

MARKOV DECISION PROCESSES WITH RECURSIVE RISK MEASURES

NICOLE BÄUERLE AND ALEXANDER GLAUNER

ABSTRACT. In this paper, we consider risk-sensitive Markov Decision Processes (MDPs) with Borel state and action spaces and unbounded cost under both finite and infinite planning horizons. Our optimality criterion is based on the recursive application of static risk measures. This is motivated by recursive utilities in the economic literature, has been studied before for the entropic risk measure and is extended here to an axiomatic characterization of suitable risk measures. We derive a Bellman equation and prove the existence of Markovian optimal policies. For an infinite planning horizon, the model is shown to be contractive and the optimal policy to be stationary. Moreover, we establish a connection to distributionally robust MDPs, which provides a global interpretation of the recursively defined objective function. Monotone models are studied in particular.

KEY WORDS: Risk-Sensitive Markov Decision Process; Risk Measure; Robustness
AMS SUBJECT CLASSIFICATIONS: 90C40, 91G70

1. INTRODUCTION

In this paper, we extend Markov Decision Processes (MDPs) to a recursive application of static risk measures. Our framework is such that it is applicable for a wide range of practical models. In particular we consider Borel state and action spaces, unbounded cost functions and rather general risk measures.

In standard MDP theory we are concerned with minimizing the expected discounted cost of a controlled dynamic system over a finite or infinite time horizon. The expectation has the nice property that it can be iterated which yields a recursive solution theory for these kind of problems, see e.g. the textbooks by [24, 17, 9] for a mathematical treatment. However, there are applications where the simple expectation, which does not reflect the true risk of a decision, might not be the best choice to evaluate decisions. In particular when the management of cash flows is concerned, economists prefer to use dynamic utilities to compare their performance. An early axiomatic treatment of a dynamic utility which takes into account the revealed information is [21]. Later, the focus was more on an extension of static risk measures to dynamic risk measures. We mention here the following axiomatic approaches [15, 25, 13, 34] just to name some of them. These approaches do not consider a control. For an overview up to 2011 see [1]. Later, besides the axiomatic characterization another important aspect has been time-consistency of the dynamic risk measures, see e.g. [11, 10] for the situation without control and [31, 30] for the situation with control. In the latter reference it is shown that the only time-consistent risk measures are those which iterate static ones. See also [19] for different ways to apply dynamic risk measures.

Approaches to establish a theory for controlled dynamic risk measures have before been presented in [27, 32, 12, 3]. [27] is an axiomatic approach. The paper restricts to bounded random variables for the infinite time horizon and uses Markov risk measures to obtain time-consistency. However, some assumptions are indirect properties of the risk measures (see e.g. Theorem 2 in this paper). In [32] so-called risk maps are considered and weighted norm spaces are used to treat unbounded rewards. Concepts like sub- and uppermodules are needed to prove the main theorems. [12] also treats unbounded cost problems but restricts to coherent risk measures. Moreover, some assumptions on the existence of limits are made because the Fatou property of risk measures is not exploited there. The note [3] restricts the discussion to entropic risk measures.

There are also papers which apply recursive risk measures in specific problems. E.g. in [20] a convex combination of expectation and Expected Shortfall is used to tackle the problem of electric vehicle charging in a dynamic decision framework. The authors there compare true risk and expectation with the standard MDP problem. [28] investigate dynamic pricing problems with dynamic Expected Shortfall and also compare their findings to the standard MDP. [6] consider optimal dividend payments under dynamic entropic risk measures and [7] optimal growth models under dynamic entropic risk measures. In [33] sampling-based algorithms for coherent risk measures are constructed.

In this paper now we restrict to a recursive application of static risk measures and use unbounded cost functions. The risk measures may be rather general and we state the needed properties for every result. In contrast to the earlier literature our assumptions are in most cases assumptions on the model data alone. We also treat the important case of monotone models where comonotonicity of the risk measures is crucial.

In more detail the structure of our paper is as follows: In the next section, we summarize some important concepts of risk measures. We consider in particular distortion risk measures. In Section 3, we introduce our Markov Decision model. The finite-horizon optimization problem is then considered in Section 4. The aim is to minimize recursive risk measures over a finite time horizon. We show here that for proper coherent risk measures with the Fatou property local bounding functions are sufficient for the well-posedness of the optimization problem. Otherwise global bounding function may be necessary. Under some continuity and compactness conditions on the MDP data we show that an optimal policy exists which is Markovian and the value of the problem can be computed recursively. In Section 5, we consider the problem with an infinite time horizon. Here the first result, which states a fixed point property of the value function and the existence of an optimal stationary policy, is under the condition of coherence of the risk measure. In Section 6, we briefly discuss the relation to distributionally robust MDP. In Section 7, we consider MDP with monotonicity properties. Here we can work with semicontinuous model data. Another special case arises when the cost function is bounded from below. Then, under the monotonicity assumptions the monetary risk measure does not have to be coherent but comonotonic additive to obtain the same results. In the last section, we illustrate our results with some examples: We show that in a monotone recursive Value-at-Risk model, the optimal policy is myopic. Moreover, we consider stopping problems, casino games and a cash balance problem where structural properties of the standard MDP formulation still hold.

2. RISK MEASURES

Let a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a real number $p \in [1, \infty)$ be fixed. With $q \in (1, \infty]$ we denote the conjugate index satisfying $\frac{1}{p} + \frac{1}{q} = 1$ under the convention $\frac{1}{\infty} = 0$. Henceforth, $L^p = L^p(\Omega, \mathcal{A}, \mathbb{P})$ denotes the vector space of real-valued random variables which have an integrable p -th moment. We follow the convention of the actuarial literature that positive realizations of random variables represent losses and negative ones gains. A *risk measure* is a functional $\rho : L^p \rightarrow \bar{\mathbb{R}}$. The following properties will be important.

Definition 2.1. A risk measure $\rho : L^p \rightarrow \bar{\mathbb{R}}$ is

- a) *law-invariant* if $\rho(X) = \rho(Y)$ for X, Y with the same distribution.
- b) *monotone* if $X \leq Y$ implies $\rho(X) \leq \rho(Y)$.
- c) *translation invariant* if $\rho(X + m) = \rho(X) + m$ for all $m \in \mathbb{R}$.
- d) *normalized* if $\rho(0) = 0$.
- e) *finite* if $\rho(L^p) \subseteq \mathbb{R}$.
- f) *comonotonic additive* if $\rho(X + Y) = \rho(X) + \rho(Y)$ for all comonotonic X, Y .
- g) *positive homogeneous* if $\rho(\lambda X) = \lambda \rho(X)$ for all $\lambda \in \mathbb{R}_+$.
- h) *convex* if $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda \rho(X) + (1 - \lambda)\rho(Y)$ for $\lambda \in [0, 1]$.
- i) *subadditive* if $\rho(X + Y) \leq \rho(X) + \rho(Y)$ for all X, Y .

j) said to have the *Fatou property*, if for every sequence $\{X_n\}_{n \in \mathbb{N}} \subseteq L^p$ with $|X_n| \leq Y$ \mathbb{P} -a.s. for some $Y \in L^p$ and $X_n \rightarrow X$ \mathbb{P} -a.s. for some $X \in L^p$ it holds

$$\liminf_{n \rightarrow \infty} \rho(X_n) \geq \rho(X).$$

A risk measure is called *monetary* if it is monotone and translation invariant. It appears to be consensus in the literature that these two properties are a necessary minimal requirement for any risk measure. Monetary risk measures which are additionally positive homogeneous and subadditive are referred to as *coherent*. Further, note that positive homogeneity implies normalization and makes convexity and subadditivity equivalent. The Fatou property means that the risk measure is lower semicontinuous w.r.t. dominated convergence.

Lemma 2.2 (Theorem 7.24 in [26]). *Finite and convex monetary risk measures have the Fatou property.*

Coherent risk measures satisfy a triangular inequality.

Lemma 2.3 (Prop. 6 in [23]). *For a coherent risk measure ρ and $X, Y \in L^p$ it holds*

$$|\rho(X) - \rho(Y)| \leq \rho(|X - Y|).$$

We denote by $\mathcal{M}_1(\Omega, \mathcal{A}, \mathbb{P})$ the set of probability measures on (Ω, \mathcal{A}) which are absolutely continuous with respect to \mathbb{P} and define

$$\mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P}) = \left\{ \mathbb{Q} \in \mathcal{M}_1(\Omega, \mathcal{A}, \mathbb{P}) : \frac{d\mathbb{Q}}{d\mathbb{P}} \in L^q(\Omega, \mathcal{A}, \mathbb{P}) \right\}.$$

Recall that an extended real-valued convex functional is called *proper* if it never attains $-\infty$ and is strictly smaller than $+\infty$ in at least one point. Coherent risk measures have the following dual or robust representation.

Proposition 2.4 (Theorem 7.20 in [26]). *A functional $\rho : L^p \rightarrow \bar{\mathbb{R}}$ is a proper coherent risk measure with the Fatou property if and only if there exists a subset $\mathcal{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$ such that*

$$\rho(X) = \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X], \quad X \in L^p.$$

The supremum is attained since the subset $\mathcal{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$ can be chosen $\sigma(L^q, L^p)$ -compact and the functional $\mathbb{Q} \mapsto \mathbb{E}^{\mathbb{Q}}[X]$ is $\sigma(L^q, L^p)$ -continuous.

With the dual representation we can derive a complementary inequality to subadditivity.

Lemma 2.5. *A proper coherent risk measure with the Fatou property $\rho : L^p \rightarrow \bar{\mathbb{R}}$ satisfies*

$$\rho(X + Y) \geq \rho(X) - \rho(-Y) \quad \text{for all } X, Y \in L^p.$$

Proof. By Proposition 2.4 it holds for $X, Y \in L^p$

$$\begin{aligned} \rho(X + Y) &= \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X + Y] = \sup_{\mathbb{Q} \in \mathcal{Q}} \left(\mathbb{E}^{\mathbb{Q}}[X] + \mathbb{E}^{\mathbb{Q}}[Y] \right) \\ &\geq \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[X] + \inf_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[Y] = \rho(X) - \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}}[-Y] \\ &= \rho(X) - \rho(-Y). \end{aligned} \quad \square$$

In the following, $F_X(x) = \mathbb{P}(X \leq x)$ denotes the distribution function, $S_X(x) = 1 - F_X(x)$, $x \in \mathbb{R}$, the survival function and $F_X^{-1}(u) = \inf\{x \in \mathbb{R} : F_X(x) \geq u\}$, $u \in [0, 1]$, the quantile function of a random variable X . Many established risk measures belong to the large class of distortion risk measures.

Definition 2.6. a) An increasing function $g : [0, 1] \rightarrow [0, 1]$ with $g(0) = 0$ and $g(1) = 1$ is called *distortion function*.

b) The *distortion risk measure* w.r.t. a distortion function g is defined by $\rho_g : L^p \rightarrow \bar{\mathbb{R}}$,

$$\rho_g(X) = \int_0^\infty g(S_X(x)) \, dx - \int_{-\infty}^0 1 - g(S_X(x)) \, dx$$

whenever at least one of the integrals is finite.

Distortion risk measures have many of the properties introduced in Definition 2.1, see e.g. [29].

Lemma 2.7. a) *Distortion risk measures are law invariant, monotone, translation invariant, normalized, positive homogeneous and comonotonic additive.*

b) *A distortion risk measure is subadditive if and only if the distortion function g is concave.*

There is an alternative representation of distortion risk measures in terms of Lebesgue-Stieltjes integrals based on the quantile function in lieu of the survival function of the risk X .

Remark 2.8. For a distortion risk measure ρ_g with left-continuous distortion function g it holds

$$\rho_g(X) = \int_0^1 F_X^{-1}(u) \, d\bar{g}(u), \quad (2.1)$$

where $\bar{g}(u) = 1 - g(1 - u)$, $u \in [0, 1]$, is the dual distortion function, cf. [14]. For a continuous concave distortion function $g : [0, 1] \rightarrow [0, 1]$, the dual distortion function $\bar{g} : [0, 1] \rightarrow [0, 1]$ is continuous convex and can be written as $\bar{g}(x) = \int_0^x \phi(s) \, ds$ for an increasing right-continuous function $\phi : [0, 1] \rightarrow \mathbb{R}_+$, which is called *spectrum*. By the properties of the Lebesgue-Stieltjes integral, (2.1) can then be written as

$$\rho_g(X) = \rho_\phi(X) = \int_0^1 F_X^{-1}(u) \phi(u) \, du. \quad (2.2)$$

Therefore, distortion risk measures with continuous concave distortion function are referred to as *spectral risk measures*. Note that continuity of g is an additional requirement only in 0, since an increasing concave function on $[0, 1]$ is already continuous on $(0, 1]$.

