x,\pi) := \lim_{N\to\infty} J_N(x,\pi).$$

The aim of the decision maker is to find the maximal value, i.e.

$$J^*(x) := \sup_{\pi\in\Pi} J(x,\pi), \quad x \in E$$

and a policy π^* such that $J(x,\pi^*) = J^*(x)$, $x \in E$.

In order to solve the problem we use dynamic programming. We need essentially the same assumptions as in the risk-neutral case.

A proof of the following theorem can be found in the appendix.

Theorem 2 *Assume (W) or (S). Then*

a) *There exist a unique function $V \in U(E)$ in case (W) holds and $V \in B(E)$ in case (S) holds and a decision rule $f^* \in F$ such that for all $x \in E$:*

$$\begin{aligned} V(x) &= \sup_{a\in D(x)} \left\{ r(x,a) + \beta S_u^{(x,a)}(V(X_1)) \right\} \\ &= r(x, f^*(x)) + \beta S_u^{(x,f^*(x))}(V(X_1)) \end{aligned} \tag{9}$$

where $S_u^{(x,a)}$ indicates that X_1 has the distribution $q(\cdot|x,a)$.

b) *Moreover, $V(x) = J^*(x) = J(x, f^*)$ for all $x \in E$, i.e. $f^* \in F$ is an optimal stationary policy.*

If u is an exponential utility then we obtain in the previous case that S_u is the entropic risk measure (Example 1 a)) and the optimality equation (9) reduces to (see Asienkiewicz and Jaśkiewicz 2017)

$$V(x) = \sup_{a\in D(x)} \left\{ r(x,a) - \frac{\beta}{\gamma} \ln \left\{ \int \exp(-\gamma V(y)) q(dy|x,a) \right\} \right\} \tag{10}$$

for $x \in E$. The expression in brackets on the r.h.s is also referred to as *risk-sensitive Koopmans operator* (see Miao 2020; Sargent and Stachurski 2023). By applying the exponential function on both sides, the equation for $\gamma < 0$ can also be written as

$$\tilde{V}(x) = \sup_{a\in D(x)} \left\{ e^{-\gamma r(x,a)} \left(\int \tilde{V}(y) q(dy|x,a) \right)^{\beta} \right\} \tag{11}$$

with $\tilde{V}(x) = e^{-\gamma V(x)}$ which yields a multiplicative Bellman equation.

A discounted recursive entropic cost linear quadratic Gaussian regulator problem with the infinite time horizon has been treated in Hansen and Sargent (1995). Conditional consistency of the recursive entropic risk measure is discussed in Dowson et al.

(2020). An efficient learning algorithm for recursive Optimized Certainty Equivalents based on the value iteration and upper confidence bounds can be found in Xu et al. (2023); Fei et al. (2021) where the latter concentrates on the entropic risk measure.

CVaR optimization (which is according to Example 1 b) another special case) for a finite time horizon applied at the terminal wealth has been considered in Rudloff et al. (2014); Pflug and Pichler (2016) and for the infinite time horizon in Uğurlu (2018). The authors also discuss time-consistency issues of optimal policies. Shapiro et al. (2013) consider risk averse approaches (in terms of a weighted criterion of expectation and CVaR) to multistage (linear) stochastic programming problems based on the Stochastic Dual Dynamic Programming method. For further computational approaches see Kozmík and Morton (2015). The recursive CVaR is very popular for applications (see Sect. 7).

Some papers have studied the more general class of convex risk measure for a nested application to stochastic dynamic decision problems. For example Shen et al. (2013, 2014); Chu and Zhang (2014); Bäuerle and Glauner (2022b) consider the infinite time horizon, unbounded cost functions and establish optimality equations and existence of optimal policies. Martyr et al. (2022) consider an iterated $\mathbb{G}$-expectation for non-Markovian optimal switching problems. In Dowson et al. (2022) the problem is tackled as a multistage stochastic program. Algorithms based on stochastic dual dynamic programming and the special role of the entropic risk measure in this class are discussed in Shapiro (2021); Dupačová and Kozmík (2015). Philpott et al. (2013) use inner, outer approximations based on dynamic programming. Further algorithms can be found in Le Tallec (2007); Tamar et al. (2016); Guigues (2016); Huang et al. (2021). Algorithms for a finite time horizon and convex risk measures based on reinforcement learning are studied in Coache and Jaimungal (2023).

There are further recursive risk-sensitive preferences in the literature which are not covered by our model. Kreps and Porteus (1978) and Epstein and Zin (1989) propose an alternative specification of lifetime value that separates and independently parametrizes temporal elasticity of substitution and risk aversion. To be more precise Kreps and Porteus (1978) consider finite time horizon recursive preferences with the conditional Certainty Equivalent[3] defined with $u(x) = x^{1-\gamma}$, $\gamma > 0$ and $\gamma \neq 1$, see (1). Here the parameter γ is responsible for the level of relative risk aversion. Epstein and Zin (1989) generalize their approach to the infinite time horizon and suggest the following form of aggregation:

$$v_n(x) := \left((1-\beta)(r(x, f_n(x)))^{1-\rho} + \beta \left(\int (v_{n+1}(y))^{1-\gamma} q(dy|x, f_n(x)) \right)^{\frac{1-\rho}{1-\gamma}} \right)^{\frac{1}{1-\rho}}.$$

The function v_n denotes the future payoff from period $n \in \mathbb{N}_0$ onwards when the process is governed by a Markovian policy $(f_n) \in \Pi^M$. Moreover, we assume that $\rho > 0$ and $\rho \neq 1$. The value $1/\rho$ represents a *Constant Elasticity of Intertemporal Substitution* (CES). Therefore, the Epstein-Zin aggregator (named from their authors)

[3] We mean here (like in the case of a conditional OCE) a Certainty Equivalent that maps a random variable that is measurable with the next period's information into a random variable that is measurable with respect to the current period's information.

is also called a CES time aggregator. Epstein and Zin (1989) obtain a remarkable result for the existence of recursive utilities across the broad set of parameters γ and ρ. Their results have been further strengthened by Ozaki and Streufert (1996) who provide an extensive analysis of existence and uniqueness of recursive utilities by introducing the notion of biconvergence. This concept requires that returns can be sufficiently discounted from above and sufficiently discounted from below. Moreover, their results are useful for studying dynamic programming with non-additive stochastic objectives in a pretty general setting. The Epstein-Zin time aggregator has also been examined by Weil (1993) but with the conditional Certainty Equivalent defined by an exponential utility function. The function v_n is there given as follows

$$v_n(x) := \left((1-\beta)(r(x, f_n(x)))^{1-\rho} + \beta \left(-\frac{1}{\gamma} \ln \int \exp(-\gamma v_{n+1}(y)) q(dy|x, f_n(x)) \right)^{1-\rho} \right)^{\frac{1}{1-\rho}}.$$

The aforementioned recursive preferences are very popular among economists (see for instance Sargent and Stachurski 2023; Miao 2020 and references cited therein) who put a lot of criticism on the standard expected discounted utility. To learn more on this subject the reader is referred to the notes following Chapter 7 in Sargent and Stachurski (2023). It is worthy to mention that the CES time aggregator and different conditional Certainty Equivalents have been also exploited within dynamic programming framework by a number of authors, see the references in Ren and Stachurski (2018), Chapter 8 in Sargent and Stachurski (2023). Finally, Epstein-Zin preferences in specific parametrizations applied as discrete-time recursive utilities indeed converge to a continuous-time analogue called continuous-time differential utility, see Kraft et al. (2013).

Marinacci and Montrucchio (2010) propose a new class of Thompson aggregators and study a class of quasi-arithmetic Certainty Equivalent operators that generalize those of Kreps and Porteus (1978). Based on specific properties of such operators and the time aggregator they provide a comprehensive analysis of existence, uniqueness and global attractivity of a continuation value process. Particularly, they make use of monotonicity and concavity of the Thompson aggregator and subhomogeneity of the quasi-arithmetic operator. These facts allow them to define a contraction within the Thompson metric.

Bloise and Vailakis (2018) develop an approach to convex programs for bounded recursive utilities. Their technique relies upon the theory of monotone concave operators. An extension is given in Bloise et al. (2021). Iwamoto (1999), on the other hand, treats optimization problems with nested recursive utilities given by applying appropriate functions. A dynamic programming approach is used to solve the problems.

Further extensions include Feinstein and Rudloff (2017) and Schlosser (2020). In the latter paper a multi-valued dynamic programming approach is considered that allows to control the moments of the distributions of future rewards. The former paper is devoted to the development of set-valued risk measures and the recursive algorithms for a dynamic setting.

5 Markov decision processes with risk-sensitive discounted reward

Instead of applying the Optimized Certainty Equivalent recursively one can also apply it to the discounted sum of the rewards. Within such a framework the optimal policies need not be time-consistent. We say that a multiperiod stochastic decision problem is time-consistent, if resolving the problem at later stages (i.e., after observing some random outcomes), the original solutions remain optimal for the later stages. For a recent survey of different approaches to dynamic decision problems with risk measures and their connection to time-consistency, see Homem-de-Mello and Pagnoncelli (2016). We only mention here a stream of references Kreps (1977a, b), Iwamoto (2004), Pflug and Ruszczyński (2005), Pflug (2006), Ruszczyński (2010), Osogami (2011), Shapiro (2012), Philpott et al. (2013) that contributed to this issue among others. Below we provide a simple example that illustrates the problem of time-consistency in the approach taken in this section.

We use the same MDP model as in the previous section. For a fixed history $\omega = (x_0, a_0, x_1, a_1, \ldots) \in H_\infty$ let us define the sum of the discounted rewards by

$$R_\beta^\infty(\omega) := \sum_{k=0}^{\infty} \beta^k r(x_k, a_k)$$

where we always assume that the initial state $x_0 = x$. We also put

$$S_u^\pi(R_\beta^\infty) = \sup_{\eta \in \mathbb{R}} \left\{ \eta + \int_{H_\infty} u(R_\beta^\infty(\omega) - \eta) \mathbb{P}_x^\pi(d\omega) \right\}, \tag{12}$$

where with a little abuse of notation R_β^∞ in (12) is now understood as a random variable on $(\Omega, \mathcal{F})$ with the distribution $\mathbb{P}_x^\pi$ supported on H_∞. In other words, S_u^π indicates that the distribution of R_β^∞ is $\mathbb{P}_x^\pi$. Then we consider the following problem.

