Math. Program., Ser. B (2010) 125:235–261
DOI 10.1007/s10107-010-0393-3

FULL LENGTH PAPER

Risk-averse dynamic programming for Markov decision processes

Andrzej Ruszczyński

Received: 29 April 2009 / Accepted: 9 November 2009 / Published online: 21 July 2010

Abstract We introduce the concept of a Markov risk measure and we use it to formulate risk-averse control problems for two Markov decision models: a finite horizon model and a discounted infinite horizon model. For both models we derive risk-averse dynamic programming equations and a value iteration method. For the infinite horizon problem we develop a risk-averse policy iteration method and we prove its convergence. We also propose a version of the Newton method to solve a nonsmooth equation arising in the policy iteration method and we prove its global convergence. Finally, we discuss relations to min–max Markov decision models.

Keywords Dynamic risk measures · Markov risk measures · Value iteration · Policy iteration · Nonsmooth Newton's method · Min-max Markov models

Mathematics Subject Classification (2000) Primary 49L20 · 90C40 · 91B30; Secondary 91A25 · 93E20

1 Introduction

Dynamic programming is one of classical areas of operations research. Initiated by Bellman [4], it underwent rapid development in the last five decades, both in theory and applications. Several excellent textbooks and monographs [5,6,21,27,32,39,50] discuss various aspects of this vast area.

Classical dynamic programming models are concerned with expected performance criteria. However, in many practical problems the expected values may not be appropri-

A. Ruszczyński (✉)
Department of Management Science and Information Systems,
Rutgers University, Piscataway, NJ 08854, USA
e-mail: rusz@business.rutgers.edu

ate to measure performance. Models with risk aversion were, so far, represented by separable utility functions; see, among others, [9,12,13,23,24] and the references therein.

The need to put mathematical framework into the theory risk-averse preferences was one of motivations for the development of the general theory of risk measures. Starting from the seminal publication of Artzner et al. [1], the theory developed new tools to evaluate risk of uncertain outcomes and stochastic processes. Particularly relevant for us are the duality theory [1,10,15–17,34,43,45,46], and the theory of conditional and dynamic risk measures [2,8,11,14,18,19,28,38,48,40,45,47].

Our plan is to adapt concepts and methods of the modern theory of risk measures to dynamic programming models for *Markov* decision processes. The adaptation is not straightforward, and new ideas and techniques need to be developed.

This work is not a survey paper, but rather an original contribution. In Sect. 2 we quickly review fundamental concepts of controlled Markov models. Section 3 has a synthetic character. We present the concepts of conditional and dynamic risk measures, and their time-consistency, and we derive a nested structure of a time-consistent dynamic risk measure. The main contributions of the paper are contained in Sects. 4–9. In Sect. 4 we introduce the concepts of a risk transition mapping and of a Markov risk measure and we analyze their properties. In Sect. 5 we derive dynamic programming equations for finite horizon problems with Markov risk measures. Section 6 is devoted to the construction of a discounted measure of risk for infinite cost sequences. It is used in an infinite horizon Markov problem in Sect. 7, where the corresponding dynamic programming equations are developed. We also present there a risk-averse version of the value iteration method. In Sect. 8 we propose and analyze a risk-averse policy iteration method. In Sect. 9 we present a specialized nonsmooth Newton method for solving a nonsmooth equation arising in the evaluation step of the policy iteration method, and we prove its global monotonic convergence. Finally, in Sect. 10 we discuss relations of our results to min-max Markov decision problems.

2 Controlled Markov models

We quickly review the main concepts of controlled Markov models and we introduce relevant notation. Our presentation is close to that of [21]. Let $\mathscr{X}$ and $\mathscr{U}$ be two Borel spaces (Polish spaces equipped with their Borel σ-algebras $\mathscr{B}(\mathscr{X})$ and $\mathscr{B}(\mathscr{Y})$, respectively), and let $U : \mathscr{X} \rightrightarrows \mathscr{U}$ be a measurable multifunction. We call $\mathscr{X}$ the *state space*, $\mathscr{U}$ the *control space*, and $U(\cdot)$ the *control set*. We also introduce the graph of the multifunction U,

$$\operatorname{graph}(U) = \{(x, u) \in \mathscr{X} \times \mathscr{U} : u \in U(x)\}.$$