Due to Hölder's inequality, spectral risk measures $\rho_\phi : L^p \rightarrow \bar{\mathbb{R}}$ with spectrum $\phi \in L^q$ fulfill

$$|\rho_\phi(X)| = \left| \int_0^1 F_X^{-1}(u) \phi(u) \, du \right| \leq \int_0^1 |F_X^{-1}(u)| \phi(u) \, du = (\mathbb{E}|F_X^{-1}(U)|^p)^{\frac{1}{p}} (\mathbb{E}|\phi(U)|^q)^{\frac{1}{q}} < \infty,$$

where $U \sim \mathcal{U}([0, 1])$ is arbitrary. Hence, they have the Fatou property by Lemma 2.2.

Example 2.9. The most widely used risk measure in finance and insurance *Value-at-Risk*

$$\text{VaR}_\alpha(X) = F_X^{-1}(\alpha), \quad \alpha \in (0, 1),$$

is a distortion risk measure with distortion function $g(u) = \mathbb{1}_{(1-\alpha, 1]}(u)$. Since the distortion function is not concave, Value-at-Risk is not coherent and especially not a spectral risk measure. The lack of coherence can be overcome by using *Expected Shortfall*

$$\text{ES}_\alpha(X) = \frac{1}{1-\alpha} \int_\alpha^1 F_X^{-1}(u) \, du, \quad \alpha \in [0, 1).$$

The corresponding distortion function $g(u) = \min\{\frac{u}{1-\alpha}, 1\}$ is concave and Expected Shortfall thus coherent. It is also spectral with $\phi(u) = \frac{1}{1-\alpha} \mathbb{1}_{[\alpha, 1]}(u)$. Due to the bounded spectrum, ES has the Fatou property. The well-known *entropic risk measure*

$$\rho_\gamma(X) = \frac{1}{\gamma} \log \mathbb{E} [e^{\gamma X}], \quad \gamma > 0,$$

is an example of a law-invariant and convex monetary risk measure which does not belong to the distortion class. For random variables with existing moment-generating function it has the Fatou property directly by dominated convergence.

To the best of our knowledge, it has surprisingly not been investigated in the literature whether Value-at-Risk has the Fatou property.

Lemma 2.10. *Value-at-Risk has the Fatou property.*

Proof. Assume the contrary. Then there exists a sequence $\{X_n\}_{n \in \mathbb{N}} \subseteq L^p$ with $|X_n| \leq Y$ \mathbb{P} -a.s. for some $Y \in L^p$ and $X_n \rightarrow X$ \mathbb{P} -a.s. for some $X \in L^p$ such that

$$\liminf_{n \rightarrow \infty} \text{VaR}_\alpha(X_n) < \text{VaR}_\alpha(X).$$

I.e. there is an $\epsilon > 0$ such that for every $\delta \in (0, \epsilon)$

$$\liminf_{n \rightarrow \infty} F_{X_n}^{-1}(\alpha) \leq F_X^{-1}(\alpha) - \delta.$$

Hence, there exists a subsequence $\{F_{X_{N_k}}^{-1}(\alpha)\}_{k \in \mathbb{N}}$ such that for all $k \in \mathbb{N}$ and $\delta \in (0, \epsilon)$

$$F_{X_{N_k}}^{-1}(\alpha) \leq F_X^{-1}(\alpha) - \delta$$

or equivalently by the properties of generalized inverses $\alpha \leq F_{X_{N_k}}(F_X^{-1}(\alpha) - \delta)$. Since F_X has at most countably many discontinuities, we can choose $\delta_0 \in (0, \epsilon)$ such that $F_X^{-1}(\alpha) - \delta_0$ is a point of continuity of F_X . Then, by the definition of convergence in distribution

$$\alpha \leq \lim_{k \rightarrow \infty} F_{X_{N_k}}(F_X^{-1}(\alpha) - \delta_0) = F_X(F_X^{-1}(\alpha) - \delta_0).$$

Again by the properties of generalized inverses, this is equivalent to $F_X^{-1}(\alpha) \leq F_X^{-1}(\alpha) - \delta_0$, a contradiction. \square

3. THE MARKOV DECISION MODEL

We consider the following standard Markov Decision Process with general Borel state and action spaces. The *state space* E is a Borel space with Borel σ -algebra $\mathcal{B}(E)$ and the *action space* A is a Borel space with Borel σ -Algebra $\mathcal{B}(A)$. The possible state-action combinations at time n form a measurable subset D_n of $E \times A$ such that D_n contains the graph of a measurable mapping $E \rightarrow A$. The x -section of D_n ,

$$D_n(x) = \{a \in A : (x, a) \in D_n\},$$

is the set of admissible actions in state $x \in E$ at time n . Note that the sets $D_n(x)$ are non-empty. We assume that the dynamics of the MDP are given by measurable *transition functions* $T_n : D_n \times \mathcal{Z} \rightarrow E$ and depend on *disturbances* Z_1, Z_2, \dots which are independent random elements on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in a measurable space $(\mathcal{Z}, \mathfrak{B})$. When the current state is x_n , the controller chooses action $a_n \in D_n(x_n)$ and z_{n+1} is the realization of Z_{n+1} , then the next state is given by

$$x_{n+1} = T_n(x_n, a_n, z_{n+1}).$$

The *one-stage cost function* $c_n : D_n \times E \rightarrow \mathbb{R}$ gives the cost $c_n(x, a, x')$ for choosing action a if the system is in state x at time n and the next state is x' . The *terminal cost function* $c_N : E \rightarrow \mathbb{R}$ gives the cost $c_N(x)$ if the system terminates in state x .

The model data is supposed to have the following continuity and compactness properties.

- Assumption 3.1.**
- (i) The sets $D_n(x)$ are compact and $E \ni x \mapsto D_n(x)$ are upper semicontinuous, i.e. if $x_k \rightarrow x$ and $a_k \in D_n(x_k)$, $k \in \mathbb{N}$, then (a_k) has an accumulation point in $D_n(x)$.
 - (ii) The transition functions T_n are continuous in (x, a) .
 - (iii) The one-stage cost functions c_n and the terminal cost function c_N are lower semicontinuous.

Under a finite planning horizon $N \in \mathbb{N}$, we consider the model data for $n = 0, \dots, N-1$. The decision model is called *stationary* if D , T do not depend on n , the disturbances are identically distributed, the one-stage cost functions are of the form $c_n = \beta^n c$, and the terminal cost function is $\beta^N c_N$, where $\beta \in (0, 1]$ is a discount factor. In that case, Z denotes a representative of the disturbance distribution. For a non-stationary model one may think of the discount factor being included in the cost functions. If the model is stationary and the terminal cost is zero, we allow for an *infinite time horizon* $N = \infty$.

For $n \in \mathbb{N}_0$ we denote by \mathcal{H}_n the set of *feasible histories* of the decision process up to time n

$$h_n = \begin{cases} x_0, & \text{if } n = 0, \\ (x_0, a_0, x_1, \dots, x_n), & \text{if } n \geq 1, \end{cases}$$

where $a_k \in D_k(x_k)$ for $k \in \mathbb{N}_0$. In order for the controller's decisions to be implementable, they must be based on the information available at the time of decision making, i.e. be functions of the history of the decision process.

- Definition 3.2.**
- a) A measurable mapping $d_n : \mathcal{H}_n \rightarrow A$ with $d_n(h_n) \in D_n(x_n)$ for every $h_n \in \mathcal{H}_n$ is called *decision rule* at time n . A finite sequence $\pi = (d_0, \dots, d_{N-1})$ is called *N -stage policy* and a sequence $\pi = (d_0, d_1, \dots)$ is called *policy*.
 - b) A decision rule at time n is called *Markov* if it depends on the current state only, i.e. $d_n(h_n) = d_n(x_n)$ for all $h_n \in \mathcal{H}_n$. If all decision rules are Markov, the (N -stage) policy is called *Markov*.
 - c) An (N -stage) policy π is called *stationary* if $\pi = (d, \dots, d)$ or $\pi = (d, d, \dots)$, respectively, for some Markov decision rule d .

With $\Pi \supseteq \Pi^M \supseteq \Pi^S$ we denote the sets of all policies, Markov policies and stationary policies. It will be clear from the context if N -stage or infinite stage policies are meant. An admissible policy always exists as D contains the graph of a measurable mapping.

Since risk measures are defined as real-valued mappings of random variables, we will work with a functional representation of the decision process. The law of motion does not need to be specified explicitly. We define for an initial state $x_0 \in E$ and a policy $\pi \in \Pi$

$$X_0^\pi = x_0, \quad X_{n+1}^\pi = T(X_n^\pi, d_n(H_n^\pi), Z_{n+1}).$$

Here, the process $(H_n^\pi)_{n \in \mathbb{N}_0}$ denotes the history of the decision process viewed as a random element, i.e.

$$H_0^\pi = x_0, \quad H_1^\pi = (X_0^\pi, d_0(X_0^\pi), X_1^\pi), \quad \dots, \quad H_n^\pi = (H_{n-1}^\pi, d_{n-1}(H_{n-1}^\pi), X_n^\pi).$$

Under a Markov policy the recourse on the random history of the decision process is not needed.

4. COST MINIMIZATION UNDER A FINITE PLANNING HORIZON

For a finite planning horizon $N \in \mathbb{N}$, we consider the non-stationary decision model. In the classical context of the risk-neutral expected cost criterion, the value of a policy $\pi \in \Pi$ at time $n = 0, \dots, N$ given $h_n \in \mathcal{H}_n$ is defined as

$$V_{n\pi}(h_n) = \mathbb{E}_{nh_n} \left[\sum_{k=n}^{N-1} c_k(X_k^\pi, d_k(H_k^\pi), X_{k+1}^\pi) + c_N(X_N^\pi) \right], \quad n = 0, \dots, N,$$

where \mathbb{E}_{nh_n} is the conditional expectation given $H_n^\pi = h_n$. Under suitable integrability conditions, $V_{n\pi}$ satisfies the value iteration

$$V_{n\pi}(h_n) = \mathbb{E} \left[c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) \right],$$

see e.g. Theorem 2.3.4 in [9]. In order to take risk-sensitive preferences of the controller into account, the approach here is to replace the factorization of conditional expectation in the value iteration by a risk measure, meaning that static risk measures are recursively applied at each stage. In the special case of the entropic risk measure, this approach has been studied by [3] in

an abstract setting and by [6, 7] in applications to optimal dividend payments and stochastic optimal growth. Their choice of the risk measure is motivated by the fact that the entropic risk measure coincides with the certainty equivalent of an exponential utility function. In the economic literature, recursive utilities have been widely studied. For a literature overview we refer the reader to [22].

Let $p \in [1, \infty)$ with conjugate index $q \in [1, \infty]$ and let $\rho_0, \dots, \rho_{N-1} : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \bar{\mathbb{R}}$ be monetary risk measures. We define the *value of a policy* $\pi = (d_0, \dots, d_{N-1}) \in \Pi$ at time $n = 0, \dots, N$ given history $h_n \in \mathcal{H}_n$ recursively as

$$V_{N\pi}(h_N) = c_N(x_N),$$

$$V_{n\pi}(h_n) = \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1}))\right).$$

In the special case that the one-stage cost functions c_n do not depend on the next state of the decision process, the value of a policy simplifies to

$$V_{n\pi}(h_n) = c_n(x_n, d_n(h_n)) + \rho_n(V_{n+1\pi}(h_n, d_n(h_n), X_{n+1}^\pi)), \quad h_n \in \mathcal{H}_n,$$

due to the translation invariance of monetary risk measures.

Remark 4.1. For the recursive definition of the policy values to be meaningful, we need to make sure that the risk measures are applied to elements of $L^p(\Omega, \mathcal{A}, \mathbb{P})$. This has two aspects: integrability will be ensured by Assumption 4.2, but first of all $V_{n\pi}$ needs to be a measurable function for all $\pi \in \Pi$ and $n = 0, \dots, N$. For most risk measures with practical relevance, this is fulfilled:

- In the risk-neutral case, i.e. for $\rho = \mathbb{E}$, and also for the entropic risk measure ρ_γ the measurability is obvious.
- For distortion risk measures, the measurability is guaranteed, too. To see this, we proceed backwards. For N there is nothing to show and if $V_{n+1\pi}$ is measurable, the function

$$f(h_n, z) = c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), z)) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), z))$$

is measurable as a composition of measurable maps. Then, Fubini's theorem yields that the survival function of $f(h_n, Z_{n+1})$

$$S(t|h_n) = \int \mathbb{1}\{f(h_n, Z_{n+1}(\omega)) > t\} \mathbb{P}(d\omega)$$

is measurable. A distortion function g is increasing and hence measurable. So again by Fubini's theorem we obtain the measurability of

$$V_{n\pi}(h_n) = \rho_g(f(h_n, Z_{n+1})) = \int_0^\infty g(S(t|h_n)) dt - \int_{-\infty}^0 1 - g(S(t|h_n)) dt$$

since the integrands are non-negative and compositions of measurable maps.

- For proper coherent risk measures with the Fatou property one can insert the dual representation of Proposition 2.4. Then, an optimal measurable selection argument as in Theorem 3.6 in [5] yields the measurability.