Problem 2 *For initial wealth $x \in E$ and policy $\pi \in \Pi$ we define the total discounted risk-sensitive reward by*

$$J(x, \pi) := S_u^\pi(R_\beta^\infty).$$

The aim of the decision maker is to find the maximal value, i.e.

$$J_\infty(x) = \sup_{\pi \in \Pi} S_u^\pi(R_\beta^\infty), \quad x \in E$$

and a policy $\pi^* \in \Pi$ such that $J(x, \pi^*) = J_\infty(x)$, $x \in E$.

A comparison between the obtained values when a coherent risk measure is applied outside or recursively (without control problem), can be found in Iancu et al. (2015). Note that in case of no discounting ($\beta = 1$) Problem 1 and Problem 2 are equivalent. This follows from (P2) and (P4). However, discounting ensures that the value of the problem is finite since we have bounded rewards. Without discounting it depends on

the distribution of $(X_k, A_k)_k$ whether the expectations are finite. The motivation or interpretation of applying the risk measure outside is somewhat easier than for the recursive application of the risk measure. It can be deduced in particular from the different representations of S_u in Example 1.

In order to solve Problem 2 note that by definition of S_u^π

$$\begin{aligned}\sup_{\pi\in\Pi} S_u^\pi(R_\beta^\infty) &= \sup_{\pi\in\Pi}\sup_{\eta\in\mathbb{R}}\left\{\eta + \mathbb{E}_x^\pi[u(R_\beta^\infty - \eta)]\right\} \\ &= \sup_{\eta\in\mathbb{R}}\left\{\eta + \sup_{\pi\in\Pi}\mathbb{E}_x^\pi[u(R_\beta^\infty - \eta)]\right\}.\end{aligned} \tag{13}$$

Thus, we essentially have to solve $\sup_{\pi\in\Pi}\mathbb{E}^\pi[u(R_\beta^\infty - \eta)]$ first. The challenge here is that there is no obvious optimality equation for solving the problem. A way to work around this is to enlarge the state space. This has been done in Bäuerle and Rieder (2014). More precisely, it is helpful to introduce a new MDP on an extended state space $\widetilde{E} := E \times [-\eta, \infty) \times [0, 1]$. Decision rules f are now measurable mappings from $\widetilde{E}$ to A respecting $f(x, y, z) \in D(x)$ for every $(x, y, z) \in \widetilde{E}$. Denote this set of decision rules by $\widetilde{F}$. Policies are defined in an obvious way and with a little abuse of notation denote the set of all policies in this new MDP by Π. For any policy $\pi \in \Pi$ let

$$\begin{aligned}V_\infty^\pi(x, y, z) &:= \mathbb{E}_x^\pi[u(zR_\beta^\infty + y)], \\ V_\infty(x, y, z) &:= \sup_{\pi\in\Pi} V_\infty^\pi(x, y, z)\end{aligned} \tag{14}$$

be the value functions on an extended state space. Thus, we are looking for $V_\infty(x, -\eta, 1)$ which is the value of the inner optimization problem in (13). Let us denote $U(\widetilde{E})$ to be the set of all upper semicontinuous functions v with $v(x, \cdot, \cdot)$ is continuous and increasing in both variables for all x, and $v(x, y, z) \geq u(y)$. Moreover, denote

$$\overline{b}(y, z) := u(zd/(1-\beta) + y), \quad \underline{b}(y, z) := u(z\underline{d}/(1-\beta) + y)$$

where $\underline{d}$ is a lower bound for r (possibly zero). The next theorem summarizes the solution.

Theorem 3 *Assume (W). Then*

a) *There exist a unique function $V \in U(\widetilde{E})$ with $\underline{b} \leq V \leq \overline{b}$ and a decision rule $\widetilde{f}^* \in \widetilde{F}$ such that for all $(x, y, z) \in \widetilde{E}$:*

$$\begin{aligned}V(x, y, z) &= \sup_{a\in D(x)}\left\{\int V(x', zr(x, a) + y, z\beta)q(dx'|x, a)\right\} \\ &= \int V(x', zr(x, \widetilde{f}^*(x, y, z)) + y, z\beta)q(dx'|x, \widetilde{f}^*(x, y, z)).\end{aligned}$$

Moreover, $V(x, y, z) = V_\infty(x, y, z)$ for every $(x, y, z) \in \widetilde{E}$.

b) *There exist an optimal* η^* *in* (13) *and a policy* $\pi^* = (g_0^*, g_1^*, \ldots)$ *with*

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

Moreover, π^* *is an optimal policy for Problem 2.*

If we denote the operator $T : U(\tilde{E}) \to U(\tilde{E})$ by

$$Tv(x, y, z) := \sup_{a \in D(x)} \Big\{ \int v(x', zr(x, a) + y, z\beta) q(dx'|x, a) \Big\},$$

then it can also be shown that $T^n \underline{b} \uparrow V_\infty$ and $T^n \overline{b} \downarrow V_\infty$ for $n \to \infty$. This implies that value iteration works here and yields numerical bounds on the value function. In Bäuerle and Rieder (2014) it has also been shown that the policy improvement converges.

If u is an exponential utility we obtain in the previous case that S_u^π is related to the entropic risk measure (Example 1 a)). Here we can drop the component y and obtain $J_\infty(x) = V_\infty(x, 1)$ where $V_\infty(x, z) = \sup_{\pi \in \Pi} S_u^\pi(zR_\beta^\infty)$ satisfies in this case

$$V_\infty(x, z) = \sup_{a \in D(x)} \Big\{ zr(x, a) - \frac{1}{\gamma} \ln \int \exp(-\gamma V_\infty(x', z\beta)) q(dx'|x, a) \Big\}.$$

Note here the difference to the optimality equation given in (10) where we use the nested application of the entropic risk measure. In case $\beta = 1$ the value function V_∞ does not depend on z and both equations coincide.

Next we give a simple example from Jaquette (1976) to show the difference in optimal policies within the aforementioned frameworks.

Example 2 Let us consider an MDP model with $E = \{1, 2, 3\}$, $A = \{a, b_1, b_2\}$. The decision maker has only a choice in state $x = 1$, namely $D(1) = \{b_1, b_2\}$. In addition, $D(2) = D(3) = \{a\}$. The transition probabilities are:

$$q(2|1, b_1) = 1 - q(3|1, b_1) = 0.5, \quad q(2|1, b_2) = 1 - q(3|1, b_2) = 0.9.$$

From state 2 and from state 3 the process always jumps to state 1 with probability 1. The rewards are as follows:

$$r(1, b_1) = 0, \quad r(1, b_2) = 1, \quad r(2, a) = 0, \quad r(3, a) = 8.$$

Obviously, there are two stationary strategies f and g, i.e. $f(1) = b_1$, $g(1) = b_2$ and $f(2) = g(2) = f(3) = g(3) = a$. Assume that $\beta = 1/2$ and the initial state is $x_0 \equiv 1$. Then, the decision maker essentially chooses between two independent gambles every other period. The first gamble, call it X_f, gives the payoff 0 or 4 with equal probabilities whilst the second gamble, call it X_g, yields the reward 1 with

probability 0.9 or 5 with probability 0.1. Since $\mathbb{E}X_f = 2 > \mathbb{E}X_g = 1.4$, the risk-neutral decision maker prefers a stationary policy f to g. Hence, the maximal expected discounted reward is equal to

$$J_{1/2}(1) = \sum_{n=0}^{\infty} \left(\frac{1}{2}\right)^{2n} \mathbb{E}X_f = 8/3 \approx 2.6666.$$

Let us suppose that the decision maker uses the Optimized Certainty Equivalent defined in (2) with $\gamma = 1$. Consider first Problem 1. Then, equation (9) in Theorem 2 takes the following form

$$V(1) = \max\left\{-\frac{1}{2}\ln\left(\frac{1}{2}e^{-V(2)} + \frac{1}{2}e^{-V(3)}\right), 1 - \frac{1}{2}\ln\left(\frac{9}{10}e^{-V(2)} + \frac{1}{10}e^{-V(3)}\right)\right\}$$

and

$$V(2) = \frac{V(1)}{2}, \quad V(3) = 8 + \frac{V(1)}{2}.$$

Then, g is an optimal stationary policy and the maximal reward is

$$V(1) = \frac{4}{3}\left(1 + \ln\left(\sqrt{\frac{10}{9 + e^{-8}}}\right)\right) \approx 1.4035.$$

Now let us turn to Problem 2. In our case the aim is to maximize over the set of all policies $\pi \in \Pi$ the functional

$$J(1, \pi) = -\ln \mathbb{E}_1^{\pi} e^{-\sum_{k=0}^{\infty}(1/2)^k r(X_k, A_k)}.$$

This is equivalent to minimize the expression $\bar{J}(1, \pi) = \mathbb{E}_1^{\pi} e^{-\sum_{k=0}^{\infty}(1/2)^k r(X_k, A_k)}$ over the set of all $\pi \in \Pi$. Since the decision maker chooses in each period between two independent gambles X_f and X_g, then

$$\bar{J}(1, \pi) = \mathbb{E}_1^{\pi} \exp\left\{-\sum_{n=0}^{\infty}\left(\frac{1}{2}\right)^{2n} X_{2n}\right\},$$

where $X_0, X_2, \ldots$ are independent random variables with the distribution as X_f or X_g, depending whether the policy $\pi = (\pi_k)$ indicates to use f or g in period $k = 0, 2, 4, \ldots$. Clearly, $\pi_k \equiv a$ for $k = 1, 3, 5, \ldots$. Therefore,

$$\bar{J}(1, \pi) = \prod_{n=0}^{\infty} \mathbb{E}e^{-(1/2)^{2n} X_{2n}}.$$

Observe that

$$\mathbb{E}e^{-sX_f} > \mathbb{E}e^{-sX_g} \quad \Longleftrightarrow \quad \frac{1}{2} + \frac{1}{2}e^{-4s} > \frac{9}{10}e^{-s} + \frac{1}{10}e^{-5s}.$$

This holds for $s > 0.455904$. Hence, for the decision maker g is better than f in periods $2n$ for which $(1/2)^{2n} > 0.455904$. This is equivalent to $2n < 1.1332$. Summing up, the optimal policy is $(g, f, f, f \ldots)$. Obviously, the policy is not stationary and it is not time-consistent [4]. However, this policy is ultimately stationary, i.e., there is a period such that from this period onwards the policy is stationary. In fact, Jaquette (1976) proves that an MDP with a finite state space and the entropic risk measure must be ultimately stationary. This need not to be true for MDPs with an infinite state space. For other examples illustrating the lack of stationarity and time-consistency the reader is referred to Brau-Rojas et al. (1998).