We use $\mathscr{P}$ to denote the set of probability measures on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ and we endow it with the weak topology. A *stochastic kernel* is a measurable function $K : \mathscr{X} \to \mathscr{P}$. For a Borel set $B \in \mathscr{B}(\mathscr{X})$ we write the measure $[K(x)](B)$ as $K(B \mid x)$. By a *controlled kernel* we mean a measurable function $Q : \operatorname{graph}(U) \to \mathscr{P}$. This means that for all $x \in \mathscr{X}$ and all $u \in U(x)$ its value $Q(x, u)$ is a probability measure on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$. Its values are written $Q(B \mid x, u)$, where B is a Borel subset of $\mathscr{X}$, $x \in \mathscr{X}$, $u \in U(x)$. Finally a *cost function* is a measurable mapping $c : \operatorname{graph}(U) \to \mathbb{R}$.

A *controlled Markov model* is defined by a state space $\mathscr{X}$, a control space $\mathscr{U}$, and sequences of control sets U_t, controlled kernels Q_t, and cost functions c_t, $t = 1, 2, \ldots$.

For $t = 1, 2, \ldots$ we define the space $\mathscr{H}_t$ of admissible state histories up to time t as $\mathscr{H}_t = \mathscr{X}^t$. A *policy* is a sequence of measurable functions $\pi_t : \mathscr{H}_t \to \mathscr{U}$, $t = 1, 2, \ldots$, such that $\pi_t(x_1, \ldots, x_t) \in U_t(x_t)$ for all $(x_1, \ldots, x_t) \in \mathscr{H}_t$. A *Markov policy* is a sequence of measurable functions $\pi_t : \mathscr{X} \to \mathscr{U}$, $t = 1, 2, \ldots$, such that $\pi_t(x) \in U_t(x)$ for all $x \in \mathscr{X}$, *i. e.*, the mappings π_t are measurable selections of U_t. A Markov policy is *stationary* if $\pi_t = \pi_1$, $t = 2, 3, \ldots$.

To simplify the presentation of our ideas, we restrict our considerations to deterministic policies. In a more general setting one can define $\overline{\mathscr{H}}_t = \text{graph}(U_1) \times \cdots \times \text{graph}(U_t) \times \mathscr{X}$ and consider *mixed policies* defined as mappings from $\overline{\mathscr{H}}_t$ to the set of probability measures on $(\mathscr{U}, \mathscr{B}(\mathscr{U}))$, and such that

$$[\pi_t(x_1, u_1, \ldots, x_{t-1}, u_{t-1}, x_t)](U_t(x_t)) = 1.$$

All our considerations can be cast in this framework, with more complex notation, but with virtually no conceptual difference.

Consider the canonical sample space $\Omega = \mathscr{X}^\infty$ with the product σ-algebra $\mathscr{F}$. Let P_1 be the initial distribution of the state $x_1 \in \mathscr{X}$. Suppose we are given a policy $\Pi = \{\pi_t\}_{t=1}^\infty$, $t = 1, 2, \ldots$. The Ionescu Tulcea theorem (see, e.g., [6]) states that there exists a unique probability measure P^Π on $(\Omega, \mathscr{F})$ such that for every measurable set $B \subset \mathscr{X}$ and all $h_t \in \mathscr{H}_t$, $t = 1, 2, \ldots$,

$$P^\Pi(x_1 \in B) = P_1(B);$$
$$P^\Pi(x_{t+1} \in B \mid h_t) = Q_t(B \mid x_t, \pi_t(h_t)).$$

To simplify our notation, from now on we focus on the case when the initial state x_1 is fixed. It will be obvious how to modify our results for a random initial state.

We start from the following two basic optimization problems for controlled Markov processes. The first one is the *finite horizon expected value problem*, in which, for a given $T \geq 1$, we want to find a policy $\Pi = \{\pi_1, \ldots, \pi_T\}$ so as to minimize the expected cost:

$$\min_\Pi \mathbb{E}\left[\sum_{t=1}^{T} c_t(x_t, u_t) + c_{T+1}(x_{T+1})\right], \tag{1}$$

where $u_t = \pi_t(x_1, \ldots, x_t)$ and $c_{T+1} : \mathscr{X} \to \mathbb{R}$ is a measurable function.