Throughout, it is implicitly assumed that the risk measures are chosen such that all policy values are measurable.

The *value functions* are given by

$$V_n(h_n) = \inf_{\pi \in \Pi} V_{n\pi}(h_n), \quad h_n \in \mathcal{H}_n,$$

for $n = 0, \dots, N$ and the controller's optimization objective is

$$V_0(x) = \inf_{\pi \in \Pi} V_{0\pi}(x), \quad x \in E.$$

In order to have well-defined value functions, we need some finiteness conditions instead of the usual integrability conditions. Moreover, we require some basic properties for the risk measures.

Assumption 4.2. (i) There exist $\underline{\epsilon}, \bar{\epsilon} \geq 0$ with $\underline{\epsilon} + \bar{\epsilon} = 1$ and measurable functions $\mathbf{b} : E \rightarrow (-\infty, -\underline{\epsilon}]$ and $\bar{\mathbf{b}} : E \rightarrow [\bar{\epsilon}, \infty)$ such that it holds for all policies $\pi \in \Pi$ and all $n = 0, \dots, N$

$$\mathbf{b}(x_n) \leq V_{n\pi}(h_n) \leq \bar{\mathbf{b}}(x_n), \quad h_n \in \mathcal{H}_n.$$

(ii) We define $\mathbf{b} : E \rightarrow [1, \infty)$, $\mathbf{b}(x) = \bar{\mathbf{b}}(x) - \mathbf{b}(x)$. For all $n = 0, \dots, N-1$ and $(\bar{x}, \bar{a}) \in D_n$ there exists an $\epsilon > 0$ and measurable functions $\Theta_{n,1}^{\bar{x}, \bar{a}}, \Theta_{n,2}^{\bar{x}, \bar{a}} : \mathcal{Z} \rightarrow \mathbb{R}_+$ such that $\Theta_{n,1}^{\bar{x}, \bar{a}}(Z_{n+1}), \Theta_{n,2}^{\bar{x}, \bar{a}}(Z_{n+1}) \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ and

$$|c_n(x, a, T_n(x, a, z))| \leq \Theta_{n,1}^{\bar{x}, \bar{a}}(z), \quad \mathbf{b}(T_n(x, a, z)) \leq \Theta_{n,2}^{\bar{x}, \bar{a}}(z)$$

for all $z \in \mathcal{Z}$ and $(x, a) \in B_\epsilon(\bar{x}, \bar{a}) \cap D_n$. Here, $B_\epsilon(\bar{x}, \bar{a})$ is the closed ball around (\bar{x}, \bar{a}) w.r.t. an arbitrary product metric on $E \times A$.

(iii) The monetary risk measures $\rho_0, \dots, \rho_{N-1} : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \bar{\mathbb{R}}$ are law invariant and have the Fatou property.

$\underline{\mathbf{b}}, \bar{\mathbf{b}}$ are called (*global*) *lower* and *upper bounding function*, respectively, while \mathbf{b} is referred to as (*global*) *bounding function*. Since $\underline{\mathbf{b}}$ is non-positive and $\bar{\mathbf{b}}$ is non-negative it holds

$$\underline{\mathbf{b}}(x_n) \leq -V_{n\pi}^-(h_n) \leq V_{n\pi}(h_n) \leq V_{n\pi}^+(h_n) \leq \bar{\mathbf{b}}(x_n), \quad h_n \in \mathcal{H}_n,$$

and consequently $|V_{n\pi}(h_n)| \leq \mathbf{b}(x_n)$. Bold print is used to distinguish these global bounding functions from the usual local (stage-wise) bounding functions used for risk-neutral MDP. Such local bounding functions can be introduced for the risk-sensitive recursive optimality criterion, too, if the risk measures have additional properties. Note that without any further properties on the risk measure we cannot construct global bounding functions from local ones.

Lemma 4.3. *Let $\rho_0, \dots, \rho_{N-1}$ be proper coherent risk measures with the Fatou property. If there exist $\underline{\epsilon}, \bar{\epsilon} \geq 0$ with $\underline{\epsilon} + \bar{\epsilon} = 1$, measurable functions $\underline{b} : E \rightarrow (-\infty, -\underline{\epsilon}]$, $\bar{b} : E \rightarrow [\bar{\epsilon}, \infty)$ and a constant $\alpha \in (0, 1)$ such that*

$$\begin{aligned} \rho_n(c_n(x, a, T_n(x, a, Z_{n+1}))) &\geq \underline{b}(x), & \rho_n(-\bar{b}(T_n(x, a, Z_{n+1}))) &\leq -\alpha \underline{b}(x), \\ \rho_n(c_n(x, a, T_n(x, a, Z_{n+1}))) &\leq \bar{b}(x), & \rho_n(\bar{b}(T_n(x, a, Z_{n+1}))) &\leq \alpha \bar{b}(x), \end{aligned}$$

for all $n = 0, \dots, N-1$ and $(x, a) \in D_n$ as well as $\underline{b}(x) \leq c_N(x) \leq \bar{b}(x)$ for all $x \in E$, then

$$\underline{\mathbf{b}} = \frac{1}{1-\alpha} \underline{b}, \quad \bar{\mathbf{b}} = \frac{1}{1-\alpha} \bar{b} \quad \text{and} \quad \mathbf{b} = \frac{1}{1-\alpha} b$$

are global bounding functions satisfying Assumption 4.2 (i).

Proof. We proceed by backward induction. At time N we have

$$\underline{\mathbf{b}}(x_N) \leq \underline{b}(x_N) \leq c_N(x_N) \leq \bar{b}(x_N) \leq \bar{\mathbf{b}}(x_N), \quad h_N \in \mathcal{H}_N.$$

Assuming the assertion holds for time $n+1$ it follows for time n :

$$\begin{aligned} V_{n\pi}(h_n) &= \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\ &\geq \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + \frac{1}{1-\alpha} \underline{b}(T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\ &\geq \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1}))\right) - \frac{1}{1-\alpha} \rho_n\left(-\bar{b}(T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\ &\geq \underline{b}(x_n) + \frac{\alpha}{1-\alpha} \underline{b}(x_n) = \underline{\mathbf{b}}(x_n). \end{aligned}$$

The second inequality is by Lemma 2.5. Regarding the upper bounding function one can argue similarly using the subadditivity of ρ_n instead.

$$\begin{aligned}
 V_{n\pi}(h_n) &= \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\
 &\leq \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + \frac{1}{1-\alpha}\bar{b}(T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\
 &\leq \rho_n\left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1}))\right) + \frac{1}{1-\alpha}\rho_n\left(\bar{b}(T_n(x_n, d_n(h_n), Z_{n+1}))\right) \\
 &\leq \bar{b}(x_n) + \frac{\alpha}{1-\alpha}\bar{b}(x_n) = \bar{\mathbf{b}}(x_n). \quad \square
 \end{aligned}$$

Remark 4.4. a) Concerning the requirements on a local lower bounding function in Lemma 4.3 it should be noted that $\rho_n(-\underline{b}(T_n(x, a, Z_{n+1}))) \leq -\alpha\underline{b}(x)$ is a stronger assumption than

$$\rho_n(\underline{b}(T_n(x, a, Z_{n+1}))) \geq \alpha\underline{b}(x). \quad (4.1)$$

Indeed, since $\underline{b} \leq 0$ the monotonicity and normalization of ρ_n yields $\rho_n(\underline{b}(T_n(x, a, Z_{n+1}))) \leq 0$. Consequently, we have by Lemma 2.3

$$\begin{aligned}
 -\rho_n(\underline{b}(T_n(x, a, Z_{n+1}))) &= |\rho_n(\underline{b}(T_n(x, a, Z_{n+1})))| \leq \rho_n\left(|\underline{b}(T_n(x, a, Z_{n+1}))|\right) \\
 &= \rho_n\left(-\underline{b}(T_n(x, a, Z_{n+1}))\right) \leq -\alpha\underline{b}(x).
 \end{aligned}$$

Multiplying with (-1) yields (4.1).

- b) If the one-stage cost functions are bounded and the monetary risk measures $\rho_0, \dots, \rho_{N-1}$ normalized, the local bounding functions \underline{b}, \bar{b} can be chosen constant. Where we have used Lemma 2.5 or subadditivity in the proof of Lemma 4.3, one can then simply argue with translation invariance. Note that normalization is no structural restriction for monetary risk measures due to the translation invariance.

With the bounding function \mathbf{b} we define the function space

$$\mathbb{B}_{\mathbf{b}} = \{v : E \rightarrow \mathbb{R} \mid v \text{ measurable with } \lambda \in \mathbb{R}_+ \text{ s.t. } |v(x)| \leq \lambda \mathbf{b}(x) \text{ for all } x \in E\}.$$

Endowing $\mathbb{B}_{\mathbf{b}}$ with the weighted supremum norm

$$\|v\|_{\mathbf{b}} = \sup_{x \in E} \frac{|v(x)|}{\mathbf{b}(x)}$$

makes $(\mathbb{B}_{\mathbf{b}}, \|\cdot\|_{\mathbf{b}})$ a Banach space, cf. Proposition 7.2.1 in [18]. In case we have local bounding functions as in Lemma 4.3, it holds

$$\begin{aligned}
 \mathbb{B}_{\mathbf{b}} &= \{v : E \rightarrow \mathbb{R} \mid v \text{ measurable with } \lambda \in \mathbb{R}_+ \text{ s.t. } |v(x)| \leq \lambda \mathbf{b}(x) \text{ for all } x \in E\} \\
 &= \{v : E \rightarrow \mathbb{R} \mid v \text{ measurable with } \lambda \in \mathbb{R}_+ \text{ s.t. } |v(x)| \leq \lambda b(x) \text{ for all } x \in E\} \\
 &= \mathbb{B}_b
 \end{aligned}$$

and the weighted supremum norms $\|\cdot\|_{\mathbf{b}}, \|\cdot\|_b$ are equivalent.

Lemma 4.5. *Let $v \in \mathbb{B}_{\mathbf{b}}$ and $n \in \{0, \dots, N-1\}$. Under Assumptions 3.1 (i) and 4.2 (ii) each sequence of random variables*

$$C_k = c_n(x_k, a_k, T_n(x_k, a_k, Z_{n+1})) + v(T_n(x_k, a_k, Z_{n+1}))$$

induced by a convergent sequence $\{(x_k, a_k)\}_{k \in \mathbb{N}}$ in D_n has an L^p -bound \bar{C} , i.e. $|C_k| \leq \bar{C} \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ for all $k \in \mathbb{N}$.

Proof. There exists a constant $\lambda \in \mathbb{R}_+$ such that $|v| \leq \lambda b$. Since D_n is closed by Lemma A.2.2 in [9], the limit point (x_0, a_0) of $\{(x_k, a_k)\}_{k \in \mathbb{N}}$ lies in D_n . Let $\epsilon > 0$ be the constant from Assumption 4.2 (ii) corresponding to (x_0, a_0) . Since the sequence is convergent, there exists $m \in \mathbb{N}$ such that $(x_k, a_k) \in B_\epsilon(x_0, a_0) \cap D_n$ for all $k > m$. For the finite number of points

$(x_0, a_0), (x_1, a_1), \dots, (x_m, a_m)$ there exist bounding functions $\Theta_{n,1}^{x_i, a_i}, \Theta_{n,2}^{x_i, a_i}$ by Assumption 4.2 (ii). Thus, the random variable

$$\bar{C} = \max_{i=0, \dots, m} \left(\Theta_{n,1}^{x_i, a_i}(Z) + \lambda \Theta_{n,2}^{x_i, a_i}(Z) \right)$$

is an L^p -bound as desired. \square

Let us now consider specifically Markov policies $\pi \in \Pi^M$ of the controller. The subspace

$$\mathbb{B} = \{v \in \mathbb{B}_{\mathbf{b}} : v \text{ lower semicontinuous}\}$$

of $(\mathbb{B}_{\mathbf{b}}, \|\cdot\|_{\mathbf{b}})$ turns out to be the set of potential value functions under such policies. $(\mathbb{B}, \|\cdot\|_{\mathbf{b}})$ is a complete metric space since the subset of lower semicontinuous functions is closed in $(\mathbb{B}_{\mathbf{b}}, \|\cdot\|_{\mathbf{b}})$. When we consider intervals $[\underline{v}, \bar{v}] \subseteq \mathbb{B}$ with $\underline{v}, \bar{v} : E \rightarrow \mathbb{R}$ s.t. $\underline{v}(x) \leq \bar{v}(x)$ for all $x \in E$, they are to be understood pointwise

$$[\underline{v}, \bar{v}] = \{v \in \mathbb{B} : \underline{v}(x) \leq v(x) \leq \bar{v}(x) \text{ for all } x \in E\}.$$

Such intervals are closed even w.r.t. pointwise convergence and therefore form a complete metric space as a closed subset of $(\mathbb{B}, \|\cdot\|_{\mathbf{b}})$. In the sequel, the interval

$$I = [\mathbf{b}, \bar{\mathbf{b}}]$$

will be of interest. We define the following operators on $\mathbb{B}_{\mathbf{b}}$ and especially on \mathbb{B} .