The first studies of this entropic setting are due to Howard and Matheson (1972) and Jaquette (1976). Linear-quadratic problems with a finite time horizon and the entropic risk measure are considered in Jacobson (1973), Whittle (1981). A more general approach can be found in Chung and Sobel (1987) where fixed point theorems for the whole distribution of the infinite time horizon discounted reward in a finite MDP are considered. In Collins and McNamara (1998) the authors deal with a finite time horizon problem where they maximize a strictly concave functional of the distribution of the terminal state. Coraluppi and Marcus (1999) connect the problem with the entropic risk measure to a minimax payoff criterion for finite state MDPs. A turnpike theorem for a risk sensitive MDP model with stopping is shown in Denardo and Rothblum (2006). Though Di Masi and Stettner (1999) consider the average reward criterion, they also solve as a by-product the infinite time horizon discounted model with Borel state and action spaces.

Numerical methods for the MDP with the entropic risk measure and finite and infinite time horizons are given in Hau et al. (2023). A finite time horizon non-discounted MDP with Borel state and action spaces and with entropic risk measure is considered in Chapman and Smith (2021). General Certainty Equivalents for MDPs with Borel state and action spaces and finite and infinite time horizons are treated in Bäuerle and Rieder (2014). Partially observable MDPs with the entropic risk measures are examined in James et al. (1994), Fernández-Gaucherand and Marcus (1997), Bäuerle and Rieder (2015), Bäuerle and Rieder (2017).

The special case of optimizing the CVaR of R_β^∞ with bounded rewards has been considered in Bäuerle and Ott (2011). A numerical algorithm and the connection to robust optimization problems is discussed in Chow et al. (2015), Ding and Feinberg (2022). Unbounded cost problems with CVaR are treated in Uğurlu (2017). In Chapman et al. (2023) the authors minimize the CVaR of a maximum random cost over a finite time horizon. Kadota et al. (2006) maximize the expected utility of the total discounted reward subject to multiple expected utility constraints.

[4] To be more precise, we have no time-consistency within the class of policies where decisions are only based on the current wealth. However, when we consult Theorem 3, we see that there is some stationarity of the optimal policy on the extended state space.

6 Markov decision processes with other risk-sensitive payoff criteria

In this section we focus on other payoff criteria than those considered in Sects 4 and 5. We start with average risk-sensitive payoff criteria when a controller is equipped with a constant Arrow-Pratt's risk coefficient, i.e. she evaluates her future income using an exponential utility function. However, sometimes instead of a reward r in the MDP we shall study a cost $c : D \to \mathbb{R}_+$. This is because the papers published so far with this criterion mainly deal with a minimization problem and moreover, the cost minimization is not equivalent to the reward maximization when changing the sign in the cost function as in the risk-neutral case (see also Remark 1).

Problem 3 *For an initial state $x \in E$ and a policy $\pi \in \Pi$ we shall consider the following cost functional:*

$$\mathcal{J}(x,\pi) = \limsup_{n\to\infty} \frac{1}{\gamma n} \ln \mathbb{E}_x^\pi \left[\exp\left(\sum_{k=0}^{n-1} \gamma c(X_k, A_k) \right) \right]$$

for $\gamma > 0$.

Here in order to ensure that the average risk-sensitive cost is well-defined, let us assume as before that c is bounded. The objective is to find the minimal cost

$$\xi(x) := \inf_{\pi\in\Pi} \mathcal{J}(x,\pi).$$

The policy π^* is optimal for the *ergodic risk-sensitive control problem* if

$$\mathcal{J}(x,\pi^*) = \inf_{x\in E} \xi(x), \quad x \in E.$$

Note that then the optimal cost $\xi(x)$ must be independent of x.

The paper of Howard and Matheson (1972)[5] is a pioneering work that deals with the aforementioned problem for MDPs with finite state and action spaces. They assume that the Markov chain is aperiodic and comprises one communicating class under any stationary policy. A Perrron-Frobenius theory of positive matrices allows them to establish a solution to the optimality equation which is of the form

$$\xi_o + h(x) = \min_{a\in D(x)} \left\{ c(x,a) + \frac{1}{\gamma} \int \exp(\gamma h(y)) q(dy|x,a) \right\} \tag{15}$$

for every $x \in E$. Here ξ_o is a real number and $h : E \to \mathbb{R}$ is a given function. If the equation holds, it is possible to prove two points. Firstly, the optimal cost is $\xi(x) = \xi_o/\gamma$ for every $x \in E$. Secondly, the minimizer of the r.h.s. in (15) (if exists), say f_*, defines an optimal stationary policy $f_* \in F$ which means that $\frac{\xi_o}{\gamma} = \mathcal{J}(x, f_*)$ for every $x \in E$. It should be noted that the optimal cost need not be constant (unlike in

[5] In their paper the maximization problem is studied.

the risk-neutral case) if the Markov chain induced by a stationary policy has transient states, consult with Brau-Rojas et al. (1998) for counterexamples. The communication properties of the Markov chains in the analysis of the ergodic risk-sensitive control problem are underlined in Cavazos-Cadena and Hernández-Hernández (2002). Since then the finite state space models have been extensively developed and the Perron-Frobenius theory has been employed, see among others (Sladkỳ 2018, 2008; Rothblum 1984; Cavazos-Cadena and Hernández-Hernández 2009) and references cited therein. In addition, the Perron-Frobenius theory provides a link between risk-sensitive control and the Donsker-Varadhan theory of large deviations. It is known that, under suitable recurrence conditions, the occupation measure of a Markov process satisfies the large deviation principle with rate function given by the convex conjugate of a long run expected rate of an exponential growth function. Such a variational formula for the optimal growth rate of reward in the spirit of the Donsker-Varadhan formula is given in Anantharam and Borkar (2017) where the existence of a Perron-Frobenius eigenvalue and an associated eigenfunction is analyzed by the nonlinear Krein-Rutman theorem. For further results in this direction the reader is referred to Cavazos-Cadena (2018), Arapostathis et al. (2016).

A nice characterization of an optimal cost via a minimization problem in a finite dimensional Euclidean space is given in Cavazos-Cadena and Hernández-Hernández (2005) where the transition law of the Markov chain satisfies a simultaneous Doeblin condition. This result is generalized to an MDP model on a Borel state space in Cavazos-Cadena and Salem-Silva (2010).

The second approach for solving ergodic risk-sensitive control problems is based on an approximation technique. This can be done either by discounted risk-sensitive cost models (Cavazos-Cadena and Fernández-Gaucherand 2000; Cavazos-Cadena and Cruz-Suárez 2017; Huang and Chen 2024) (as in Problem 2) or by certain discounted risk-sensitive dynamic games, see Cavazos-Cadena and Hernández-Hernández (2002, 2011), Hernández-Hernández and Marcus (1999, 1996) for a countable state space case and Di Masi and Stettner (2000, 1999), Jaśkiewicz (2007a, b) for a general state space case. This technique leads via the vanishing discount factor approach to the optimality equation or to the optimality inequality, (when the sign '$=$' in (15) is replaced by '$\geq$'). For instance, the existence of a solution to the optimality inequality is established in Hernández-Hernández and Marcus (1999), Jaśkiewicz (2007a) where the so-called uniform Tauberian theorem was used, see Jaśkiewicz (2007a) and Proposition 1 in Jaśkiewicz and Nowak (2014). The essential ingredient in this approach is the variational formula for the logarithmic moment-generating function (see Fleming and Hernández-Hernández 1997; Dai Pra et al. 1996; Dembo and Zeitouni 1998). It should be noted that in contrast to the risk-neutral case to get a solution to the optimality equation or inequality one needs to assume except ergodicity conditions that the absolute vale of the risk coefficient is sufficiently small. This condition is either imposed explicitly or implicitly, i.e. other conditions in fact enforce this requirement, see Example 1 in Jaśkiewicz (2007a). There is only one exception: the so-called invariant models in which the transition probabilities are independent of the state space, see Jaśkiewicz (2007b). A further discussion on the conditions when the optimality equation or the optimal inequality hold is provided in Cavazos-Cadena (2010).

The ergodic risk-sensitive control problem is also attacked from different sides. Borkar and Meyn (2002) apply an ergodic multiplicative theorem and assume a simple growth condition on the one-stage cost function. They establish the optimality equation for a countable state Markov decision chain. The very recent results for countable state space models have been developed in Biswas and Pradhan (2022); Chen and Wei (2023). Finally, an approximation by uniformly ergodic Markov controlled processes for a general state space model under minorization condition is studied in Di Masi and Stettner (2007). A mutual relationship between the aforementioned works, an extensive discussion of other results and a list of further references are given in the excellent survey of Biswas and Borkar (2023). Finally, we would like to mention that the nested form of an average risk-sensitive reward is discussed in Shen et al. (2013).