The second problem is the *infinite horizon discounted expected value problem*. For a given $\alpha \in (0, 1)$, our aim is to find a policy $\Pi = \{\pi_t\}_{t=1}^\infty$ so as to minimize the expected discounted cost:

$$\min_\Pi \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c_t(x_t, u_t)\right]. \tag{2}$$

Under more specific but still fairly general assumptions, both problems have solutions in form of Markov policies (see, e.g., [21]). Moreover, the second problem has a stationary optimal policy, if the underlying Markov model is stationary. In both cases,

the optimal policies can be constructed by solving appropriate dynamic programming equations.

Our intention is to introduce risk aversion to both problems, and to replace the expected value operators in (1) and (2) by more general risk measures.

In order to motivate our research, let us consider a simple example.

Example 1 Suppose a device can be in one of the states $i = 1, 2, \ldots, n$, with 1 representing "new," n representing "irreparable," and intermediate states corresponding to decreasing quality levels. At each state $i = 1, 2, \ldots, n$ we have a set $U(i)$ of possible control decisions u. For $i = 2, \ldots, n-1$,

$$U(i) = \{\text{"do nothing"}, \text{"perform maintenance"}, \text{"replace"}\};$$

$U(1)$ contains only the first two controls, and $U(n) = \{\text{"replace"}\}$. Corresponding transition probabilities $q_{ij}(u), i = 1, \ldots, n, j = 1, \ldots, n, u \in U(i)$, describe the evolution of the controlled Markov chain. Costs $c_i(u), i = 1, \ldots, n, u \in U(i)$, are associated with each state-control pair. Our objective is to minimize expected discounted cost of operating the system over infinite horizon. For nonnegative costs and any transition kernel $q_{ij}(u)$, such problem has an optimal solution given by some Markov policy $\hat{\pi}$.

Suppose a possibility is offered at each state $i = 1, \ldots, n-1$ to purchase insurance against transition to the "irreparable" state in the next stage. Such an insurance would decrease the replacement cost by some amount $C > 0$. The price $w(i)$ of the insurance depends on the current state i. With this option available, all combinations of previously available controls with "yes" or "no" decisions to purchase insurance are possible control values. Consequently, the set of possible controls increases to 6 at states $2, \ldots, n-1$, and to 4 at state 1. We also have to augment our model with a new state $n+1$: "irreparable insured." Suppose insurance does not change any transition probabilities, except that transition to state n is replaced by an equally likely transition to state $n+1$. As the insurance seller is a profit-making business, it is reasonable to assume that

$$w(i) > Cp_{in}(\hat{u}(i)), \quad i = 1, \ldots, n-1.$$

Then in the expected value model the insurance will never be purchased, because the expected profit of the insurer is our expected loss. Similar situations arise in financial models, where options can be purchased. A question arises, why warranty or insurance are purchased in real life? In order to understand the motivation for such actions, we need to introduce risk aversion to the preference model of the decision maker.

3 Dynamic risk measures: time consistency

Consider a probability space $(\Omega, \mathscr{F}, P)$, a filtration $\mathscr{F}_1 \subset \cdots \subset \mathscr{F}_T \subset \mathscr{F}$, and an adapted sequence of random variables $Z_t, t = 1, \ldots, T$. We assume that $\mathscr{F}_1 = \{\Omega, \emptyset\}$, and thus Z_1 is in fact deterministic. In our all considerations we interpret the variables Z_t as stage-wise costs.

Define the spaces $\mathscr{Z}_t = \mathscr{L}_p(\Omega, \mathscr{F}_t, P)$, $p \in [1, \infty]$, $t = 1, \ldots, T$, and $\mathscr{Z}_{t,T} = \mathscr{Z}_t \times \cdots \times \mathscr{Z}_T$.

The fundamental question in the theory of dynamic risk measures is the following: *how do we evaluate the risk of the subsequence $Z_t, \ldots, Z_T$ from the perspective of stage t?* This motivates the following definition.

Definition 1 A mapping $\rho_{t,T} : \mathscr{Z}_{t,T} \to \mathscr{Z}_t$, where $1 \le t \le T$, is called a *conditional risk measure*, if it has the following monotonicity property:

$$\rho_{t,T}(Z) \le \rho_{t,T}(W) \text{ for all } Z, W \in \mathscr{Z}_{t,T} \text{ such that } Z \le W. \tag{3}$$

Here and elsewhere in the paper, inequalities between random vectors are understood component-wise and in the almost sure sense.