Definition 4.6. For $v \in \mathbb{B}_{\mathbf{b}}$ and a Markov decision rule d let

$$\begin{aligned} L_n v(x, a) &= \rho_n \left(c_n(x, a, T_n(x, a, Z_{n+1})) + v(T_n(x, a, Z_{n+1})) \right), & (x, a) \in D_n, \\ \mathcal{T}_{nd} v(x) &= L_n v(x, d(x)), & x \in E, \\ \mathcal{T}_n v(x) &= \inf_{a \in D_n(x)} L_n v(x, a), & x \in E. \end{aligned}$$

Note that the operators are monotone in v . Under a Markov policy $\pi = (d_0, \dots, d_{N-1}) \in \Pi^M$, the value iteration can be expressed with the operators. In order to distinguish from the history-dependent case, we denote policy values here with J . Setting $J_{N\pi}(x) = c_N(x)$, $x \in E$, we obtain for $n = 0, \dots, N-1$ and $x \in E$

$$J_{n\pi}(x) = \rho_n \left(c_n(x, d_n(x), T_n(x, d_n(x), Z_{n+1})) + J_{n+1\pi}(T_n(x, d_n(x), Z_{n+1})) \right) = \mathcal{T}_{nd_n} J_{n+1\pi}(x).$$

Let us further define for $n = 0, \dots, N-1$ the Markov value function

$$J_n(x) = \inf_{\pi \in \Pi^M} J_{n\pi}(x), \quad x \in E.$$

The next result shows that V_n satisfies a Bellman equation and proves that an optimal policy exists and is Markov.

Theorem 4.7. *Let Assumptions 3.1 and 4.2 be satisfied. Then, for $n = 0, \dots, N$, the value function V_n only depends on x_n , i.e. $V_n(h_n) = J_n(x_n)$ for all $h_n \in \mathcal{H}_n$, lies in $I = [\mathbf{b}, \bar{\mathbf{b}}] \subseteq \mathbb{B}$ and satisfies the Bellman equation*

$$\begin{aligned} J_N(x) &= c_N(x), \\ J_n(x) &= \mathcal{T}_n J_{n+1}(x), \quad x \in E. \end{aligned}$$

Furthermore, for $n = 0, \dots, N-1$ there exist Markov decision rules d_n^* such that $\mathcal{T}_n J_{n+1} = \mathcal{T}_{nd_n^*} J_{n+1}$ and every sequence of such minimizers constitutes an optimal policy $\pi = (d_0^*, \dots, d_{N-1}^*)$.

Proof. The proof is by backward induction. At time N we have $V_N = J_N = c_N$ which is in \mathbb{B} by Assumptions 3.1 (iii) and 4.2 (i). Assuming the assertion holds at time $n+1$, we obtain for

time n :

$$\begin{aligned}
V_n(h_n) &= \inf_{\pi \in \Pi} V_{n\pi}(h_n) \\
&= \inf_{\pi \in \Pi} \rho_n \left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1\pi}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) \right) \\
&\geq \inf_{\pi \in \Pi} \rho_n \left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + V_{n+1}(h_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) \right) \\
&= \inf_{\pi \in \Pi} \rho_n \left(c_n(x_n, d_n(h_n), T_n(x_n, d_n(h_n), Z_{n+1})) + J_{n+1}(T_n(x_n, d_n(h_n), Z_{n+1})) \right) \\
&= \inf_{a_n \in D(x_n)} \rho_n \left(c_n(x_n, a_n, T_n(x_n, a_n, Z_{n+1})) + J_{n+1}(T_n(x_n, a_n, Z_{n+1})) \right). \tag{4.2}
\end{aligned}$$

The last equality holds since the minimization does not depend on the entire policy but only on $a_n = d_n(h_n)$. Here, objective and constraint depend on the history of the process only through x_n . Thus, given existence of a minimizing Markov decision rule d_n^* , (4.2) equals $\mathcal{T}_{nd_n^*} J_{n+1}(x_n)$. Again by the induction hypothesis there exists an optimal Markov policy $\pi^* \in \Pi^M$ such that $J_{n+1} = J_{n+1\pi^*}$. Hence, we have

$$V_n(h_n) \geq \mathcal{T}_{nd_n^*} J_{n+1}(x_n) = \mathcal{T}_{nd_n^*} J_{n+1\pi^*}(x_n) = J_{n\pi^*}(x_n) \geq J_n(x_n) \geq V_n(h_n).$$

It remains to show the existence of a minimizing Markov decision rule d_n^* and that $J_n \in \mathbb{B}$. We want to apply Proposition 2.4.3 in [9]. The set-valued mapping $E \ni x \mapsto D_n(x)$ is compact-valued and upper semicontinuous. Next, we show that $D_n \ni (x, a) \mapsto L_n v(x, a)$ is lower semicontinuous for every $v \in \mathbb{B}$. Let $\{(x_k, a_k)\}_{k \in \mathbb{N}}$ be a convergent sequence in D_n with limit $(x^*, a^*) \in D_n$. The function $D_n \ni (x, a) \mapsto c_n(x, a, T_n(x, a, Z_{n+1}(\omega))) + v(T_n(x, a, Z_{n+1}(\omega)))$ is lower semicontinuous for every $\omega \in \Omega$ as a composition of a continuous and a lower semicontinuous one. Consequently,

$$\begin{aligned}
&\liminf_{k \rightarrow \infty} \inf_{\ell \geq k} c_n(x_\ell, a_\ell, T_n(x_\ell, a_\ell, Z_{n+1})) + v(T_n(x_\ell, a_\ell, Z_{n+1})) \\
&= \liminf_{k \rightarrow \infty} c_n(x_k, a_k, T_n(x_k, a_k, Z_{n+1})) + v(T_n(x_k, a_k, Z_{n+1})) \\
&\geq c_n(x^*, a^*, T_n(x^*, a^*, Z_{n+1})) + v(T_n(x^*, a^*, Z_{n+1})). \tag{4.3}
\end{aligned}$$

The sequence $\{C_k\}_{k \in \mathbb{N}}$ with

$$C_k(\omega) = \inf_{\ell \geq k} c_n(x_\ell, a_\ell, T_n(x_\ell, a_\ell, Z_{n+1})) + v(T_n(x_\ell, a_\ell, Z_{n+1}))$$

is measurable as the ω -wise infimum of a countable number of random variables and increasing for every $\omega \in \Omega$. By Lemma 4.5, there exists a nonnegative random variable $\bar{C} \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ such that $|C_k| \leq \bar{C}$ for all $k \in \mathbb{N}$. Hence, $\{C_k\}_{k \in \mathbb{N}}$ converges almost surely to some $C^* \in L^p(\Omega, \mathcal{A}, \mathbb{P})$. The Fatou property of the risk measure ρ_n implies

$$\begin{aligned}
\liminf_{k \rightarrow \infty} L_n v(x_k, a_k) &= \liminf_{k \rightarrow \infty} \rho_n \left(c_n(x_k, a_k, T_n(x_k, a_k, Z_{n+1})) + v(T_n(x_k, a_k, Z_{n+1})) \right) \\
&\geq \liminf_{k \rightarrow \infty} \rho_n(C_k) \\
&\geq \rho_n(C^*) \\
&\geq \rho_n \left(c_n(x^*, a^*, T_n(x^*, a^*, Z_{n+1})) + v(T_n(x^*, a^*, Z_{n+1})) \right) \\
&= L_n v(x^*, a^*).
\end{aligned}$$

The last inequality follows from (4.3) and the monotonicity of ρ_n . So we have shown the lower semicontinuity of $D_n \ni (x, a) \mapsto L_n v(x, a)$. Proposition 2.4.3 in [9] yields the existence of a minimizing Markov decision rule d_n^* and that $J_n = \mathcal{T} J_{n+1}$ is lower semicontinuous. Furthermore, J_n is bounded by \mathbf{b} and $\bar{\mathbf{b}}$ according to Assumption 4.2 (i). Thus, $J_n \in I$ and the proof is complete. \square

5. COST MINIMIZATION UNDER AN INFINITE PLANNING HORIZON

In this section, we consider the risk-sensitive recursive cost minimization problem with an infinite planning horizon. This is reasonable if the terminal period is unknown or if one wants to approximate a model with a large but finite planning horizon. Solving the infinite horizon problem will turn out to be easier since it admits a stationary optimal policy. We study the stationary version of the decision model with no terminal cost. Therefore, the risk measure may no longer vary over time. We also require coherence as an additional property. Recall that if ρ is finite on $L^p(\Omega, \mathcal{A}, \mathbb{P})$, the Fatou property is already implied by coherence. Within the class of distortion risk measures requiring coherence essentially means a restriction to spectral risk measures. For spectral risk measures, finiteness is guaranteed if the spectrum ϕ lies in L^q . Due to coherence we can work with local bounding functions, see Lemma 4.3. We will see that if the one-stage cost function is bounded, coherence can be dropped as a requirement on the risk measure. Then, all distortion risk measures with the Fatou property are admissible. For clarity, all assumptions of this section are summarized below.

Assumption 5.1. (i) There exist $\alpha, \underline{\epsilon}, \bar{\epsilon} \geq 0$ with $\underline{\epsilon} + \bar{\epsilon} = 1$ and measurable functions $\underline{b} : E \rightarrow (-\infty, -\underline{\epsilon}]$, $\bar{b} : E \rightarrow [\bar{\epsilon}, \infty)$ such that for all $(x, a) \in D$

$$\begin{aligned} \rho(c(x, a, T(x, a, Z))) &\geq \underline{b}(x), & \rho(-\bar{b}(T(x, a, Z))) &\leq -\alpha \bar{b}(x), \\ \rho(c(x, a, T(x, a, Z))) &\leq \bar{b}(x), & \rho(\bar{b}(T(x, a, Z))) &\leq \alpha \bar{b}(x). \end{aligned}$$

(ii) We define $b : E \rightarrow [1, \infty)$, $b(x) = \bar{b}(x) - \underline{b}(x)$. For all $(\bar{x}, \bar{a}) \in D$ there exists an $\epsilon > 0$ and measurable functions $\Theta_1^{\bar{x}, \bar{a}}, \Theta_2^{\bar{x}, \bar{a}} : \mathcal{Z} \rightarrow \mathbb{R}_+$ such that $\Theta_1^{\bar{x}, \bar{a}}(Z), \Theta_2^{\bar{x}, \bar{a}}(Z) \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ and

$$|c(x, a, T(x, a, z))| \leq \Theta_1^{\bar{x}, \bar{a}}(z), \quad b(T(x, a, z)) \leq \Theta_2^{\bar{x}, \bar{a}}(z)$$

for all $z \in \mathcal{Z}$ and $(x, a) \in B_\epsilon(\bar{x}, \bar{a}) \cap D$. Here, $B_\epsilon(\bar{x}, \bar{a})$ is the closed ball around (\bar{x}, \bar{a}) w.r.t. an arbitrary product metric on $E \times A$.

(iii) The law-invariant risk measure $\rho : L^p(\Omega, \mathcal{A}, \mathbb{P}) \rightarrow \bar{\mathbb{R}}$ is proper, coherent and has the Fatou property.

(iv) The discount factor β satisfies $\alpha\beta < 1$.

Due to discounting, the global bounding functions corresponding to $\underline{b}, \bar{b}, b$ are given by

$$\underline{\mathbf{b}} = \frac{1}{1 - \alpha\beta} \underline{b}, \quad \bar{\mathbf{b}} = \frac{1}{1 - \alpha\beta} \bar{b} \quad \text{and} \quad \mathbf{b} = \frac{1}{1 - \alpha\beta} b. \quad (5.1)$$

This can be seen as in the proof of Lemma 4.3.

Since the model with infinite planning horizon will be derived as a limit of the one with finite horizon, the consideration can be restricted to Markov policies $\pi = (d_1, d_2, \dots) \in \Pi^M$ due to Theorem 4.7. When calculating limits, it is convenient to index the value functions with the distance to the time horizon rather than the point in time. This is also referred to as *forward form* of the value iteration.