Parallel to the theoretical results much effort was put on developing efficient algorithms to solve ergodic risk-sensitive control problem. The value iterations are established in Bielecki et al. (1999b); Cavazos-Cadena and Montes-de Oca (2003) for stationary models and in Cavazos-Cadena and Montes-De-Oca (2005) for non-stationary models. A Q-learning algorithm is proposed in Borkar (2002) and a version of an actor-critic algorithm is considered in Borkar (2001). However, these algorithms do not incorporate any approximation of the value function in order to defeat the curse of dimensionality. Such an approximation in terms of linear combination of a moderate number of basis functions is developed in Basu et al. (2008). The learning scheme iteratively learns coefficients in the linear combination instead of learning the whole value function. The other tools are applied in Arapostathis and Borkar (2021) and Borkar (2017) where equivalent linear and dynamic programs are derived. The former work deals with minimization of the asymptotic growth rate of the cumulative cost whereas the latter one uses a variational representation for asymptotic growth rate of risk-sensitive reward obtained in Anantharam and Borkar (2017). This technique allows to link the average risk-sensitive reward with linear programming without assuming irreducibility of the Markov chain.

Except for the average cost/reward criteria defined with the help of an exponential utility function, there are papers that deal with other average risk-sensitive payoff criteria for which traditional dynamic programming fails. For example in Cavazos-Cadena and Hernández-Hernández (2016) a finite-state irreducible risk-sensitive MDP is considered where the usual exponential utility is replaced by an arbitrary utility function (see also Stettner 2023). The authors prove a connection to the exponential utility criterion. Xia (2020) studies the optimization of the mean-variance combined metric assuming that the finite state Markov decision chain is *ergodic* under any stationary policy. More precisely, for $f \in F$, and an initial state $x \in E$ he defines

$$\mathcal{J}^0(x, f) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_x^f \left[\sum_{k=0}^{n-1} \Big(r(X_k, A_k) - \lambda (r(X_k, A_k) - \mathcal{J}^{av}(x, f))^2 \Big) \right]$$

where $\lambda > 0$ is a trade-off parameter and

$$\mathcal{J}^{av}(x, f) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_x^f \left[\sum_{k=0}^{n-1} r(X_k, A_k) \right].$$

Note that $\mathcal{J}^{av}$ and $\mathcal{J}^0$ are independent of an initial state, because of the ergodicity condition. The objective is to find a stationary policy $f_* \in F$ which maximizes the associated value, i.e. $f_* \in \arg\max_{f\in F} \mathcal{J}^0(x, f)$ for all $x \in E$. Since the optimality equation does not hold, the theory of sensitivity-based optimization is utilized. A version of value iteration algorithm is proposed to find an optimal policy. The theory of sensitivity-based optimization is also applied in Xia and Glynn (2022) to the ergodic Markov decision chains when the CVaR measure is used. In this work Xia and Glynn (2022) consider the cost functions and aim at the cost functional

$$CVaR_\alpha^f = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} CVaR_\alpha^f(c_k)$$

where

$$CVaR_\alpha^f(c_k) = \mathbb{E}_x^f \left[c(X_k, A_k) | c(X_k, A_k) \geq F^{-1}_{c(X_k,A_k)}(\alpha) \right]$$

and $F^{-1}_{c(X_k,A_k)}(\alpha)$ denotes the upper α-quantile of the random variable $c(X_k, A_k)$. The objective is to find an optimal policy, i.e. $f_* \in F$ such that $f_* \in \arg\min_{f\in F} CVaR_\alpha^f$. In particular, the authors establish the local optimality equation and develop a policy iteration procedure that turns out to be more efficient than solving the bilevel MDP problem examined among others for risk-sensitive discounted rewards in Bäuerle and Ott (2011).

At the end let us mention the undiscounted models, i.e models in which the discount factor $\beta = 1$ and the time horizon is infinite. MDPs with non-positive payoffs and an entropic risk measure are studied in Jaśkiewicz (2008). The aim is to show the existence of an optimal stationary policy and the convergence of the value iteration algorithm. In Çavuş and Ruszczyński (2014), on the other hand, a recursive undiscounted cost is defined with the aid of Markov risk measures. For the so-called uniformly risk transient Markov decision process the optimality equation is established and the existence of an optimal stationary policy.

7 Applications

In this section we summarize some applications of the risk-sensitive criterion in dynamic, discrete-time optimization problems. This is not a complete list but simply a biased selection of examples. We start with the *entropic risk measure*.

7.1 Entropic risk criterion

One area of applications where the entropic risk criterion is used is *financial mathematics and economics*. In Bielecki et al. (1999a) the authors consider an investment problem in a financial market with a factor process given by a Markov chain $(X_t)_{t\in\mathbb{N}}$.

The evolution of the wealth is defined by

$$x_{t+1} = x_t \left[e^r + \pi_t \cdot (Z_{t+1} - e^r 1) \right]$$

where r is a fixed interest rate, $(Z_t)_t$ are the relative price vectors, conditionally independent given the states of the Markov chain at time t and $t+1$ and $(\pi_t)_t$ are the proportions of wealth invested in the risky assets. The aim is to maximize

$$\liminf_{T\to\infty} -\frac{2}{\theta}\frac{1}{T} \ln \mathbb{E}_x^\pi \exp\Big(-\frac{\theta}{2} \ln X_T \Big) \tag{16}$$

over all investment strategies. Here, θ in (0,1) is a risk-sensitivity parameter. Under some irreducibility assumptions an optimal investment strategy is stationary and is characterized by the optimality equation given in (15).

Stettner (1999) considers a similar problem which however stems from a discretized version of a continuous Black-Scholes model with several factors. The optimization criterion is again (16). Under a uniform ergodicity condition an optimal investment strategy is characterized via the optimality equation. The cases with (proportional) and without transaction cost are considered. The model with proportional transaction cost and consumption is taken up in Stettner (2005). Finally, the assumptions are further relaxed in Pitera and Stettner (2023) for the same optimization criterion.

Bäuerle and Jaśkiewicz (2018) consider a stochastic optimal growth model with nested entropic risk measures. The model is as follows: an agent obtains the output x_t, which is divided between consumption a_t and investment (saving) $y_t = x_t - a_t$. From consumption a_t the agent receives utility $u(a_t)$. Investment is used for production with input y_t yielding output

$$x_{t+1} = f(y_t, \xi_t)$$

where $(\xi_t)_t$ is a sequence of i.i.d. shocks and f a production function. The criterion of Problem 1 is used for the aggregation of the utilities. The value function and an optimal policy are again characterized via the optimality equation. Properties of the optimal consumption strategy are also shown. The problem is solved explicitly for special utility and production functions. The results are extended in Goswami et al. (2022) to include regime switches.

Other applications in economics touch the problem of precautionary savings, which is one of the most studied issues in the theory of choice under uncertainty. For example, Luo and Young (2010) study the consumption-savings behavior of households who have risk-sensitive preferences and suffer from limited information-processing capacity (rational inattention). The value iteration is as for Problem 1 given by

$$V(x) = \sup_c \Big\{ -\frac{1}{2}(c - \bar{c}) - \frac{\beta}{\gamma} \ln \mathbb{E}[\exp(-\gamma V(X_1))] \Big\}$$

where x is the present value of lifetime resources, c is consumption and $\bar{c}$ denotes a bliss point. The authors solve the model explicitly and show that rational inattention

increases precautionary savings by interacting with income uncertainty and risk sensitivity. They show that the model displays a wide range of observational equivalence properties, implying that consumption and savings data cannot distinguish between risk sensitivity, robustness, or the discount factor, in any combination. Bommier and Le Grand (2019), on the other hand, examine non-stationary models of precautionary savings with recursive risk-sensitive preferences (as in Problem 1) of the infinitely-lived agents. Agents are endowed with an exogenous income process $(Z_t)_t$. The value function in period t is given by the equation

$$V_t(x_t, z^t) = \max_{a_t \in \mathbb{R}} \left\{ \tilde{u}(a_t) - \frac{\beta}{\gamma} \ln \mathbb{E}_t[\exp(-\gamma V_{t+1}(x_{t+1}, z^t, Z_{t+1}))] \right\}$$

where x_t is the wealth at time t, a_t is the consumption at time t and $z^t = (z_0, \ldots, z_t)$ is the realized exogenous income trajectory. Here, $\tilde{u}$ is the one-stage utility of a household. It is assumed that the function $(z_0, \ldots, z_t) \to \mathbb{P}(Z_{t+1} \geq \bar{z} | z_0, \ldots, z_t)$ is non-decreasing. Moreover, $a_t > 0$, $x_t + Z_t - y_t = a_t$, $x_{t+1} = r_{t+1} y_t$, where y_t is investment and r_{t+1} is the deterministic (but time varying) gross interest rate between periods t and $t+1$. Additionally, the constraint $y_t \geq \bar{y}_t(z^t)$ allows to borrow the agent, but no more what she can repay in the worst scenario. The main result announces that the greater risk aversion (the greater absolute values of γ) implies a higher propensity to save at any time. This leads to the conclusion that the greater risk aversion implies greater accumulated wealth or larger precautionary savings. It should be stressed out that this is not the case when other recursive preferences are considered, for instance, the Epstein-Zin-Weil preferences, see Epstein and Zin (1989), Weil (1990) or the preferences developed in Weil (1993). The reader is referred to the numerical results obtained in Bommier and Le Grand (2019) that confirm the aforementioned conclusions.

It is worth mentioning that Pareto optimal consumption allocations is studied by Anderson (2005), who also assumes that the agents have recursive risk-sensitive preferences defined by an exponential utility function.

Nested entropic risk measures are used in actuarial theory as well. In this matter the reader is referred to the works of Bäuerle and Jaśkiewicz (2015, 2017). In the latter paper, within the recursive preference framework they determine the optimal dividend strategy for an insurance company and derive a policy improvement algorithm.