The monotonicity requirement in Definition 1 is weaker than the monotonicity condition of [8,25], because we consider the sequence of stage-wise costs, rather than the sequence of positions, or cumulative costs, $C_t = \sum_{\tau=1}^{t} Z_t$, $t = 1, \ldots, T$. This is discussed in detail in [19, Remark 4.1]. Our main motivation to look at stage-wise costs, rather than positions, is the application to discrete-time dynamic optimization.

The value of the conditional risk measure $\rho_{t,T}(Z_t, \ldots, Z_T)$ can be interpreted as a fair one-time $\mathscr{F}_t$-measurable charge we would be willing to incur at time t, instead of the sequence of random future costs $Z_t, \ldots, Z_T$.

Much work on dynamic measures of risk focused on the case when we have just one final cost Z_T and we are evaluating it from the perspective of earlier stages t; see, *inter alia* [14,18,28,48]. Another approach is to define time-consistency directly through the dynamic programming principle in optimization models, as it is done for a portfolio problem in [7]. Our view is close to [19], who consider risk of a sequence of payoffs.

Definition 2 A *dynamic risk measure* is a sequence of conditional risk measures $\rho_{t,T} : \mathscr{Z}_{t,T} \to \mathscr{Z}_t$, $t = 1, \ldots, T$.

The key issue associated with dynamic preferences is the question of their consistency over time. It has been studied in various contexts in the past (see, *inter alia*, [2,29,30]); here, we adapt the perspective which is closest to that of [8]. The following definition is similar to [8, Prop. 4.4], which we use as the starting point, due to its intuitive appeal.

Definition 3 A dynamic risk measure $\{\rho_{t,T}\}_{t=1}^{T}$ is called *time-consistent* if for all $1 \le \tau < \theta \le T$ and all sequences $Z, W \in \mathscr{Z}_{\tau,T}$ the conditions

$$Z_k = W_k, \ k = \tau, \ldots, \theta - 1 \quad \text{and} \quad \rho_{\theta,T}(Z_\theta, \ldots, Z_T) \le \rho_{\theta,T}(W_\theta, \ldots, W_T) \tag{4}$$

imply that

$$\rho_{\tau,T}(Z_\tau, \ldots, Z_T) \le \rho_{\tau,T}(W_\tau, \ldots, W_T). \tag{5}$$

In words, if Z will be at least as good as W from the perspective of some future time θ, and they are identical between now (τ) and θ, then Z should not be worse than W

from today's perspective. A similar view is adapted in [40], but with equality, rather than inequality $\leq$, between risk measures in (4)–(5).

For a dynamic risk measure $\{\rho_{t,T}\}_{t=1}^{T}$ we can define a broader family of conditional risk measures, by setting

$$\rho_{\tau,\theta}(Z_\tau,\ldots,Z_\theta)=\rho_{\tau,T}(Z_\tau,\ldots,Z_\theta,0,\ldots,0),\quad 1\leq\tau\leq\theta\leq T. \tag{6}$$

We can derive the following structure of a time-consistent dynamic risk measure.

Theorem 1 *Suppose a dynamic risk measure $\{\rho_{t,T}\}_{t=1}^{T}$ satisfies for all $Z\in\mathscr{Z}$ and all $t=1,\ldots,T$ the conditions:*

$$\rho_{t,T}(Z_t,Z_{t+1},\ldots,Z_T)=Z_t+\rho_{t,T}(0,Z_{t+1},\ldots,Z_T), \tag{7}$$

$$\rho_{t,T}(0,\ldots,0)=0. \tag{8}$$

Then it is time-consistent if and only if for all $1\leq\tau\leq\theta\leq T$ and all $Z\in\mathscr{Z}_{1,T}$ the following identity is true:

$$\rho_{\tau,T}\left(Z_\tau,\ldots,Z_\theta,\ldots,Z_T\right)=\rho_{\tau,\theta}\left(Z_\tau,\ldots,Z_{\theta-1},\rho_{\theta,T}(Z_\theta,\ldots,Z_T)\right). \tag{9}$$