Definition 5.2. For $v \in \mathbb{B}_b$ and a Markov decision rule d let

$$\begin{aligned} \mathcal{T}_d v(x) &= \rho\left(c(x, d(x), T(x, d(x), Z)) + \beta v(T(x, d(x), Z))\right), & x \in E, \\ \mathcal{T} v(x) &= \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z)) + \beta v(T(x, a, Z))\right), & x \in E. \end{aligned}$$

The value of a policy $\pi = (d_0, d_1, \dots) \in \Pi^M$ up to a planning horizon $N \in \mathbb{N}$ now is

$$J_{N\pi}(x) = \mathcal{T}_{d_0} \circ \dots \circ \mathcal{T}_{d_{N-1}} 0(x), \quad x \in E. \quad (5.2)$$

In a non-stationary formulation the discounting is included in the one-stage cost functions and therefore calibrated w.r.t. the fixed reference time zero. If the value functions are considered at a later point in time, the non-stationary and stationary version differ by a discounting factor:

$$J_n^{\text{non-stat}}(x) = \beta^n J_{N-n}^{\text{stat}}(x), \quad x \in E, \quad n = 0, \dots, N.$$

The reformulation (5.2) makes it necessary to write the value iteration in terms of the *shifted policy* $\vec{\pi} = (d_1, d_2, \dots)$ corresponding to $\pi = (d_0, d_1, \dots) \in \Pi^M$:

$$J_{N\pi}(x) = \mathcal{T}_{d_0} J_{N-1\vec{\pi}}(x) = \rho\left(c(x, d_0(x), T(x, d_0(x), Z)) + \beta J_{N-1\vec{\pi}}(T(x, d_0(x), Z))\right), \quad x \in E.$$

The value function under planning horizon $N \in \mathbb{N}$ is given by

$$J_N(x) = \inf_{\pi \in \Pi^M} J_{N\pi}(x), \quad x \in E,$$

By Theorem 4.7, the value function satisfies the Bellman equation

$$J_N(x) = \mathcal{T} J_{N-1}(x) = \mathcal{T}^N 0(x), \quad x \in E. \quad (5.3)$$

When the planning horizon is infinite, we define the value of a policy $\pi \in \Pi^M$ as

$$J_{\infty\pi}(x) = \lim_{N \rightarrow \infty} J_{N\pi}(x), \quad x \in E. \quad (5.4)$$

Hence, the optimality criterion considered in this section is

$$J_{\infty}(x) = \inf_{\pi \in \Pi^M} J_{\infty\pi}(x), \quad x \in E. \quad (5.5)$$

The next lemma shows that the infinite horizon policy value (5.4) and value function (5.5) are well-defined.

Lemma 5.3. *Under Assumption 5.1, the sequence $\{J_{N\pi}\}_{N \in \mathbb{N}}$ converges pointwise for every Markov policy $\pi \in \Pi^M$ and the limit $J_{\infty\pi}$ is bounded by $\underline{\mathbf{b}}$ and $\bar{\mathbf{b}}$.*

Proof. First, we show by induction that for all $N \in \mathbb{N}$

$$J_{N\pi}(x) \geq J_{N-1\pi}(x) + (\alpha\beta)^{N-1} \underline{\mathbf{b}}(x), \quad x \in E. \quad (5.6)$$

For $N = 1$ it holds by Assumption 5.1 (i) that $J_{1\pi}(x) \geq \underline{\mathbf{b}}(x) = J_{0\pi}(x) + (\alpha\beta)^0 \underline{\mathbf{b}}(x)$. For $N \geq 2$ it follows

$$\begin{aligned} J_{N\pi}(x) &= \rho\left(c(x, d_0(x), T(x, d_0(x), Z)) + \beta J_{N-1\vec{\pi}}(T(x, d_0(x), Z))\right) \\ &\geq \rho\left(c(x, d_0(x), T(x, d_0(x), Z)) + \beta J_{N-2\vec{\pi}}(T(x, d_0(x), Z)) + \beta(\alpha\beta)^{N-2} \underline{\mathbf{b}}(T(x, d_0(x), Z))\right) \\ &\geq \rho\left(c(x, d_0(x), T(x, d_0(x), Z)) + \beta J_{N-2\vec{\pi}}(T(x, d_0(x), Z))\right) \\ &\quad - \beta(\alpha\beta)^{N-2} \rho\left(-\underline{\mathbf{b}}(T(x, d_0(x), Z))\right) \\ &\geq \rho\left(c(x, d_0(x), T(x, d_0(x), Z)) + \beta J_{N-2\vec{\pi}}(T(x, d_0(x), Z))\right) + (\alpha\beta)^{N-1} \underline{\mathbf{b}}(x) \\ &= J_{N-1\pi}(x) + (\alpha\beta)^{N-1} \underline{\mathbf{b}}(x). \end{aligned}$$

The first inequality is by the induction hypothesis, the second one is by Lemma 2.5 together with the positive homogeneity of ρ and the third one is due to Assumption 5.1 (i). Thus, (5.6) holds. Applying this inequality repeatedly for $N, N-1, \dots, m$ yields

$$J_{N\pi}(x) \geq J_{m\pi}(x) + \sum_{k=m}^{N-1} (\alpha\beta)^k \underline{\mathbf{b}}(x) \geq J_{m\pi}(x) + \sum_{k=m}^{\infty} (\alpha\beta)^k \underline{\mathbf{b}}(x), \quad (5.7)$$

where $\delta_m(x) = \sum_{k=m}^{\infty} (\alpha\beta)^k \underline{\mathbf{b}}(x)$ are non-positive functions with $\lim_{m \rightarrow \infty} \delta_m(x) = 0$. Hence, the sequence of functions $\{J_{N\pi}\}_{N \in \mathbb{N}}$ is weakly increasing and therefore convergent to a limit function $J_{\infty\pi}$ by Lemma A.1.4 in [9]. The global bounds (5.1) also apply to the limit $J_{\infty\pi}$. \square

Lemma 5.4. *Given Assumption 5.1, the Bellman operator \mathcal{T} is a contraction on $I = [\underline{\mathbf{b}}, \bar{\mathbf{b}}]$ with modulus $\alpha\beta \in (0, 1)$.*

Proof. Let $v \in I$. It has been established in the proof of Theorem 4.7 that $\mathcal{T}v$ is lower semicontinuous. Furthermore,

$$\begin{aligned} \mathcal{T}v(x) &\geq \mathcal{T}\mathbf{b}(x) = \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z)) + \frac{\beta}{1 - \alpha\beta} \underline{b}(T(x, a, Z))\right) \\ &\geq \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z))\right) - \frac{\beta}{1 - \alpha\beta} \rho\left(-\underline{b}(T(x, a, Z))\right) \geq \underline{b}(x) + \frac{\alpha\beta}{1 - \alpha\beta} \underline{b}(x) = \mathbf{b}(x). \end{aligned}$$

The second inequality is by Lemma 2.5 together with the positive homogeneity of ρ and the third one is due to Assumption 5.1 (i). Regarding the upper bounding function one can argue similarly, using the subadditivity of ρ instead of Lemma 2.5:

$$\begin{aligned} \mathcal{T}v(x) &\leq \mathcal{T}\bar{\mathbf{b}}(x) = \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z)) + \frac{\beta}{1 - \alpha\beta} \bar{b}(T(x, a, Z))\right) \\ &\leq \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z))\right) + \frac{\beta}{1 - \alpha\beta} \rho\left(\bar{b}(T(x, a, Z))\right) \leq \bar{b}(x) + \frac{\alpha\beta}{1 - \alpha\beta} \bar{b}(x) = \bar{\mathbf{b}}(x). \end{aligned}$$

Hence, the operator \mathcal{T} is an endofunction on I and it remains to verify the Lipschitz constant $\alpha\beta$. For $v_1, v_2 \in I$ it holds

$$\begin{aligned} |\mathcal{T}v_1(x) - \mathcal{T}v_2(x)| &\leq \sup_{a \in D(x)} |Lv_1(x, a) - Lv_2(x, a)| \\ &\leq \beta \sup_{a \in D(x)} \rho\left(|v_1(T(x, a, Z)) - v_2(T(x, a, Z))|\right) \\ &\leq \beta \sup_{a \in D(x)} \rho\left(\|v_1 - v_2\|_b \bar{b}(T(x, a, Z))\right) \\ &= \beta \|v_1 - v_2\|_b \sup_{a \in D(x)} \rho\left(\bar{b}(T(x, a, Z)) - \underline{b}(T(x, a, Z))\right) \\ &\leq \beta \|v_1 - v_2\|_b \sup_{a \in D(x)} \left[\rho\left(\bar{b}(T(x, a, Z))\right) + \rho\left(-\underline{b}(T(x, a, Z))\right)\right] \\ &\leq \alpha\beta \|v_1 - v_2\|_b \left[\bar{b}(x) - \underline{b}(x)\right] \\ &= \alpha\beta \|v_1 - v_2\|_b b(x). \end{aligned}$$

Dividing by $b(x)$ and taking the supremum over $x \in E$ on the left hand side completes the proof. Note that the second inequality is by Lemma 2.3, the fourth one due to the subadditivity of ρ and the last one by Assumption 5.1 (i). \square

Under a finite planning horizon $N \in \mathbb{N}$ we have characterized the value function with the Bellman equation (5.3). We will show that this is compatible with the optimality criterion of the infinite horizon model (5.5). To this end, we define the *limit value function*

$$J(x) = \lim_{N \rightarrow \infty} J_N(x), \quad x \in E.$$

Note that the limit exists since it follows from (5.7) that $J_N \geq J_m + \delta_m$ for all $N \geq m$ which implies the convergence.

Theorem 5.5. *Let Assumptions 3.1 and 5.1 be satisfied. Then it holds:*

- a) *The limit value function J is the unique fixed point of the Bellman operator \mathcal{T} in $I = [\mathbf{b}, \bar{\mathbf{b}}]$.*
- b) *There exists a Markov decision rule d^* such that $\mathcal{T}_{d^*} J = \mathcal{T}J$.*
- c) *Each stationary policy $\pi^* = (d^*, d^*, \dots)$ induced by a Markov decision rule d^* as in b) is optimal for optimization problem (5.5) and it holds $J_\infty = J$.*

Proof. a) The fact that J is the unique fixed point of the operator \mathcal{T} in I follows directly from Banach's Fixed Point Theorem using Lemma 5.4.

- b) The existence of a minimizing Markov decision rule follows from the respective result in the finite horizon case, cf. Theorem 4.7.

c) Let d^* be a Markov decision rule as in part b) and $\pi^* = (d^*, d^*, \dots)$. Then it holds

$$J(x) \leq J_\infty(x) \leq J_{\infty\pi^*}(x), \quad x \in E.$$

The second inequality holds by definition. Regarding the first one note that for any $\pi \in \Pi^M$ we have $J_N(x) \leq J_{N\pi}(x)$ for all $N \in \mathbb{N}_0$. Letting $N \rightarrow \infty$ yields $J(x) \leq J_{\infty\pi}(x)$. Since $\pi \in \Pi^M$ was arbitrary we get $J(x) \leq \inf_{\pi \in \Pi^M} J_{\infty\pi}(x) = J_\infty(x)$. It remains to show

$$J_{\infty\pi^*}(x) \leq J(x), \quad x \in E. \quad (5.8)$$

To that end, we will prove by induction that for all $N \in \mathbb{N}_0$ and $x \in E$

$$J(x) \geq J_{N\pi^*}(x) + \frac{(\alpha\beta)^N}{1 - \alpha\beta} \mathbf{b}(x). \quad (5.9)$$

Letting $N \rightarrow \infty$ in (5.9) yields (5.8) and concludes the proof. For $N = 0$ equation (5.9) reduces to $J(x) \geq \frac{1}{1 - \alpha\beta} \mathbf{b}(x) = \mathbf{b}(x)$, which holds by part a). For $N \geq 1$ the induction hypothesis yields

$$\begin{aligned} J(x) &= \mathcal{T}_{d^*} J(x) \geq \mathcal{T}_{d^*} \left(J_{N-1\pi^*} + \frac{(\alpha\beta)^{N-1}}{1 - \alpha\beta} \mathbf{b} \right) (x) \\ &\geq \rho \left(c(x, d^*(x), T(x, d^*(x), Z)) + \beta J_{N-1\pi^*}(T(x, d^*(x), Z)) \right) \\ &\quad - \beta \frac{(\alpha\beta)^{N-1}}{1 - \alpha\beta} \rho \left(-\mathbf{b}(T(x, d^*(x), Z)) \right) \\ &\geq J_{N\pi^*}(x) + \frac{(\alpha\beta)^N}{1 - \alpha\beta} \mathbf{b}(x). \end{aligned}$$

The second inequality is by Lemma 2.5 together with the positive homogeneity of ρ and the last one is by Assumption 5.1 (i). \square

Let us now consider the special case that the one-stage cost is bounded, i.e.

(B) there exist $\underline{b} \in \mathbb{R}_-$ and $\bar{b} \in \mathbb{R}_+$ such that $b = \bar{b} - \underline{b} > 0$ and $\underline{b} \leq c(x, a, T(x, a, Z)) \leq \bar{b}$ \mathbb{P} -f.s. for all $(x, a) \in D$.

Then, Assumption 5.1 (i) is satisfied with $\alpha = 1$ and part (ii) is obvious. Part (iv) of the assumption reduces to $\beta < 1$.