The next prominent applications can be found in the *operations research* area. The paper of Bouakiz and Sobel (1992) is one of the first paper that uses the exponential utility function to the multiperiod news vendor inventory model. The authors minimize the risk-sensitive discounted cost, i.e. as in Problem 2. It is shown that the base-stock policy is optimal and depends on the length of a time horizon, discount factor and risk parameter. For the infinite time horizon an optimal policy is ultimately stationary. Their considerations are extended to models with dependent demands in Choi and Ruszczyński (2011) where an asymptotic behavior of the solution when the degree of risk aversion coefficient converges to zero or infinity is analyzed. Another interesting issue from the area of *revenue management* can be found in Barz and Waldmann (2007). The approach is explained in the setting of optimal airline ticket booking where the airline has to decide whether or not to accept a request for a certain fare given

the remaining capacity. The target function is the one from Problem 2. The optimal strategy is computed and compared to the risk-neutral setting. Further applications to revenue management with different risk-averse target functions can be found in Schlosser (2015, 2016). A survey of risk-sensitive and robust revenue management problems the reader may find in Gönsch (2017), where among other issues the capacity control and dynamic pricing are considered. Finally, Denardo et al. (2007) consider the multiarmed bandit problem with an exponential utility and criterion as in Problem 2. They show the optimality of some kind of index policy using analytical arguments.

Applications in *computer science and engineering* are as follows. One of the first papers is Koenig and Simmons (1994). The authors discuss goal reaching problems (e.g. for robots) under risk-sensitive criteria. They obtain the following optimality equation (there is no discounting):

$$V(x) = \inf_a \left\{ \sum_{y \in E \setminus G} q(y|x,a) e^{\gamma c(x,a,y)} V(y) + \sum_{y \in G} q(y|x,a) e^{\gamma (c(x,a,y)+r(y))} \right\}$$

where $G \subset E$ denotes the set of the goal states, $c(x, a, y)$ is the cost of executing action a in state x and proceeding state y and r is the terminal reward function. Solution algorithms, in particular under change of measure are discussed and some block world problems are considered. In Medina et al. (2012), Befekadu et al. (2015) the authors consider a finite time horizon linear-quadratic problem with target function like in Problem 2 with an exponential utility. In Medina et al. (2012) the setting is to optimize a human-robot interaction such that the physically coupled human-robot follows a desired trajectory. Befekadu et al. (2015) study the impact of cyber-attacks in control systems with partial observation. In Guo et al. (2018) the authors consider risk-sensitive scheduling problems of data packets where large inter-delivery times are penalized.

Further, Mazouchi et al. (2022) investigate risk-averse preview-based Q-learning planner for navigation of autonomous vehicles on a multi-lane road. The criterion is that of Problem 2 with an exponential utility function.

7.2 CVaR risk criterion

Another popular optimization criterion is the CVaR.

We start with some examples from *operations research and engineering*. Gönsch et al. (2018) consider dynamic pricing with a risk-averse seller maximizing the CVaR over the selling horizon. The aim is to dynamically adjust the price during the selling horizon in order to sell a fixed capacity of a perishable product where demand is stochastic such that the total expected/risk averse revenue is maximized. As optimization criterion they use the CVaR of the cumulated revenue. More precisely, they consider the setting of Sect. 5 with a finite time horizon and CVaR, i.e.

$$\max_{\pi \in \Pi} CVaR_\alpha \Big(\sum_{k=1}^{N} A_k 1_{[Y_k \geq A_k]} \Big) =: V_N(x)$$

where A_k is the price offered at time k by the firm. The state x is the remaining good and $(Y_k)_k$ are i.i.d. continuous random variables which represent the willingness to pay of a potential customer arriving in period k. The authors use recursive algorithms to solve the problem, based on specific properties of the CVaR given by $V_0(x, \alpha) = 0$ for $x \geq 0$ and

$$V_t(x, \alpha) = \max_a CVaR_\alpha \left\{ 1_{[Y_t \geq a]} \left(a + V_{t-1}(x-1, \alpha z_{t-1,x-1}) \right) + 1_{[Y_t < a]} V_{t-1}(x, \alpha z_{t-1,x}) \right\}$$

where $z_{t-1,x-1}$ and $z_{t-1,x}$ are certain constants arising from CVaR minimization. A nested formulation with CVaR is considered in Schur et al. (2019).

Wozabal and Rameseder (2020) consider multi-stage stochastic programming approaches to optimize the bidding strategy of a virtual power plant operating on the Spanish spot market for electricity. They consider different setups among others a nested CVaR approach.

Maceira et al. (2015) deal with hydrothermal generation planning in Brazil. The aim is to optimize the system operation taking into account the expected value of thermal generation and possible load curtailment costs over a given set of inflow scenarios to the reservoirs in the future. Risk aversion is crucial here to avoid unacceptable amounts of load curtailment in critical inflow scenarios. The authors use nested CVaR and dual stochastic dynamic programming to solve the problem.

The PhD thesis of Ott (2010) treats several problems of surveillance of critical infrastructures treated as stochastic dynamic optimization problems. The author uses CVaR as criterion in the total discounted cost problems and average cost problems.

Jiang and Powell (2016) investigate a dynamic decision problem faced by the manager of an electric vehicle charging station, who aims to satisfy the charging demand of the customer while minimizing cost. Since the total time needed to charge the electric vehicle up to capacity is often less than the amount of time that the customer is away, there are opportunities to exploit electricity spot price variations. The authors formulate this problem as a combination of nested CVaR and expectation over a finite time horizon. They identify structural properties of an optimal policy and propose an approximation algorithm based on regression and polynomial optimization to solve the problem.

Zhang et al. (2016) consider five decompositions of nested CVaR application in multistage stochastic linear programming. They apply the proposed formulations to a water management problem in the area of the southeastern portion of Tucson, AZ to best use the limited water resources available to that region.

Finally, Ahmed et al. (2007) solve a multiperiod inventory model with nested approach of coherent risk measures. For a finite time horizon they prove that the optimal policy has a similar structure as that of the expected value problem. Moreover, an analyis of monotonicity properties of the optimal order quantity with respect to the degree of risk aversion for certain risk measures like CVaR is conducted.

Applications in *financial mathematics and economics* are as follows. Staino and Russo (2020) treat portfolio optimization problems with nested CVaR when asset log returns are stage-wise dependent by a single-factor. Using a cubic spline interpolation the authors numerically solve the problem with a finite time horizon by backward

recursion. A dynamic mean-risk problem, where the risk constraint is given by the CVaR is considered in Bäuerle and Mundt (2009). The financial market is a binomial model which allows for explicit solutions. Since the problem is solved via a Lagrange function, the CVaR appears in the optimization criterion. It is applied to the cumulated gain/loss and the problem is solved by recursion explicitly.

An application in *biology* is given in Bushaj et al. (2022) where the authors apply a mean-CVaR multistage, stochastic mixed-integer programming model to optimize a manager's decisions about the surveillance and control of a non-native forest insect, the emerald ash borer.

As mentioned before, this is just a selection of applications. Further examples can be found in the literature.

Funding Open Access funding enabled and organized by Projekt DEAL.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose

Appendix A Proof of Theorem 2

First we show the statements under assumption (W). Let $v \in U(E)$ and define

$$Lv(x) = \sup_{a \in D(x)} \left\{ r(x,a) + \beta S_u^{(x,a)}(v(X_1)) \right\}$$

where X_1 has distribution $q(\cdot|x,a)$. We first prove that $L : U(E) \to U(E)$. Note that by (P1) and (P4) we get for every $x \in E$

$$Lv(x) = \sup_{a \in D(x)} \left\{ r(x,a) + \beta S_u^{(x,a)}(v(X_1)) \right\} \geq 0 + \sup_{a \in D(x)} \beta S_u^{(x,a)}(0) \geq 0.$$

On the other hand, we have again by (P1) and (P4) that

$$Lv(x) \leq d + \sup_{a \in D(x)} \beta S_u^{(x,a)}(\|v\|) = d + \beta \|v\|.$$

Now we show that Lv is upper semicontinuous. For this purpose we prove for $v \in U(E)$ that

$$(x, a, \eta) \to r(x, a) + \beta\eta + \beta \int u(v(y) - \eta)q(dy|x, a) \tag{A1}$$

is upper semicontinuous. Clearly, $(x, a, \eta) \to r(x, a) + \beta\eta$ is upper semicontinuous. For the second part assume that (x_n, a_n, η_n) is a sequence which converges to (x_0, a_0, η_0) as $n \to \infty$ where $x_n \in E$, $a_n \in D(x_n)$, $\eta_n \in \mathbb{R}$ for $n \in \mathbb{N}_0$. Set $\phi_n(y) := u(v(y) - \eta_n)$ for $n \in \mathbb{N}$. Since u is continuous and non-decreasing, ϕ_n are upper semicontinuous. Making use of the Fatou lemma for weakly convergent measures (see Lemma 3.6 in Balbus et al. 2015) we get that

$$\limsup_{n\to\infty} \int \phi_n(y)q(dy|x_n, a_n) \le \int \phi^*(y)q(dy|x_0, a_0)$$

with $\phi^*(x) = \sup\{\limsup_{n\to\infty} \phi_n(y_n) : y_n \to x\}$. The supremum is taken over all sequences (y_n) converging to x. In our case, for any $y_n \to x$

$$\limsup_{n\to\infty} \phi_n(y_n) = \limsup_{n\to\infty} u(v(y_n) - \eta_n) \le u(v(x) - \eta_0).$$

Hence, $\phi^*(x) = u(v(x) - \eta_0)$. This proves that the function in (A1) is upper semicontinuous.

Next we conclude by Proposition 2.1 in Ben-Tal and Teboulle (2007) that the supremum over all $\eta \in \mathbb{R}$ in the definition of the Optimized Certainty Equivalent can be restricted to the compact set, for example $[0, \|v\|]$. This is the support of the random variable $v(X_1)$. Hence, by Proposition 2.4.3 in Bäuerle and Rieder (2011) the function

$$Lv(x) = \sup_{a\in D(x)} \sup_{\eta\in[0,\|v\|]} \left\{ r(x, a) + \beta\eta + \beta \int u(v(y) - \eta)q(dy|x, a) \right\}.$$

is upper semicontinuous.