Proof We adapt the argument of [8, Prop. 4.4]. Consider two sequences:

$$\begin{aligned}Z&=(Z_\tau,\ldots,Z_{\theta-1},Z_\theta,Z_{\theta+1},\ldots,Z_T),\\ W&=\left(Z_\tau,\ldots,Z_{\theta-1},\rho_{\theta,T}(Z_\theta,\ldots,Z_T),0,\ldots,0\right).\end{aligned}$$

Suppose the measure $\{\rho_{t,T}\}_{t=1}^{T}$ is time-consistent. Then, by (7) and (8),

$$\begin{aligned}\rho_{\theta,T}(W_\theta,\ldots,W_T)&=\rho_{\theta,T}\left(\rho_{\theta,T}(Z_\theta,\ldots,Z_T),0,\ldots,0\right)\\ &=\rho_{\theta,T}(Z_\theta,\ldots,Z_T)+\rho_{\theta,T}(0,\ldots,0)=\rho_{\theta,T}(Z_\theta,\ldots,Z_T).\end{aligned}$$

Using Definition 3 we get $\rho_{\tau,T}(Z)=\rho_{\tau,T}(W)$. Equation (6) then yields (9).

Conversely, suppose the identity (9). Consider Z and W satisfying conditions (4). Then, by Definition 1, we have

$$\begin{aligned}&\rho_{\tau,T}\left(Z_\tau,\ldots,Z_{\theta-1},\rho_{\theta,T}(Z_\theta,\ldots,Z_T),0,\ldots,0\right)\\ &\quad\leq\rho_{\tau,T}\left(W_\tau,\ldots,W_{\theta-1},\rho_{\theta,T}(W_\theta,\ldots,W_T),0,\ldots,0\right).\end{aligned}$$

Using Eq. (9) we obtain (5). □

We may remark that condition (7) is a form of the *translation property*, discussed in various settings in [2,19,38]. Our version is weaker, because Z_t is $\mathscr{F}_t$-measurable.

In fact, relations corresponding to (9) are usually used to define time-consistency of dynamic risk measures: see, e.g., [8,25] and the references therein.

If the risk measure is time-consistent and satisfies (7) and (8), then we obtain the chain of equations:

$$\begin{aligned}\rho_{t,T}\left(Z_t,\ldots,Z_{T-1},Z_T\right) &= \rho_{t,T}\left(Z_t,\ldots,\rho_{T-1,T}(Z_{T-1},Z_T)\right)\\ &= \rho_{t,T-1}\left(Z_t,\ldots,Z_{T-1}+\rho_{T-1,T}(0,Z_T)\right).\end{aligned}$$

In the first equation we used the identity (9), and in the second one condition (7). Define one-step conditional risk measures $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t, t = 1,\ldots,T-1$ as follows:

$$\rho_t(Z_{t+1}) = \rho_{t,t+1}(0, Z_{t+1}).$$

Proceeding in this way, we obtain for all $t = 1,\ldots,T$ the following recursive relation:

$$\begin{aligned}\rho_{t,T}(Z_t,\ldots,Z_T) = Z_t + \rho_t\big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-2}\\ \times\ (Z_{T-1} + \rho_{T-1}(Z_T))\cdots\big)\big).\end{aligned} \tag{10}$$

It follows that a time-consistent dynamic risk measure is completely defined by one-step conditional risk measures $\rho_t, t = 1,\ldots,T-1$. For $t = 1$ formula (10) defines a risk measure of the *entire* sequence $Z \in \mathscr{Z}_{1,T}$ (with a deterministic Z_1).

It may be worth mentioning that the particular structure (10) of a dynamic risk measure has been introduced in a constructive way in [47]. Here, it has been derived from general principles of time-consistency and conditions (7)–(8).

Another important property of a dynamic risk measure is the local property, discussed in detail in [8,25,28].

Definition 4 A conditional risk measure $\rho_{\tau,\theta}$, $1 \le \tau \le \theta \le T$, has the *local property* if for all sequences $Z \in \mathscr{Z}_{\tau,\theta}$ and all events $A \in \mathscr{F}_\tau$ we have

$$\rho_{\tau,\theta}(\mathbb{1}_A Z) = \mathbb{1}_A \rho_{\tau,\theta}(Z).$$

It means that if event $A \in \mathscr{F}_\tau$ did not happen, then the risk of the future costs $\mathbb{1}_A Z$ is zero. Local property does not follow from time-consistency, as defined here. In [49] the local property is called time-consistency, but we shall follow here the terminology established in the dynamic risk measurement literature.