Corollary 5.6. *Given (B), Lemmata 5.3, 5.4 and in case Assumption 3.1 is satisfied, Theorem 5.5 hold for any normalized monetary risk measure with the Fatou property.*

Proof. The steps in the proofs that were justified by Lemma 2.5, subadditivity, positive homogeneity or Assumption 5.1 (i) now hold due to translation invariance and normalization. Nothing else has to be changed. \square

6. CONNECTION TO DISTRIBUTIONALLY ROBUST MDP

We consider the stationary version of the decision model with no terminal cost under both finite and infinite horizon in this section. If the planning horizon is finite, stationarity is only assumed for convenience and everything can be transferred to a non-stationary setting purely by notational changes. Let the risk measure ρ be proper and coherent with the Fatou property. By inserting the dual representation of Proposition 2.4 in the Bellman equation, we get

$$\begin{aligned} J_N(x) &= 0, \\ J_n(x) &= \inf_{a \in D(x)} \sup_{\mathbb{Q} \in \mathcal{Q}} \mathbb{E}^{\mathbb{Q}} \left[c(x, a, T(x, a, Z)) + \beta J_{n+1}(T(x, a, Z)) \right], \quad x \in E, \end{aligned}$$

i.e. the Bellman equation of a distributionally robust MDP as considered in [5]. Under some minor technical assumptions we have indeed a special case of the distributionally robust MDP and thus obtain a global interpretation of the recursively defined risk-sensitive optimality criterion.

Due to the independence of the disturbances we can w.l.o.g. assume that the underlying probability space has a product structure $(\Omega, \mathcal{A}, \mathbb{P}) = \bigotimes_{n=1}^{\infty} (\Omega_1, \mathcal{A}_1, \mathbb{P}_1)$ with $Z_n(\omega) = Z_n(\omega_n)$ only depending on component ω_n of $\omega = (\omega_1, \omega_2, \dots) \in \Omega$. For a policy $\pi = (d_0, d_1, \dots) \in \Pi^M$ of the controller and $\gamma = (\gamma_0, \gamma_1, \dots)$, where $\gamma_n : D \rightarrow \mathcal{Q}$ is measurable, we define the transition kernel

$$Q_n^{\pi\gamma}(B|x, a) = \int \mathbb{1}_B(T(x, d_n(x), Z(\omega))) \gamma_n(d\omega|x, d_n(x)), \quad B \in \mathcal{B}(E), \quad x \in E,$$

and the law of motion $\mathbb{Q}_x^{\pi\gamma} = \delta_x \otimes Q_0^{\pi\gamma} \otimes Q_1^{\pi\gamma} \otimes \dots$. The set of all possible laws of motion under policy $\pi \in \Pi^M$ is denoted by $\mathfrak{Q}_\pi = \{\mathbb{Q}_x^{\pi\gamma} : \gamma \in \Gamma\}$ with Γ being the set of all possible γ .

Theorem 6.1. *Let Assumption 5.1 be fulfilled with the following tightening in part (i):*

$$\rho(c^-(x, a, T(x, a, Z))) \leq -b(x), \quad \rho(c^+(x, a, T(x, a, Z))) \leq \bar{b}(x), \quad (x, a) \in D.$$

Furthermore, let the underlying probability space have a product structure as above and let the probability measure \mathbb{P}_1 on $(\Omega_1, \mathcal{A}_1)$ be separable. Then, for $N \in \mathbb{N} \cup \{\infty\}$ it holds

$$J_N(x) = \inf_{\pi \in \Pi^M} \sup_{\mathbb{Q} \in \mathfrak{Q}_\pi} \mathbb{E}^{\mathbb{Q}} \left[\sum_{k=0}^{N-1} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right] \quad (6.1)$$

Proof. We need to verify Assumption 3.1 in [5]. Then, the assertion follows from Theorem 3.10 therein for the finite horizon case and from Theorem 4.18 in [16] for the infinite horizon case. Part (i) holds since we have for all $\mathbb{Q} \in \mathfrak{Q}$ and $(x, a) \in D$

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}} [-c^-(x, a, T(x, a, Z))] &\geq \inf_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [-c^-(x, a, T(x, a, Z))] = - \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [c^-(x, a, T(x, a, Z))] \\ &= -\rho(c^-(x, a, T(x, a, Z))) \geq b(x), \\ \mathbb{E}^{\mathbb{Q}} [b(T(x, a, Z))] &\geq \inf_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [b(T(x, a, Z))] = - \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [-b(T(x, a, Z))] \\ &= -\rho(-b(T(x, a, Z))) \geq \alpha b(x), \\ \mathbb{E}^{\mathbb{Q}} [c^+(x, a, T(x, a, Z))] &\leq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [c^+(x, a, T(x, a, Z))] = \rho(c^+(x, a, T(x, a, Z))) \leq \bar{b}(x), \\ \mathbb{E}^{\mathbb{Q}} [\bar{b}(T(x, a, Z))] &\leq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}} [\bar{b}(T(x, a, Z))] = \rho(\bar{b}(T(x, a, Z))) \leq \alpha \bar{b}(x). \end{aligned}$$

Part (ii) equals Assumption 5.1 (ii). Finally, part (iii) holds since $\rho(X) = \max_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}^{\mathbb{Q}}[X]$ by Proposition 2.4 where $\mathfrak{Q} \subseteq \mathcal{M}_1^q(\Omega, \mathcal{A}, \mathbb{P})$ is weak* compact and therefore norm bounded by the Banach-Alaoglu Theorem 6.21 in [2]. \square

It is readily checked that for a fixed policy $\pi \in \Pi^M$ of the controller $\tilde{\rho}(X) = \sup_{\mathbb{Q} \in \mathfrak{Q}_\pi} \mathbb{E}^{\mathbb{Q}}[X]$, $X \in L^p(\Omega, \mathcal{A}, \mathbb{P})$, defines a coherent risk measure. If the stage-wise applied risk measure ρ is spectral and the model data has certain monotonicity properties, one can choose \mathfrak{Q} to be independent of π , cf. Lemma 6.8 and subsequent remarks in [5]. In this case, the recursive minimization of spectral risk measures is equivalent to the minimization of a non-standard coherent risk measure applied to the total cost.

Besides, one can reformulate (6.1) to

$$J_N(x) = \inf_{\pi \in \Pi^M} \sup_{\gamma \in \Gamma} \mathbb{E}^{\pi\gamma} \left[\sum_{k=0}^{N-1} \beta^k c(X_k, d_k(X_k), X_{k+1}) \right]$$

with the interpretation of a Stackelberg game of the controller against a theoretical opponent (nature) selecting the most adverse disturbance distribution in each scenario. Here, $\gamma = (\gamma_0, \gamma_1, \dots)$ with $\gamma_n : D \rightarrow \mathcal{Q}$ is a Markov policy of nature. This game is extensively studied in [5]. From this perspective we get another global interpretation of the recursively defined objective function as robust minimization of the expected total cost.

7. RELAXED ASSUMPTIONS FOR MONOTONE MODELS

The model has been introduced in Section 3 with a general Borel space as state space. However, in many applications the state space is simply \mathbb{R} . In this case, the assumption on the transition function can be relaxed to semicontinuity when the transition and one-stage cost function have some form of monotonicity. For notational convenience, we consider the stationary model with no terminal cost under both finite and infinite horizon in this section. We replace Assumption 3.1 by

- Assumption 7.1.** (i) The state space is the real line $E = \mathbb{R}$.
 (ii) The sets $D(x)$ are compact and $\mathbb{R} \ni x \mapsto D(x)$ is upper semicontinuous and decreasing, i.e. $D(x) \supseteq D(y)$ for $x \leq y$.
 (iii) The transition functions T is lower semicontinuous in (x, a) and increasing in x .
 (iv) The one-stage cost function c is lower semicontinuous in (x, a, x') and increasing in (x, x') .

How do the modified continuity assumptions affect the validity of the results in Sections 4 and 5? Lemmata 4.3, 4.5, 5.3 and 5.4 were proven without using the continuity of T . Thus, only Theorems 4.7 and 5.5 need to be looked at.

Proposition 7.2. *Let the new continuity and monotonicity Assumptions 7.1 be satisfied. Then,*

- a) *under Assumption 4.2, the assertion of Theorem 4.7 remains true.*
- b) *under Assumption 5.1, the assertion of Theorem 5.5 remains true.*

In both cases, the value functions are increasing and the set of potential value functions can be replaced by $\mathbb{B} = \{v \in \mathbb{B}_{\mathbf{b}} : v \text{ lower semicontinuous and increasing}\}$.

Proof. In the proof of Theorem 4.7, the continuity of T is only used to show that $D \ni (x, a) \mapsto Lv(x, a)$ is lower semicontinuous for every $v \in \mathbb{B}$. Due to the monotonicity assumptions,

$$D \ni (x, a) \mapsto c(x, a, T(x, a, Z(\omega))) + \beta v(T(x, a, Z(\omega)))$$

is lower semicontinuous for every $\omega \in \Omega$. Now, the lower semicontinuity of $D \ni (x, a) \mapsto Lv(x, a)$ and the existence of a minimizing decision rule follow as in the proof of Theorem 4.7. The fact that $\mathcal{T}v$ is increasing for every $v \in \mathbb{B}$ follows as in Theorem 2.4.14 in [9]. Theorem 5.5 uses the continuity of T only indirectly through Theorem 4.7. \square

With the real line as state space, a simple separation condition is sufficient for Assumptions 4.2 (ii) or 5.1 (ii).

Lemma 7.3. *Let there be upper semicontinuous functions $\vartheta_1, \vartheta_2 : D \rightarrow \mathbb{R}_+$ and measurable functions $\Theta_1, \Theta_2 : \mathcal{Z} \rightarrow \mathbb{R}_+$ which fulfill $\Theta_1(Z), \Theta_2(Z) \in L^p(\Omega, \mathcal{A}, \mathbb{P})$ and*

$$|c(x, a, T(x, a, z))| \leq \vartheta_1(x, a) + \Theta_1(z), \quad b(T(x, a, z)) \leq \vartheta_2(x, a) + \Theta_2(z)$$

for every $(x, a, z) \in D \times \mathcal{Z}$. Then Assumptions 4.2 (ii) and 5.1 (ii) are satisfied.

Proof. Let $(\bar{x}, \bar{a}) \in D$. We can choose $\epsilon > 0$ arbitrarily. The set $S = [\bar{x} - \epsilon, \bar{x} + \epsilon] \times D(\bar{x} - \epsilon)$ is compact w.r.t. the product topology by the Tychonoff Product Theorem 2.61 in [2]. Moreover, $B_\epsilon(\bar{x}, \bar{a}) \cap D \subseteq S$ since the set-valued mapping $D(\cdot)$ is decreasing. Due to upper semicontinuity there exist $(x_i, a_i) \in S$ such that $\vartheta_i(x_i, a_i) = \sup_{(x,a) \in S} \vartheta_i(x, a)$, $i = 1, 2$. Hence, one can define

$$\Theta_i^{\bar{x}, \bar{a}}(\cdot) = \vartheta_i(x_i, a_i) + \Theta_i(\cdot), \quad i = 1, 2$$

and Assumptions 4.2 (ii) and 5.1 (ii) are satisfied. \square

A monotone model not only allows for weaker assumptions on the transition function, but also requirements regarding the risk measure may be relaxed. In the following, we study two such cases: local bounding and infinite horizon cost minimization with bounded below cost.

Firstly, the existence of a global upper and lower bounding function can be guaranteed by suitable local bounding functions as in Lemma 4.3. However due to the monotonicity properties of the model, the risk measure does not need to be coherent. E.g. spectral risk measures can be replaced by other distortion risk measures.

Lemma 7.4. *Let Assumption 7.1 be satisfied and the monetary risk measure ρ be positive homogeneous and comonotonic additive. If there exist $\underline{\epsilon}, \bar{\epsilon} \geq 0$ with $\underline{\epsilon} + \bar{\epsilon} = 1$, increasing functions $\underline{b} : \mathbb{R} \rightarrow (-\infty, -\underline{\epsilon}]$, $\bar{b} : \mathbb{R} \rightarrow [\bar{\epsilon}, \infty)$ and a constant $\alpha > 0$ such that $\alpha\beta \in (0, 1)$ and*

$$\begin{aligned} \rho(c(x, a, T(x, a, Z))) &\geq \underline{b}(x), & \rho(\underline{b}(T(x, a, Z))) &\geq \alpha\underline{b}(x), \\ \rho(c(x, a, T(x, a, Z))) &\leq \bar{b}(x), & \rho(\bar{b}(T(x, a, Z))) &\leq \alpha\bar{b}(x), \end{aligned}$$

for all $(x, a) \in D$, then

$$\mathbf{b} = \frac{1}{1 - \alpha\beta} \underline{b} \quad \text{and} \quad \bar{\mathbf{b}} = \frac{1}{1 - \alpha\beta} \bar{b}$$

are global lower/ upper bounding functions and Assumption 4.2 (i) holds.