Finally we prove that L is contracting. Let $v_1, v_2 \in U(E)$. Then due to (P1) and (P2) and $v_1 \le v_2 + \|v_1 - v_2\|$, we obtain:

$$\begin{aligned} Lv_1(x) - Lv_2(x) &\le \beta \sup_{a\in D(x)} \left(S_u^{(x,a)}(v_1(X_1)) - S_u^{(x,a)}(v_2(X_1)) \right) \\ &\le \beta \sup_{a\in D(x)} \left(S_u^{(x,a)}(\|v_1 - v_2\| + v_2(X_1)) - S_u^{(x,a)}(v_2(X_1)) \right) \\ &= \beta\|v_1 - v_2\|. \end{aligned}$$

Interchanging the roles of v_1 and v_2 yields $\|Lv_1 - Lv_2\| \le \beta\|v_1 - v_2\|$. Finally since $U(E)$ equipped with the supremum norm is complete, the Banach fixed point theorem implies that there exists $V \in U(E)$ such that $V = LV$.

It remains to show that V is the value function. Observe that for all $(x, a) \in D$ we immediately have

$$V(x) \geq r(x, a) + \beta S_u^{(x,a)}(v(X_1)).$$

Let $(\pi_k)_{k\in\mathbb{N}_0} \in \Pi$ be any policy. Then for all $k = 1, \ldots, N$ we obtain $V(x_k) \geq L_{\pi_k} V(h_k)$. Making use of this inequality by iteration we infer that

$$V(x) \geq (L_{\pi_0} \circ \ldots \circ L_{\pi_N})V(x) \geq (L_{\pi_0} \circ \ldots \circ L_{\pi_N})\mathbf{0}(x) = J_{N+1}(x, \pi).$$

Letting $N \to \infty$ implies $V(x) \geq J(x, \pi)$ for all policies $\pi \in \Pi$ which in turn gives

$$V(x) \geq \sup_{\pi\in\Pi} J(x, \pi) \quad \text{for every} \quad x \in E. \tag{A2}$$

For the reverse inequality by Proposition 2.4.3 in Bäuerle and Rieder (2011) it follows that firstly the function

$$(x, a) \to \sup_{\eta\in[0,\|V\|]} \left\{ r(x, a) + \beta\eta + \beta \int u(V(y) - \eta)q(dy|x, a) \right\}$$

is upper semicontinuous and secondly, there exists $f^* \in F$ such that $V = L_{f^*}V$. Thus, again by iteration we have $V = L_{f^*}^{(N)}V$ where $L_{f^*}^{(N)}$ denotes the composition of L_{f^*} with itself N times. Hence, putting $r(x, f^*(x)) = r_{f^*}(x)$ we get

$$\begin{aligned} V(x) &\leq L_{f^*}^{(N-1)}\Big(r_{f^*} + \beta\|V\|\Big)(x) \\ &= L_{f^*}^{(N-2)}\Big(r_{f^*} + \beta S_u^{(\cdot, f^*(\cdot))}(r_{f^*}(X_1)) + \beta^2\|V\|\Big)(x) \\ &\leq \ldots \leq J_N(x, f^*) + \beta^N\|V\|. \end{aligned}$$

Letting $N \to \infty$ yields that $V(x) \leq J(x, f^*)$ for every $x \in E$. This fact and (A2) finish the proof.

Assume now that (S) holds. It suffices to show that $L : B(E) \to B(E)$. Let $v \in B(E)$. Assume that $(a_n, \eta_n) \to (a_0, \eta_0)$ as $n \to \infty$ for $a_n \in D(x)$ and $\eta_n \in \mathbb{R}$. Then, by condition (S) and Proposition 18 on p. 270 in Royden (1988) we have that

$$\int u(v(y) - \eta_n)q(dy|x, a_n) \to \int u(v(y) - \eta_0)q(dy|x, a_0) \quad \text{as } n \to \infty.$$

Hence, the function

$$(a, \eta) \to \left\{ r(x, a) + \beta\eta + \beta \int u(v(y) - \eta)q(dy|x, a) \right\}$$

is upper semicontinuous for each $x \in E$. Again the measurable selection theorem (see Theorem A.2.4 in Bäuerle and Rieder 2011) and the fact that by Proposition 2.1 in

Ben-Tal and Teboulle (2007) the supremum over all $\eta \in \mathbb{R}$ in $S_u^{(x,a)}$ can be replaced by the supremum over the set $[0, \|v\|]$, imply that $Lv \in B(E)$. Now the remaining part proceeds along the same lines with obvious changes, i.e. the fixed point of L is found in $B(E)$. Note that for the proof to hold, the concavity of $S_u^{(x,a)}$ is not necessary (only continuity) whereas we use all other properties (P1), (P2) and (P4).

Appendix B Proof of Theorem 3

The proof of part a) is essentially Theorem 3 in Bäuerle and Rieder (2014). The only difference is that we have a maximization problem here instead of a minimization problem.

For part b) note again that R_β^∞ is bounded and thus the maximization over η in the definition of S_u can be restricted to a compact set by Proposition 2.1 in Ben-Tal and Teboulle (2007). In other words, we have to solve in the second step for large $K > 0$

$$\sup_{\eta \in [-K,K]} \Big\{ \eta + V_\infty(x, -\eta, 1) \Big\}.$$

But from part a) we know that V_∞ is continuous in η which implies the existence of an η^* with

$$\sup_{\eta \in [-K,K]} \Big\{ \eta + V_\infty(x, -\eta, 1) \Big\} = \eta^* + V_\infty(x, -\eta^*, 1)$$

and thus the statement.

References

Ahmed S, Çakmak U, Shapiro A (2007) Coherent risk measures in inventory problems. Eur J Oper Res 182:226–238

Anantharam V, Borkar VS (2017) A variational formula for risk-sensitive reward. SIAM J Control Optim 55(2):961–988

Anderson EW (2005) The dynamics of risk-sensitive allocations. J Econ Theory 125(2):93–150

Arapostathis A, Borkar VS (2021) Linear and dynamic programs for risk-sensitive cost minimization. In: Proceedings of the 60th IEEE conference on decision and control. IEEE, pp 3042–3047

Arapostathis A, Borkar VS, Kumar SK (2016) Risk-sensitive control and an abstract Collatz–Wielandt formula. J Theor Probab 29(4):1458–1484

Arrow KJ (1971) The theory of risk aversion. In: Essays in the theory of risk-bearing. North Holland, pp 90–120

Asienkiewicz H, Jaśkiewicz A (2017) A note on a new class of recursive utilities in Markov decision processes. Applicationes Mathematicae 44:149–161

Balbus Ł, Jaśkiewicz A, Nowak AS (2015) The dynamics of risk-sensitive allocations. J Optim Theory Appl 165:295–315

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

Basu A, Bhattacharyya T, Borkar VS (2008) A learning algorithm for risk-sensitive cost. Math Oper Res 33(4):880–898

Bäuerle N, Glauner A (2022) Distributionally robust Markov decision processes and their connection to risk measures. Math Oper Res 47(3):1757–1780

Bäuerle N, Glauner A (2022) Markov decision processes with recursive risk measures. Eur J Oper Res 296(3):953–966
Bäuerle N, Jaśkiewicz A (2015) Risk-sensitive dividend problems. Eur J Oper Res 242(1):161–171
Bäuerle N, Jaśkiewicz A (2017) Optimal dividend payout model with risk sensitive preferences. Insurance Math Econom 73:82–93
Bäuerle N, Jaśkiewicz A (2018) Stochastic optimal growth model with risk sensitive preferences. J Econ Theory 173:181–200
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, Berlin
Bäuerle N, Rieder U (2014) More risk-sensitive Markov decision processes. Math Oper Res 39(1):105–120
Bäuerle N, Rieder U (2015) Partially observable risk-sensitive stopping problems in discrete time. In: Piunovskiy AB (ed) Modern trends of controlled stochastic processes: theory and Applications, vol II. Luniver Press, pp 12–31
Bäuerle N, Rieder U (2017) Partially observable risk-sensitive Markov decision processes. Math Oper Res 42(4):1180–1196
Befekadu GK, Gupta V, Antsaklis PJ (2015) Risk-sensitive control under Markov modulated denial-of-service (DoS) attack strategies. IEEE Trans Autom Control 60(12):3299–3304
Ben-Tal A, Teboulle M (2007) An old-new concept of convex risk measures: the optimized certainty equivalent. Math Financ 17(3):449–476
Bernoulli D (1954) Exposition of a new theory on the measurement of risk. Econometrica 22:23–36
Bielecki T, Hernández-Hernández D, Pliska SR (1999) Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. Math Methods Oper Res 50:167–188
Bielecki T, Hernandez-Hernandez D, Pliska SR (1999b) Value iteration for controlled Markov chains with risk sensitive cost criterion. In: Proceedings of the 38th IEEE conference on decision and control. IEEE, pp 126–130
Biswas A, Borkar VS (2023) Ergodic risk-sensitive control—a survey. Annu Rev Control 55:118–141
Biswas A, Pradhan S (2022) Ergodic risk-sensitive control of Markov processes on countable state space revisited. ESAIM: Control Optim Cal Variat 28:26
Bloise G, Vailakis Y (2018) Convex dynamic programming with (bounded) recursive utility. J Econ Theory 173:118–141
Bloise G, Le Van C, Vailakis Y (2021) Do not blame Bellman: It is Koopmans' fault. SSRN 3943709
Bommier A, Le Grand F (2019) Risk aversion and precautionary savings in dynamic settings. Manage Sci 65(3):1386–1397
Borkar VS (2001) A sensitivity formula for risk-sensitive cost and the actor-critic algorithm. Syst Control Lett 44(5):339–346
Borkar VS (2002) Q-learning for risk-sensitive control. Math Oper Res 27(2):294–311
Borkar VS (2017) Linear and dynamic programming approaches to degenerate risk-sensitive reward processes. In: 56th Annual IEEE conference on decision and control. IEEE, pp 3714–3718
Borkar VS, Meyn SP (2002) Risk-sensitive optimal control for Markov decision processes with monotone cost. Math Oper Res 27(1):192–209
Bouakiz M, Sobel MJ (1992) Inventory control with an exponential utility criterion. Oper Res 40(3):603–608
Brau-Rojas A, Cavazos-Cadena R, Fernández-Gaucherand E (1998) Controlled Markov chains with risk-sensitive criteria: some (counter) examples. In: Proceedings of the 37th IEEE conference on decision and control. IEEE, pp 1853–1858
Braun DA, Nagengast AJ, Wolpert DM (2011) Risk-sensitivity in sensorimotor control. Front Hum Neurosci 5:1
Bushaj S, Büyüktahtakın İE, Haight RG (2022) Risk-averse multi-stage stochastic optimization for surveillance and operations planning of a forest insect infestation. Eur J Oper Res 299(3):1094–1110
Cavazos-Cadena R (2010) Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains. Math Methods Oper Res 71(1):47–84
Cavazos-Cadena R (2018) Characterization of the optimal risk-sensitive average cost in denumerable Markov decision chains. Math Oper Res 43(3):1025–1050
Cavazos-Cadena R, Cruz-Suárez D (2017) Discounted approximations to the risk-sensitive average cost in finite Markov chains. J Math Anal Appl 450(2):1345–1362