Our considerations had, so far, quite general character. Their main objective was to derive Eq. (10), which will have fundamental importance in all our considerations.

To proceed further, we shall assume stronger properties of the one-step conditional risk measures $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t, t = 1,\ldots,T-1$, appearing in (10):

A1. $\rho_t(\lambda Z + (1-\lambda)W) \le \lambda\rho_t(Z) + (1-\lambda)\rho_t(W)\ \ \forall\,\lambda \in (0,1),\ \ Z, W \in \mathscr{Z}_{t+1}$;
A2. If $Z \le W$ then $\rho_t(Z) \le \rho_t(W),\ \forall\, Z, W \in \mathscr{Z}_{t+1}$;
A3. $\rho_t(Z+W) = Z + \rho_t(W),\ \forall\, Z \in \mathscr{Z}_t,\ \ W \in \mathscr{Z}_{t+1}$;
A4. $\rho_t(\beta Z) = \beta\rho_t(Z), \forall\, Z \in \mathscr{Z}_{t+1},\ \ \beta \ge 0$.

These axioms were introduced in the spaces $\mathscr{L}_\infty(\Omega, \mathscr{F}_t, P)$ in [48]; later they were analyzed in the general setting in [28,47]. In the special case of $\mathscr{F}_1 = \{\Omega, \emptyset\}$ these are the axioms of a *coherent measure of risk* of [1] (also introduced for $p = \infty$).

In fact, (A2) follows from Definition 1, but we repeat it here for completeness. We can also remark that the property of time consistency, conditions (7)–(8), and assumption (A3) imply a much stronger monotonicity property, than that assumed in Definition 1. Applying (A3) for $T, T-1, \dots, t$ to (10) we get

$$\rho_{t,T}(Z_t, \dots, Z_T) = \rho_t\left(\rho_{t+1}\left(\cdots \rho_{T-2}\left(\rho_{T-1}(Z_t + Z_{t+1} + \cdots + Z_{T-1} + Z_T)\right)\cdots\right)\right).$$

As all $\rho_\tau, \tau = t, \dots, T-1$, are monotone, their composition is monotone as well. Therefore

$$\rho_{t,T}(Z_t, \dots, Z_T) \le \rho_{t,T}(W_t, \dots, W_T) \quad \text{if} \quad \sum_{\tau=t}^{T} Z_\tau \le \sum_{\tau=t}^{T} W_\tau.$$

Assumptions (A2) and (A3) imply the local property of a one-step conditional risk measure; see [47, Prop. 3.2] and [28, Sect. 3].

All our further considerations will assume the form (10) of a dynamic risk measure, with one-step conditional risk measures ρ_t satisfying (A1)–(A4). The system (A1)–(A4) is written for the case when lower values of Z are preferred (for example, Z represents an uncertain cost); a similar system, with (A2) and (A3) suitably modified, can be written for the opposite preferences, and all results for any of these systems can be easily translated to the other one.

Example 2 An example of a one-step conditional risk measure is the following *mean–semideviation model* analyzed in [35,36], [46, Example 4.2], [47, Example 6.1]):

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1}|\mathscr{F}_t] + \kappa \mathbb{E}\left[((Z_{t+1} - \mathbb{E}[Z_{t+1}|\mathscr{F}_t])_+)^r \mid \mathscr{F}_t\right]^{\frac{1}{r}}. \tag{11}$$

Here $r \in [1, p]$ and $\kappa \in [0, 1]$ may be $\mathscr{F}_t$-measurable random variables. The symbol $(z)_+$ denotes $\max(0, z)$.

Example 3 Another important example is the Conditional Average Value at Risk (see, *inter alia*, [37, Sect. 4], [38, Sect. 2.2.3, 3.3.4], [41], [46, Example 4.3], [47, Example 6.2]), which is defined as follows:

$$\rho_t(Z_{t+1}) = \inf_{U \in \mathscr{Z}_t} \left\{ U + \frac{1}{\alpha}\mathbb{E}\left[(Z_{t+1} - U)_+ \mid \mathscr{F}_t\right] \right\}. \tag{12}$$

In the formula above, the infimum is understood point-wise, and the level α may be an $\mathscr{F}_t$-measurable function with values in an interval $[\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$.