Proof. We proceed by backward induction. At time N there is nothing to show. Assuming the assertion holds at time $n + 1$, it follows for time n :

$$\begin{aligned} V_{n\pi}(h_n) &= \rho\left(c(x_n, d_n(h_n), T(x_n, d_n(h_n), Z)) + \beta V_{n+1\pi}(h_n, d_n(h_n), T(x_n, d_n(h_n), Z))\right) \\ &\geq \rho\left(c(x_n, d_n(h_n), T(x_n, d_n(h_n), Z)) + \frac{\beta}{1 - \alpha\beta} \underline{b}(T(x_n, d_n(h_n), Z))\right) \\ &= \rho\left(c(x_n, d_n(h_n), T(x_n, d_n(h_n), Z))\right) + \frac{\beta}{1 - \alpha\beta} \rho\left(\underline{b}(T(x_n, d_n(h_n), Z))\right) \\ &\geq \underline{b}(x_n) + \frac{\alpha\beta}{1 - \alpha\beta} \underline{b}(x_n) = \mathbf{b}(x_n), \end{aligned}$$

$\pi \in \Pi$, $h_n \in \mathcal{H}_n$. The second equality is by the comonotonic additivity and positive homogeneity of ρ . Regarding the upper bounding function one argues analogously. \square

In Lemma 7.4, the local bounding functions are assumed to be increasing, which was not necessary in Lemma 4.3. Also note that we only have to require $\rho(\underline{b}(T(x, a, Z))) \geq \alpha\underline{b}(x)$, $(x, a) \in D$ which is weaker than the corresponding assumption for the model with general state space, cf. Lemma 4.3 and Remark 4.4.

As a second example, where the assumptions on the risk measure can be relaxed, we consider infinite horizon cost minimization with bounded below cost. For absolutely bounded cost functions we already showed in Corollary 5.6 that a coherent risk measure is not necessary to solve the infinite horizon problem. This result is very general regarding the risk measure but very restrictive concerning the one-stage cost. The monotone model allows for a middle course.

(B⁻) There exist $\underline{b} \leq 0$, $\bar{\epsilon} \geq 0$ and $\alpha \geq 1$ with $\bar{\epsilon} - \underline{b} = 1$ and an increasing function $\bar{b} : \mathbb{R} \rightarrow [\bar{\epsilon}, \infty)$ such that $c(x, a, T(x, a, Z)) \geq \underline{b}$ \mathbb{P} -f.s. and

$$\rho(c(x, a, T(x, a, Z))) \leq \bar{b}(x), \quad \rho(\bar{b}(T(x, a, Z))) \leq \alpha\bar{b}(x).$$

for all $(x, a) \in D$.

W.l.o.g. we assume $\alpha \geq 1$ since then $\rho(-\underline{b}) = -\underline{b} \leq \alpha\underline{b}$ due to translation invariance and normalization. Otherwise one would need separate alphas for the lower and upper local bounding function. If the risk measure is comonotonic additive and positive homogeneous, the objective function is globally bounded under (B⁻) due to Lemma 7.4 and Theorem 4.7 remains true. Under an infinite planning horizon, the assertion of Theorem 5.5 can be proven without requiring a coherent risk measure. When we refer to the interval $I = [\underline{\mathbf{b}}, \bar{\mathbf{b}}]$ in the following, it is to be understood as a subset of the modified function space \mathbb{B} as in Proposition 7.2.

Proposition 7.5. *Let Assumptions 7.1 and 5.1 be satisfied with the modification that part (i) is replaced by (B⁻) and part (iii) by the requirement that ρ is a law invariant, comonotonic additive and positive homogeneous monetary risk measure with the Fatou property. Then it holds:*

- a) *The sequence $\{J_{N\pi}\}_{N \in \mathbb{N}}$ converges pointwise for every Markov policy $\pi \in \Pi^M$ and the limit function $J_{\infty\pi}$ is bounded by $\underline{\mathbf{b}}$ and $\bar{\mathbf{b}}$.*

- b) The Bellman operator \mathcal{T} is a contraction on I with modulus $\alpha\beta \in (0, 1)$ and the limit value function J is the unique fixed point of \mathcal{T} in I .
- c) There exists a Markov decision rule d^* such that $\mathcal{T}_{d^*}J = \mathcal{T}J$, each stationary policy $\pi^* = (d^*, d^*, \dots)$ induced by such a Markov decision rule is optimal for optimization problem (5.5) and it holds $J_\infty = J$.

Proof. a) We show by induction that for all $N \in \mathbb{N}$

$$J_{N\pi}(x) \geq J_{N-1\pi}(x) + (\alpha\beta)^{N-1}\underline{b}, \quad x \in \mathbb{R}. \quad (7.1)$$

For $N = 1$ it holds due to (B^-) that $J_{1\pi}(x) \geq \underline{b} = J_{0\pi}(x) + (\alpha\beta)^0\underline{b}$. For $N \geq 2$ it follows with the monotonicity and translation invariance of ρ that

$$\begin{aligned} J_{N\pi}(x) &= \mathcal{T}_{d_0}J_{N-1\pi}(x) \geq \mathcal{T}_{d_0}(J_{N-2\pi} + (\alpha\beta)^{N-2}\underline{b})(x) = \mathcal{T}_{d_0}J_{N-2\pi}(x) + \beta(\alpha\beta)^{N-2}\underline{b} \\ &\geq \mathcal{T}_{d_0}J_{N-2\pi}(x) + (\alpha\beta)^{N-1}\underline{b} = J_{N-1\pi}(x) + (\alpha\beta)^{N-1}\underline{b}. \end{aligned}$$

Thus, (7.1) holds. Applying this inequality repeatedly for $N, N-1, \dots, m$ yields

$$J_{N\pi}(x) \geq J_{m\pi}(x) + \sum_{k=m}^{N-1} (\alpha\beta)^k \underline{b} \geq J_{m\pi}(x) + \sum_{k=m}^{\infty} (\alpha\beta)^k \underline{b}.$$

Since $\sum_{k=m}^{\infty} (\alpha\beta)^k \underline{b}$ is non-positive and converges to zero as $m \rightarrow \infty$, the sequence $\{J_{N\pi}\}_{N \in \mathbb{N}}$ is weakly increasing and hence convergent to a limit $J_{\infty\pi}$ by Lemma A.1.4 in [9]. Clearly, the global bounds $\underline{b}, \bar{b}(\cdot)$ also apply to the limit $J_{\infty\pi}$.

- b) Let $v \in I$. Due to Proposition 7.2 $\mathcal{T}v$ is increasing and lower semicontinuous. Furthermore, the monotonicity and translation invariance of ρ imply

$$\mathcal{T}v(x) \geq \mathcal{T}\underline{b}(x) = \mathcal{T}0(x) + \underline{b} \geq \underline{b} + \frac{\alpha\beta}{1-\alpha\beta}\underline{b} = \underline{b}.$$

Regarding the upper bounding function it follows from the comonotonic additivity and positive homogeneity of ρ that

$$\begin{aligned} \mathcal{T}v(x) &\leq \mathcal{T}\bar{b}(x) = \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z)) + \frac{\beta}{1-\alpha\beta}\bar{b}(T(x, a, Z))\right) \\ &= \inf_{a \in D(x)} \rho\left(c(x, a, T(x, a, Z))\right) + \frac{\beta}{1-\alpha\beta}\rho\left(\bar{b}(T(x, a, Z))\right) \\ &\leq \bar{b}(x) + \frac{\alpha\beta}{1-\alpha\beta}\bar{b}(x) = \bar{b}(x). \end{aligned}$$

I.e. \mathcal{T} is an endofunction on I and it remains to verify the Lipschitz constant. For $v_1, v_2 \in I$ it holds

$$\begin{aligned} &\mathcal{T}v_1(x) - \mathcal{T}v_2(x) \\ &\leq \sup_{a \in D(x)} Lv_1(x, a) - Lv_2(x, a) = \beta \sup_{a \in D(x)} \rho\left(v_1(T(x, a, Z))\right) - \rho\left(v_2(T(x, a, Z))\right) \\ &= \beta \sup_{a \in D(x)} \rho\left(v_1(T(x, a, Z)) - v_2(T(x, a, Z)) + v_2(T(x, a, Z))\right) - \rho\left(v_2(T(x, a, Z))\right) \\ &\leq \beta \sup_{a \in D(x)} \rho\left(\|v_1 - v_2\|_b b(T(x, a, Z)) + v_2(T(x, a, Z))\right) - \rho\left(v_2(T(x, a, Z))\right) \\ &= \|v_1 - v_2\|_b \beta \sup_{a \in D(x)} \rho\left(b(T(x, a, Z))\right) = \|v_1 - v_2\|_b \beta \sup_{a \in D(x)} \left[\rho\left(\bar{b}(T(x, a, Z))\right) - \underline{b}\right] \\ &\leq \alpha\beta \|v_1 - v_2\|_b [\bar{b}(x) - \underline{b}] = \alpha\beta \|v_1 - v_2\|_b b(x). \end{aligned}$$

The first equality is by comonotonic additivity and positive homogeneity. Since \underline{b} is constant, $b(\cdot) = \bar{b}(\cdot) - \underline{b}$ is an increasing function and so is v_2 . Therefore, the third

equality is again by comonotonic additivity. The last inequality is by (B^-) using $\alpha \geq 1$. Interchanging the roles of v_1 and v_2 yields

$$|\mathcal{T}v_1(x) - \mathcal{T}v_2(x)| \leq \alpha\beta\|v_1 - v_2\|_b b(x).$$

Finally, dividing by $b(x)$ and taking the supremum over $x \in \mathbb{R}$ shows that \mathcal{T} is a contraction and Banach's Fixed Point Theorem yields the assertion.

- c) The existence of a minimizing Markov decision rule follows from Proposition 7.2. With the same argument as in the proof of Theorem 5.5, the relation $J \leq J_\infty \leq J_{\infty\pi}$ holds for any policy and it remains to show that $J_{\infty\pi^*} \leq J$ for the specific policy π^* . To that end, we will prove by induction that $J \geq J_{N\pi^*} + (\alpha\beta)^N \mathbf{b}$ for all $N \in \mathbb{N}_0$. Then, letting $N \rightarrow \infty$ concludes the proof. The case $N = 0$, i.e. $J(x) \geq \frac{1}{1-\alpha\beta} \mathbf{b}$, holds by part b). For $N \geq 1$ we have

$$\begin{aligned} J(x) &= \mathcal{T}_{d^*} J(x) \geq \mathcal{T}_{d^*} (J_{N-1\pi^*} + (\alpha\beta)^{N-1} \mathbf{b})(x) = \mathcal{T}_{d^*} J_{N-1\pi^*}(x) + \beta(\alpha\beta)^{N-1} \mathbf{b} \\ &\geq \mathcal{T}_{d^*} J_{N-1\pi^*}(x) + (\alpha\beta)^N \mathbf{b} = J_{N\pi^*}(x) + (\alpha\beta)^N \mathbf{b}. \end{aligned}$$

The first inequality is by the induction hypothesis and the monotonicity of ρ , the equality thereafter is by translation invariance and the second inequality holds since $\alpha \geq 1$. \square

8. EXAMPLES

In this section, we present some applications of the results in the previous sections. In particular we show that often structural results about optimal policies which are known from the classical iterated expectation case still hold under more general risk measures.

Example 8.1 (Value-at-Risk is Myopic in Monotone Models). In a monotone model as in Section 7, where the one-stage cost function does not depend on the controller's action, i.e. $c(x, a, x') = c(x, x')$, recursive decision making with Value-at-Risk is myopic. This can be seen as follows. Let Assumptions 7.1 and 4.2 (i),(ii) be satisfied. The Bellman equation here reads

$$\begin{aligned} J_N(x) &= 0 \\ J_n(x) &= \inf_{a \in D(x)} \text{VaR}_\alpha (c(x, T(x, a, Z)) + \beta J_{n+1}(T(x, a, Z))), \quad n = 0, \dots, N-1. \end{aligned}$$

We can now interchange VaR with the increasing lower semicontinuous (i.e. left-continuous) function $h(x') = c(x, x') + \beta J_{n+1}(x')$, $x' \in \mathbb{R}$ by properties of the quantile function (see e.g. Proposition 2.2 in [4]). Doing this we obtain

$$J_n(x) = \inf_{a \in D(x)} h(\text{VaR}_\alpha(T(x, a, Z))) = h\left(\inf_{a \in D(x)} \text{VaR}_\alpha(T(x, a, Z))\right).$$

Hence, the minimizer of $a \mapsto \text{VaR}_\alpha(T(x, a, Z))$ induces an optimal decision rule for each stage. In particular the optimal policy is stationary and does not depend on time.

Note here that we can interpret a spectral risk measure as a Value-at-Risk criterion with unknown parameter α which has a prior distribution given by the density ϕ . However, since we apply it recursively at each stage, learning of the parameter is not possible.