Cavazos-Cadena R, Fernández-Gaucherand E (2000) The vanishing discount approach in Markov chains with risk-sensitive criteria. IEEE Trans Autom Control 45(10):1800–1816
Cavazos-Cadena R, Hernández-Hernández D (2002) Solution to the risk-sensitive average optimality equation in communicating Markov decision chains with finite state space: An alternative approach. Math Methods Oper Res 56:473–479
Cavazos-Cadena R, Hernández-Hernández D (2005) A characterization of the optimal risk-sensitive average cost in finite controlled Markov chains. Ann Appl Probab 15(1A):175–212
Cavazos-Cadena R, Hernández-Hernández D (2009) Necessary and sufficient conditions for a solution to the risk-sensitive Poisson equation on a finite state space. Syst Control Lett 58(4):254–258
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(1):133–146
Cavazos-Cadena R, Hernández-Hernández D (2016) A characterization of the optimal certainty equivalent of the average cost via the Arrow-Pratt sensitivity function. Math Oper Res 41(1):224–235
Cavazos-Cadena R, Montes-De-Oca R (2005) Nonstationary value iteration in controlled Markov chains with risk-sensitive average criterion. J Appl Probab 42(4):905–918
Cavazos-Cadena R, Montes-de Oca R (2003) The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space. Math Oper Res 28(4):752–776
Cavazos-Cadena R, Salem-Silva F (2010) The discounted method and equivalence of average criteria for risk-sensitive Markov decision processes on borel spaces. Appl Math Optim 61(2):167–190
Çavuş O, Ruszczyński A (2014) Risk-averse control of undicounted transient Markov models. SIAM J Control Optim 52(6):3935–3966
Chapman MP, Smith KM (2021) Classical risk-averse control for a finite-horizon Borel model. IEEE Control Syst Lett 6:1525–1530
Chapman MP, Fauß M, Smith KM (2023) On optimizing the conditional value-at-risk of a maximum cost for risk-averse safety analysis. IEEE Trans Autom Control 68(6):3720–3727
Chen X, Wei Q (2023) Risk-sensitive average optimality for discrete-time Markov decision processes. SIAM J Control Optim 61(1):72–104
Choi S, Ruszczyński A (2011) A multi-product risk-averse newsvendor with exponential utility function. Eur J Oper Res 214:78–84
Chow Y, Tamar A, Mannor S, et al (2015) Risk-sensitive and robust decision-making: a CVaR optimization approach. In: Proceedings of the 28th international conference on neural information processing systems, ACMDL, pp 1522–1530
Chu S, Zhang Y (2014) Markov decision processes with iterated coherent risk measures. Int J Control 87(11):2286–2293
Chung KJ, Sobel MJ (1987) Discounted MDP's: distribution functions and exponential utility maximization. SIAM J Control Optim 25(1):49–62
Coache A, Jaimungal S (2023) Reinforcement learning with dynamic convex risk measures. Math Financ. https://doi.org/10.1111/mafi.12388
Collins E, McNamara J (1998) Finite-horizon dynamic optimisation when the terminal reward is a concave functional of the distribution of the final state. Adv Appl Probab 30(1):122–136
Coraluppi SP, Marcus SI (1999) Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. Automatica 35(2):301–309
Dai Pra P, Meneghini L, Runggaldier WJ (1996) Connections between Stochastic control and dynamic games. Math Control Signals Syst 9:303–326
Dembo A, Zeitouni O (1998) Large deviations techniques and applications. Springer, Berlin
Denardo EV, Rothblum UG (2006) A turnpike theorem for a risk-sensitive Markov decision process with stopping. SIAM J Control Optim 45(2):414–431
Denardo EV, Park H, Rothblum UG (2007) Risk-sensitive and risk-neutral multiarmed bandits. Math Oper Res 32(2):374–394
Di Masi GB, Stettner Ł (1999) Risk-sensitive control of discrete-time Markov processes with infinite horizon. SIAM J Control Optim 38(1):61–78
Di Masi GB, Stettner Ł (2000) Infinite horizon risk sensitive control of discrete time Markov processes with small risk. Syst Control Lett 40(1):15–20
Di Masi GB, Stettner Ł (2007) Infinite horizon risk sensitive control of discrete time Markov processes under minorization property. SIAM J Control Optim 46(1):231–252
Ding R, Feinberg EA (2022) Sequential optimization of CVaR. ArXiv preprint arXiv:2211.07288

Dowson O, Morton DP, Pagnoncelli BK (2020) Multistage stochastic programs with the entropic risk measure. Optim Online https://optimization-online.org/?p=16662
Dowson O, Morton DP, Pagnoncelli BK (2022) Incorporating convex risk measures into multistage stochastic programming algorithms. Ann Oper Res. https://doi.org/10.1007/s10479-022-04977-w
Duffie D, Epstein LG (1992) Stochastic differential utility. Econometrica J Econom Soc 1:353–394
Dupačová J, Kozmík V (2015) Structure of risk-averse multistage stochastic programs. OR Spectrum 37:559–582
Epstein LG, Zin SE (1989) Substitution, risk aversion and the temporal behavior of consumption and asset returns: A theoretical framework. Econometrica 57(4):937–969
Fei Y, Yang Z, Chen Y et al (2021) Exponential Bellman equation and improved regret bounds for risk-sensitive reinforcement learning. Adv Neural Inf Process Syst 34:20436–20446
Feinstein Z, Rudloff B (2017) A recursive algorithm for multivariate risk measures and a set-valued Bellman's principle. J Global Optim 68(1):47–69
Fernández-Gaucherand E, Marcus SI (1997) Risk-sensitive optimal control of hidden Markov models: Structural results. IEEE Trans Autom Control 42(10):1418–1422
Filar J, Koos V (1997) Competitive Markov Decision Processes. Springer, Berlin
Fleming WH, Hernández-Hernández D (1997) Risk-sensitive control of finite state machines on an infinite horizon I. SIAM J Control Optim 35(5):1790–1810
Föllmer H, Schied A (2010) Convex and coherent risk measures. Encyclop Quant Financ 1:355–363
Gönsch J (2017) A survey on risk-averse and robust revenue management. Eur J Oper Res 263(2):337–348
Gönsch J, Hassler M, Schur R (2018) Optimizing Conditional Value-at-Risk in dynamic pricing. OR Spectrum 40:711–750
Goswami A, Rana N, Siu TK (2022) Regime switching optimal growth model with risk sensitive preferences. J Math Econ 101:102702
Guigues V (2016) Convergence analysis of sampling-based decomposition methods for risk-averse multistage stochastic convex programs. SIAM J Optim 26(4):2468–2494
Guo X, Singh R, Kumar P et al (2018) A risk-sensitive approach for packet inter-delivery time optimization in networked cyber-physical systems. IEEE/ACM Trans Networking 26(4):1976–1989
Hambly B, Xu R, Yang H (2023) Recent advances in reinforcement learning in finance. Math Financ 33(3):437–503
Hansen LP, Sargent TJ (1995) Discounted linear exponential quadratic Gaussian control. IEEE Trans Autom Control 40(5):968–971
Hau JL, Petrik M, Ghavamzadeh M (2023) Entropic risk optimization in discounted MDPs. In: International conference on artificial intelligence and statistics. PMLR, pp 47–76
Hernández-Hernández D, Marcus SI (1996) Risk sensitive control of Markov processes in countable state space. Syst Control Lett 29(3):147–155 (**Corrigendum in System and Control Letters (1998) 34:105–106**)
Hernández-Hernández D, Marcus SI (1999) Existence of risk-sensitive optimal stationary policies for controlled Markov processes. Appl Math Optim 40:273–285
Hernández-Lerma O, Lasserre JB (1996) Discrete-time Markov control processes, basic optimality criteria. Springer, Berlin
Homem-de-Mello T, Pagnoncelli BK (2016) Risk aversion in multistage stochastic programming: a modeling and algorithmic perspective. Eur J Oper Res 249(1):188–199
Howard RA, Matheson JE (1972) Risk-sensitive Markov decision processes. Manage Sci 18(7):356–369
Huang A, Leqi L, Lipton ZC, et al (2021) On the convergence and optimality of policy gradient for Markov coherent risk. arXiv preprint arXiv:2103.02827
Huang T, Chen J (2024) Markov decision processes under risk sensitivity: a discount vanishing approach. J Math Anal Appl 533(2):128026
Iancu DA, Petrik M, Subramanian D (2015) Tight approximations of dynamic risk measures. Math Oper Res 40(3):655–682
Iwamoto S (1999) Conditional decision processes with recursive function. J Math Anal Appl 230(1):193–210
Iwamoto S (2004) Stochastic optimization of forward recursive functions. J Math Anal Appl 292(1):73–83
Jacobson D (1973) Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. IEEE Trans Autom Control 18(2):124–131
James MR, Baras JS, Elliott RJ (1994) Risk-sensitive control and dynamic games for partially observed discrete-time nonlinear systems. IEEE Trans Autom Control 39(4):780–792