4 Markov risk measures

Consider now application of the dynamic risk measure (10) to a controlled Markov process $x_t, t = 1, \ldots, T, T+1$. Each policy Π results in a cost sequence $Z_t = c_t(x_t, u_t), t = 1, \ldots, T$, and $Z_{T+1} = c_{T+1}(x_{T+1})$. To evaluate risk of this sequence we use a dynamic time-consistent risk measure. By our results of Sect. 3, it has the form (10), which we recall here for convenience:

$$J(\Pi) = \rho_{1,T+1}(Z_1, \ldots, Z_{T+1}) = c_1(x_1, u_1) + \rho_1 \left(c_2(x_2, u_2) + \rho_2 \left(c_3(x_3, u_3)\right.\right.$$
$$\left.\left. + \cdots + \rho_{T-1} \left(c_T(x_T, u_T) + \rho_T(c_{T+1}(x_{T+1}))\right) \cdots\right)\right), \tag{13}$$

with some one-step conditional risk measures $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t, t = 1, \ldots, T$. The fundamental difficulty of this formulation is that at time t the value of $\rho_t(\cdot)$ is $\mathscr{F}_t$-measurable and is allowed to depend on the entire history of the process. For example, the multiplier κ in (11) may depend on $h_t = \{x_1, \ldots, x_t\}$. We cannot expect to obtain a Markov optimal policy, if our attitude to risk may depend on the whole past of the process.

In order to overcome these difficulties, we consider a new construction of a one-step conditional measure of risk. Its arguments are measurable functions on the state space $\mathscr{X}$, rather than on the probability space Ω. This entails additional complication, because in a controlled Markov process the probability measure on the state space is not fixed, but depends on our decisions u.

To simplify notation, from now on we write $\mathscr{B}$ for the σ-field $\mathscr{B}(\mathscr{X})$ of Borel sets in the state space $\mathscr{X}$. Let P_0 be some fixed probability measure on $(\mathscr{X}, \mathscr{B})$. Let $\mathscr{V} = \mathscr{L}_p(\mathscr{X}, \mathscr{B}, P_0), \mathscr{Y} = \mathscr{L}_q(\mathscr{X}, \mathscr{B}, P_0)$ with $p, q \in [1, \infty]$, and $1/p + 1/q = 1$. Define

$$\mathscr{M} = \left\{ m \in \mathscr{Y} : \int_{\mathscr{X}} m(x)\, P_0(dx) = 1,\ m \geq 0 \right\}.$$

We identify an element $m \in \mathscr{M}$ with a probability measure on $(\mathscr{X}, \mathscr{B})$, which has $m(x)$ as its density (Radon–Nikodym derivative) with respect to P_0. We also assume that the spaces $\mathscr{V}$ and $\mathscr{Y}$ are endowed with topologies that make them paired topological vector spaces with the bilinear form

$$\langle v, m \rangle = \int_{\mathscr{X}} v(x) m(x)\ P_0(dx). \tag{14}$$

In the sequel, we always assume that we the space $\mathscr{Y}$ (and thus $\mathscr{M}$) is endowed with the weak* topology. For $p \in [1, \infty)$ we may endow $\mathscr{V}$ with the strong (norm) topology, or with the weak topology. For $p = \infty$, the space $\mathscr{V}$ will be endowed with is weak topology defined by the form (14), that is, the weak* topology on $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$.

Definition 5 A measurable functional $\sigma : \mathscr{V} \times \mathscr{X} \times \mathscr{M} \to \mathbb{R}$ is a *risk transition mapping* associated with the controlled kernel $Q : \text{graph}(U) \to \mathscr{M}$ if

(i) For every $x \in \mathscr{X}$ and every $u \in U(x)$ the function $v \mapsto \sigma(v, x, Q(x, u))$ is a coherent measure of risk on $\mathscr{V}$;
(ii) For every $v \in \mathscr{V}$ and every measurable selection $u(\cdot)$ of $U(\cdot)$ the function $x \mapsto \sigma(v, x, Q(x, u(x))$ is an element of $\mathscr{V}$.