Example 8.2 (Stopping Problems). Let us consider the following standard stopping problem: Suppose a real-valued Markov chain $(X_n) \subset L^p$ is given by $X_{n+1} = T(X_n, Z_{n+1})$ where (Z_n) is an i.i.d. sequence of random variables. We are allowed to observe the Markov chain and when we stop it in state x we have to pay the cost $c(x)$. In case we do not stop we have to pay the fixed cost \bar{c} . We have to stop no later than time point N . Suppose Assumptions 4.2 (i) and (ii) are fulfilled. The risk measure ρ is simply monetary and finite. The Fatou property is not needed here since the existence of minimizers is immediate. The Bellman equation is

$$\begin{aligned} J_N(x) &= x, \\ J_n(x) &= \min \left\{ \rho(c(x)); \rho(\bar{c} + \beta J_{n+1}(T(x, Z_{n+1}))) \right\}, \\ &= \min \left\{ \rho(c(x)); \bar{c} + \rho(\beta J_{n+1}(T(x, Z_{n+1}))) \right\}, \quad n = 0, \dots, N-1. \end{aligned}$$

In the well-known house selling application for example X_n is the offer for a house at time n that we may buy. When we decide to buy it we have to pay the price, i.e. $c(x) = x$. In case we do not buy, we still have to pay the rent \bar{c} . Offers are here assume to be i.i.d. Thus, the Bellman equation specializes to

$$\begin{aligned} J_N(x) &= x, \\ J_n(x) &= \min \{ \rho(x); \bar{c} + \rho(\beta J_{n+1}(Z_{n+1})) \}, \quad n = 0, \dots, N-1. \end{aligned}$$

Thus, when we define

$$t_n := \sup \{ x \in \mathbb{R} : \rho(x) \leq \bar{c} + \rho(\beta J_{n+1}(Z_{n+1})) \}$$

then the optimal policy obviously is to buy at time n if $x \leq t_n$, otherwise not. Hence the optimal strategy is still a threshold policy, but the thresholds depend on ρ . For example if ρ is normalized and $\rho(X) \geq \mathbb{E}X$ (this is e.g. satisfied for Average-Value-at-Risk or the Entropic risk measures) then $t_n \geq t_n^E$ where t_n^E belongs to the case $\rho = \mathbb{E}$. Hence, under the risk measure we will accept an offer earlier.

Example 8.3 (Casino Game). Suppose we have to play N -times the same game and decide how much of our current capital we should bet. Outcomes are either a gain or a loss and given by i.i.d. random variables (Z_n) , with $\mathbb{P}(Z_n = 1) = p = 1 - \mathbb{P}(Z_n = -1)$. Note that the Z_n are bounded. We assume that the risk measure is monetary, law-invariant, positive homogeneous and has the Fatou property. We want to minimize the risk of a loss. Note that Assumptions 3.1 and 4.2 are satisfied. The Bellman equation is

$$\begin{aligned} J_N(x) &= -x, \\ J_n(x) &= \inf_{0 \leq a \leq x} \rho(J_{n+1}(T(x, a, Z_{n+1}))), \\ &= \inf_{0 \leq a \leq x} \rho(J_{n+1}(x + aZ)), \quad n = 0, \dots, N-1. \end{aligned}$$

It is then easy to see by induction that the optimal policy is stationary and given by

$$d^*(x) = \begin{cases} 0, & \rho(-Z) \geq 0, \\ x, & \rho(-Z) < 0. \end{cases}$$

From the monotonicity and law-invariance of ρ it follows that there exists $p^* \in [0, 1]$ such that

$$d^*(x) = \begin{cases} 0, & p < p^*, \\ x, & p \geq p^*. \end{cases}$$

In case $\rho(-Z) < 0$, bold-play is best and we obtain by induction that $J_n(x) = -x(1 - \rho(-Z))^{N-n}$. When ρ is additionally convex (which then means coherent) and the game is fair ($p = \frac{1}{2}$), we obtain since $0 \leq_{cx} -Z_n$ in convex order implying that $0 = \rho(0) \leq \rho(-Z)$ (see Theorem 3.4 in [8]) and it is obviously best not to play.

Example 8.4 (Cash Balance). In a cash balance problem the aim is to keep the cash level of a company close to zero, because a negative cash level means we have to pay interest and a positive cash level creates opportunity cost (see Section 2.6.2 in [9]). We assume that a convex function $L : \mathbb{R} \rightarrow \mathbb{R}_+$ with $L(0) = 0$ gives the cost of deviating from zero. The cash level is subject to random changes which are modelled as i.i.d. random variables (Z_n) . It is possible to increase or decrease the cash level at the beginning of each period by paying transfer cost. The transfer cost $c : \mathbb{R} \rightarrow \mathbb{R}_+$ are assumed to be piecewise linear:

$$c(x') = c_u(x')^+ + c_d(x')^-$$

with $c_u, c_d > 0$. Of course the state is here the cash level and we choose the action to be the new cash level. Hence the transition function is $T(x, a, z) = a - z$. We consider the infinite horizon problem and suppose that Assumption 5.1 is in force. Assumption 3.1 is here satisfied, except for the compactness of the admissible actions which are here \mathbb{R} . However, it can be seen that it

is possible to restrict to a compact level set (see Section 2.6.2 in [9] for details). The Bellman equation for the infinite horizon problem is here

$$\begin{aligned} J_\infty(x) &= \sup_{a \in \mathbb{R}} \rho \left(c(a-x) + L(a) + \beta J_\infty(a-Z) \right) \\ &= \sup_{a \in \mathbb{R}} \left\{ c(a-x) + L(a) + \beta \rho \left(J_\infty(a-Z) \right) \right\} \end{aligned}$$

Now we can proceed exactly in the same way as in the classical case (see Section 2.6.2 in [9] for details) since the functions

$$\begin{aligned} h_u(a) &:= (a-x)c_u + L(a) + \beta \rho \left(J_\infty(a-Z) \right) \\ h_d(a) &:= (x-a)c_d + \beta \rho \left(J_\infty(a-Z) \right) \end{aligned}$$

are still both convex under our assumption that ρ is convex. We obtain:

Proposition 8.5. *For the cash balance problem with infinite horizon it holds under Assumption 5.1:*

a) *There exist critical levels S_- and S_+ such that*

$$J_\infty(x) = \begin{cases} (S_- - x)c_u + L(S_-) + \beta \rho \left(J_\infty(S_- - Z) \right) & \text{if } x < S_- \\ L(x) + \beta \rho \left(J_\infty(x - Z) \right) & \text{if } S_- \leq x \leq S_+ \\ (x - S_+)c_d + L(S_+) + \beta \rho \left(J_\infty(S_+ - Z) \right) & \text{if } x > S_+. \end{cases}$$

J_∞ is convex.

b) *The stationary policy (f^*, f^*, \dots) is optimal with*

$$f^*(x) := \begin{cases} S_- & \text{if } x < S_-, \\ x & \text{if } S_- \leq x \leq S_+, \\ S_+ & \text{if } x > S_+, \end{cases} \quad (8.1)$$

Of course the switching points S_- and S_+ depend on the choice of the risk measure ρ .

REFERENCES

1. Beatrice Acciaio and Irina Penner, *Dynamic risk measures*, Advanced mathematical methods for finance, Springer, 2011, pp. 1–34.
2. Charalambos D. Aliprantis and Kim C. Border, *Infinite dimensional analysis: A hitchhiker's guide*, 3rd ed., Springer-Verlag, Berlin Heidelberg, 2006.
3. Hubert Asienkiewicz and Anna Jaśkiewicz, *A note on a new class of recursive utilities in Markov decision processes*, *Applicationes Mathematicae* **44** (2017), no. 2, 149–161.
4. Nicole Bäuerle and Alexander Glauner, *Optimal risk allocation in reinsurance networks*, *Insurance: Mathematics and Economics* **82** (2018), 37–47.
5. ———, *Distributionally robust Markov decision processes and their connection to risk measures*, arXiv:2007.13103, 2020.
6. Nicole Bäuerle and Anna Jaśkiewicz, *Optimal dividend payout model with risk sensitive preferences*, *Insurance: Mathematics and Economics* **73** (2017), 82–93.
7. ———, *Stochastic optimal growth model with risk sensitive preferences*, *Journal of Economic Theory* **173** (2018), 181–200.
8. Nicole Bäuerle and Alfred Müller, *Stochastic orders and risk measures: Consistency and bounds*, *Insurance: Mathematics and Economics* **38** (2006), no. 1, 132–148.
9. Nicole Bäuerle and Ulrich Rieder, *Markov decision processes with applications to finance*, Springer-Verlag, Berlin Heidelberg, 2011.
10. Tomasz R Bielecki, Igor Cialenco, and Marcin Pitera, *A unified approach to time consistency of dynamic risk measures and dynamic performance measures in discrete time*, *Mathematics of Operations Research* **43** (2018), no. 1, 204–221.
11. Jocelyne Bion-Nadal, *Time consistent dynamic risk processes*, *Stochastic Processes and their Applications* **119** (2009), no. 2, 633–654.
12. Shanyun Chu and Yi Zhang, *Markov decision processes with iterated coherent risk measures*, *International Journal of Control* **87** (2014), no. 11, 2286–2293.
13. Kai Detlefsen and Giacomo Scandolo, *Conditional and dynamic convex risk measures*, *Finance and Stochastics* **9** (2005), no. 4, 539–561.

14. Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang, *Remarks on quantiles and distortion risk measures*, *European Actuarial Journal* **2** (2012), no. 2, 319–328.
15. Larry G. Epstein and Martin Schneider, *Recursive multiple-priors*, *Journal of Economic Theory* **113** (2003), no. 1, 1–31.
16. Alexander Glauner, *Robust and risk-sensitive Markov decision processes with applications to dynamic optimal reinsurance*, Ph.D. thesis, Karlsruhe Institute of Technology, 2020.
17. Onésimo Hernández-Lerma and Jean B Lasserre, *Discrete-time Markov control processes: basic optimality criteria*, vol. 30, Springer Science & Business Media, 2012.
18. Onésimo Hernández-Lerma and Jean Bernard Lasserre, *Further topics on discrete-time Markov control processes*, Springer-Verlag, New York, 1999.
19. Tito Homem-de Mello and Bernardo K Pagnoncelli, *Risk aversion in multistage stochastic programming: A modeling and algorithmic perspective*, *European Journal of Operational Research* **249** (2016), no. 1, 188–199.
20. Daniel R Jiang and Warren B Powell, *Practicality of nested risk measures for dynamic electric vehicle charging*, arXiv preprint arXiv:1605.02848 (2016).
21. David M. Kreps and Evan L. Porteus, *Temporal resolution of uncertainty and dynamic choice theory*, *Econometrica* **46** (1978), no. 1, 185–200.
22. Jianjun Miao, *Economic dynamics in discrete time*, MIT Press, Cambridge, Mass., 2014.
23. Alois Pichler, *The natural Banach space for version independent risk measures*, *Insurance: Mathematics and Economics* **53** (2013), no. 2, 405–415.
24. Martin L Puterman, *Markov decision processes: discrete stochastic dynamic programming*, John Wiley & Sons, 2014.
25. Frank Riedel, *Dynamic coherent risk measures*, *Stochastic processes and their applications* **112** (2004), no. 2, 185–200.
26. Ludger Rüschendorf, *Mathematical risk analysis: Dependence, risk bounds, optimal allocations and portfolios*, Springer-Verlag, Berlin Heidelberg, 2013.
27. Andrzej Ruszczyński, *Risk-averse dynamic programming for Markov decision processes*, *Mathematical Programming* **125** (2010), no. 2, 235–261.
28. Rouven Schur, Jochen Gönsch, and Michael Hassler, *Time-consistent, risk-averse dynamic pricing*, *European Journal of Operational Research* **277** (2019), no. 2, 587–603.
29. Ekaterina N. Sereda, Efim M. Bronshtein, Svetozar T. Rachev, Frank J. Fabozzi, Wei Sun, and Stoyan V. Stoyanov, *Distortion risk measures in portfolio optimization*, *Handbook of Portfolio Construction* (John B. Guerard, ed.), Springer, New York, 2010, pp. 649–673.
30. Alexander Shapiro, *Minimax and risk averse multistage stochastic programming*, *European Journal of Operational Research* **219** (2012), no. 3, 719–726.
31. ———, *Time consistency of dynamic risk measures*, *Operations Research Letters* **40** (2012), no. 6, 436–439.
32. Yun Shen, Wilhelm Stannat, and Klaus Obermayer, *Risk-sensitive Markov control processes*, *SIAM Journal on Control and Optimization* **51** (2013), no. 5, 3652–3672.
33. Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor, *Sequential decision making with coherent risk*, *IEEE Transactions on Automatic Control* **62** (2016), no. 7, 3323–3338.
34. Stefan Weber, *Distribution-invariant risk measures, information, and dynamic consistency*, *Mathematical Finance: An International Journal of Mathematics, Statistics and Financial Economics* **16** (2006), no. 2, 419–441.

(N. Bäuerle) DEPARTMENT OF MATHEMATICS, KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT), D-76128 KARLSRUHE, GERMANY

Email address: nicole.baeuerle@kit.edu

(A. Glauner) DEPARTMENT OF MATHEMATICS, KARLSRUHE INSTITUTE OF TECHNOLOGY (KIT), D-76128 KARLSRUHE, GERMANY

Email address: alexander.glauner@kit.edu