Jaquette SC (1976) A utility criterion for Markov decision processes. Manag Sci 23(1):43–49
Jaśkiewicz A (2007) Average optimality for risk-sensitive control with general state space. Ann Appl Probab 17(2):654–675
Jaśkiewicz A (2007) A note on risk-sensitive control of invariant models. Syst Control Lett 56(11–12):663–668
Jaśkiewicz A (2008) A note on negative dynamic programming for risk-sensitive control. Oper Res Lett 36(5):531–534
Jaśkiewicz A, Nowak AS (2014) Robust Markov control processes. J Math Anal Appl 420(2):1337–1353
Jiang DR, Powell WB (2016) Practicality of nested risk measures for dynamic electric vehicle charging. ArXiv preprint arXiv:1605.02848
Kadota Y, Kurano M, Yasuda M (2006) Discounted Markov decision processes with utility constraints. Comput Math Appl 51(2):279–284
Koenig S, Simmons RG (1994) Risk-sensitive planning with probabilistic decision graphs. In: Principles of knowledge representation and reasoning. Elsevier, pp 363–373
Kozmík V, Morton DP (2015) Evaluating policies in risk-averse multi-stage stochastic programming. Math Program 152:275–300
Kraft H, Seifried FT, Steffensen M (2013) Consumption-portfolio optimization with recursive utility in incomplete markets. Finance Stochast 17:161–196
Kreps DM (1977) Decision problems with expected utility critera, I: upper and lower convergent utility. Math Oper Res 2(1):45–53
Kreps DM (1977) Decision problems with expected utility criteria, II: stationarity. Math Oper Res 2(3):266–274
Kreps DM, Porteus EL (1978) Temporal resolution of uncertainty and dynamic choice theory. Econometrica 46(1):185–200
Le Tallec Y (2007) Robust, risk-sensitive, and data-driven control of Markov decision processes. Phd thesis, Massachusetts Institute of Technology, available at https://dspace.mit.edu/handle/1721.1/38598
Luenberger DG (2014) Investement Science. Oxford University Press, Oxford
Luo Y, Young ER (2010) Risk-sensitive consumption and savings under rational inattention. Am Econ J Macroecon 2(4):281–325
Maceira MEP, Marzano L, Penna DDJ et al (2015) Application of CVaR risk aversion approach in the expansion and operation planning and for setting the spot price in the Brazilian hydrothermal interconnected system. Int J Electr Power Energy Syst 72:126–135
Mannor S, Tsitsiklis J (2011) Mean-variance optimization in Markov decision processes. In: Proceedings of the 28th international conference on machine learning. ICML, pp 177–184
Marinacci M, Montrucchio L (2010) Unique solutions for stochastic recursive utilities. J Econ Theory 145(5):1776–1804
Markowitz HM (1952) Portfolio selection. J Financ 7(1):77–91
Martyr R, Moriarty J, Perninge M (2022) Discrete-time risk-aware optimal switching with non-adapted costs. Adv Appl Probab 54(2):625–655
Mazouchi M, Nageshrao S, Modares H (2022) Automating vehicles by risk-averse preview-based Q-learning algorithm. IFAC-PapersOnLine 55(15):105–110
Medina JR, Lee D, Hirche S (2012) Risk-sensitive optimal feedback control for haptic assistance. In: IEEE international conference on robotics and automation. IEEE, pp 1025–1031
Miao J (2020) Economic Dynamics in Discrete Time. MIT press
Moldovan T, Abbeel P (2012) Risk aversion in Markov decision processes via near-optimal Chernoff bounds. Adv Neural Inf Process Syst 4:3131–3139
Osogami T (2011) Iterated risk measures for risk-sensitive Markov decision processes with discounted cost. In: Proceedings of the 27th conference on uncertainty in artificial intelligence, pp 573–580
Ott J (2010) A Markov decision model for a surveillance application and risk-sensitive Markov decision processes. PhD Thesis, Karlsruhe Institute of Technology. https://publikationen.bibliothek.kit.edu/1000020835
Ozaki H, Streufert PA (1996) Dynamic programming for non-additive stochastic objectives. J Math Econ 25(4):391–442
Pflug GC (2006) A value-of-information approach to measuring risk in multi-period economic activity. J Bank Finance 30(2):695–715
Pflug GC, Pichler A (2016) Time-inconsistent multistage stochastic programs: Martingale bounds. Eur J Oper Res 249(1):155–163

Pflug GC, Ruszczyński (2005) Measuring risk for income streams. Comput Optim Appl 32:161–178
Philpott A, de Matos V, Finardi E (2013) On solving multistage stochastic programs with coherent risk measures. Oper Res 61(4):957–970
Pitera M, Stettner Ł (2023) Discrete-time risk sensitive portfolio optimization with proportional transaction costs. Math Financ 33(4):1287–1313
Piunovskiy AB (2013) Examples in Markov decision processes. Imperial College Press, London
Powell WB (2022) Reinforcement learning and Stochastic optimization: a unified framework for sequential decisions. Wiley, Boca Raton
Pratt JW (1964) Risk aversion in the small and in the large. Econometrica 32:122–136
Puterman ML (2014) Markov decision processes: discrete stochastic dynamic programming. Wiley, Boca Raton
Ren G, Stachurski J (2018) Dynamic programming with recursive preferences: optimality and applications. ArXiv preprint arXiv:1812.05748
Rothblum UG (1984) Multiplicative Markov decision chains. Math Oper Res 9(1):6–24
Royden HL (1988) Real analysis. Prentice Hall, New Jersey
Rudloff B, Street A, Valladão DM (2014) Time consistency and risk averse dynamic decision models: Definition, interpretation and practical consequences. Eur J Oper Res 234(3):743–750
Ruszczyński A (2010) Risk-averse dynamic programming for Markov decision processes. Math Program 125:235–261
Sargent T, Stachurski J (2023) Dynamic Programming, Vol. I: Foundations. https://dp.quantecon.org
Schäl M (1975) Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 32:179–196
Schäl M (1983) Stationary policies in dynamic programming models under compactness assumptions. Math Oper Res 8(3):366–372
Schlosser R (2015) A stochastic dynamic pricing and advertising model under risk aversion. J Revenue Pricing Manag 14:451–468
Schlosser R (2016) Stochastic dynamic multi-product pricing with dynamic advertising and adoption effects. J Revenue Pric Manag 15:153–169
Schlosser R (2020) Risk-sensitive control of Markov decision processes: A moment-based approach with target distributions. Comput Oper Res 123:104997
Schur R, Gönsch J, Hassler M (2019) Time-consistent, risk-averse dynamic pricing. Eur J Oper Res 277(2):587–603
Shapiro A (2012) Minimax and risk averse multistage stochastic programming. Eur J Oper Res 219(3):719–726
Shapiro A (2021) Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. Eur J Oper Res 288(1):1–13
Shapiro A, Tekaya W, da Costa JP et al (2013) Risk neutral and risk averse stochastic dual dynamic programming method. Eur J Oper Res 224(2):375–391
Shen Y, Stannat W, Obermayer K (2013) Risk-sensitive Markov control processes. SIAM J Control Optim 51(5):3652–3672
Shen Y, Stannat W, Obermayer K (2014) A unified framework for risk-sensitive Markov control processes. In: Proceedings of the 53rd IEEE Conference on Decision and Control, IEEE, pp 1073–1078
Sladkỳ K (2008) Growth rates and average optimality in risk-sensitive Markov decision chains. Kybernetika 44(2):205–226
Sladkỳ K (2018) Risk-sensitive average optimality in Markov decision processes. Kybernetika 54(6):1218–1230
Staino A, Russo E (2020) Nested Conditional Value-at-Risk portfolio selection: a model with temporal dependence driven by market-index volatility. Eur J Oper Res 280(2):741–753
Stettner Ł (1999) Risk sensitive portfolio optimization. Math Methods Oper Res 50(3):463–474
Stettner Ł (2005) Discrete time risk sensitive portfolio optimization with consumption and proportional transaction costs. Applicationes Mathematicae 4(32):395–404
Stettner Ł (2023) Certainty equivalent control of discrete time Markov processes with the average reward functional. Syst Control Lett 181:105627
Sutton RS, Barto AG (2018) Reinforcement Learning: An Introduction. MIT Press, Cambridge
Tamar A, Chow Y, Ghavamzadeh M et al (2016) Sequential decision making with coherent risk. IEEE Trans Autom Control 62(7):3323–3338

Uğurlu K (2017) Controlled Markov decision processes with AVaR criteria for unbounded costs. J Comput Appl Math 319:24–37
Uğurlu K (2018) Robust optimal control using conditional risk mappings in infinite horizon. J Comput Appl Math 344:275–287
Von Neumann J, Morgenstern O (2007) Theory of Games and Economic Behavior (60th Anniversary Commemorative Edition). Princeton University Press, Princeton
Weil P (1990) Nonexpected utility in macroeconomics. Q J Econ 105(1):29–42
Weil P (1993) Precautionary savings and the permanent income hypothesis. Rev Econ Stud 60(2):367–383
Whittle P (1981) Risk-sensitive linear/quadratic/Gaussian control. Adv Appl Probab 13(4):764–777
Wozabal D, Rameseder G (2020) Optimal bidding of a virtual power plant on the spanish day-ahead and intraday market for electricity. Eur J Oper Res 280(2):639–655
Xia L (2020) Risk-sensitive Markov decision processes with combined metrics of mean and variance. Prod Oper Manag 29(12):2808–2827
Xia L, Glynn PW (2022) Risk-sensitive Markov decision processes with long-run CVaR criterion. ArXiv preprint arXiv:2210.08740
Xu W, Gao X, He X (2023) Regret bounds for Markov decision processes with recursive optimized certainty equivalents. ArXiv preprint arXiv:2301.12601
Zhang W, Rahimian H, Bayraksan G (2016) Decomposition algorithms for risk-averse multistage stochastic programs with application to water allocation under uncertainty. INFORMS J Comput 28(3):385–404