Suppose for every $x \in \mathscr{X}$ and every $m \in \mathscr{M}$ the risk transition mapping σ is lower semicontinuous with respect to the first argument. Then it follows from [46, Theorem 2.2] that there exists a closed convex set $\mathscr{A}(x, m) \subset \mathscr{M}$ such that for all $v \in \mathscr{V}$ we have

$$\sigma(v, x, m) = \sup_{\mu \in \mathscr{A}(x,m)} \langle v, \mu \rangle. \tag{15}$$

If, in addition, the functional $\sigma(\cdot, x, m)$ is continuous, then the set $\mathscr{A}(x, m)$ is bounded. In fact, for $p \in [1, \infty)$ the continuity of $\sigma(\cdot, x, m)$ follows from the monotonicity and convexity axioms [46, Proposition 3.1]. Moreover, the set $\mathscr{A}(x, m)$ is weakly* compact. In this case the "sup" operation in (15) can be replaced by the "max" operation.

Example 4 Consider the mean–semideviation measure defined by (11), but now with the state and the underlying probability measure as its arguments. For $r \geq 1$ we define

$$\sigma(v, x, m) = \langle v, m \rangle + \kappa(x) \left(\left\langle ((v - \langle v, m \rangle)_+)^r, m \right\rangle \right)^{\frac{1}{r}}, \tag{16}$$

with some measurable function $\kappa : \mathscr{X} \to [0, 1]$. Observe that a function $v \in \mathscr{V}$ is integrable with respect to a measure $m \in \mathscr{M}$. Therefore the first order ($r = 1$) semideviation with respect to the measure m is well-defined for all $p \in [1, \infty]$. Another important case is that of $p = \infty$. Then we can consider semideviations of any order $r \in [1, \infty)$. In this case r in (16) may be replaced by a measurable function from $\mathscr{X}$ to $[1, \infty)$.

Following the derivations of [46, Example 4.2], for $r > 1$ we obtain

$$\mathscr{A}(x, m) = \left\{ g \in \mathscr{M} : g = m\,(1 + h - \langle h, m \rangle),\ \left(\left\langle |h|^{\frac{r}{r-1}}, m \right\rangle \right)^{\frac{r-1}{r}} \leq \kappa(x),\ h \geq 0 \right\},$$

and for $r = 1$ we have

$$\mathscr{A}(x, m) = \left\{ g \in \mathscr{M} : g = m\,(1 + h - \langle h, m \rangle),\ \sup_{y \in \mathscr{X}} |h(y)| \leq \kappa(x),\ h \geq 0 \right\}.$$

Example 5 The Conditional Average Value at Risk defined in (12) has the following risk transition counterpart:

$$\sigma(v, x, m) = \inf_{u \in \mathbb{R}} \left\{ u + \frac{1}{\alpha(x)} \langle (v - u)_+, m \rangle \right\}.$$

Here $\alpha : \mathscr{X} \to [\alpha_{\min}, \alpha_{\max}]$ is measurable. Following the derivations of [46, Example 4.3], we obtain

$$\mathscr{A}(x, m) = \left\{ g \in \mathscr{M} : 0 \leq g \leq \frac{1}{\alpha(x)},\ \langle g, m \rangle = 1 \right\}.$$

Risk transition mappings allow for convenient formulation of risk-averse preferences for controlled Markov processes. From now on we assume that the controlled kernels Q_t have values in the set $\mathscr{M}$, which is the set of probability measures on $(\mathscr{X}, \mathscr{B})$ having densities with respect to P_0 in $\mathscr{Y}$.

Definition 6 A one-step conditional risk measure $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$ is a *Markov risk measure* with respect to the controlled Markov process $\{x_t\}$, if there exists a risk transition mapping $\sigma_t : \mathscr{V} \times \mathscr{X} \times \mathscr{M} \to \mathbb{R}$ such that for all $v \in \mathscr{V}$ and for all measurable $u_t \in U_t(x_t)$ we have

$$\rho_t(v(x_{t+1})) = \sigma_t(v, x_t, Q_t(x_t, u_t)).$$

From (15) we deduce that for a Markov risk measure ρ_t there exists a closed convex-valued multifunction $\mathscr{A}_t : \mathscr{X} \times \mathscr{M} \rightrightarrows \mathscr{M}$ such that

$$\rho_t(v(x_{t+1})) = \sup_{\mu \in \mathscr{A}_t(x_t, Q_t(x_t, u_t))} \langle v, \mu \rangle. \tag{17}$$

We shall call the multifunction $\mathscr{S}_t : \mathscr{X} \times \mathscr{U} \rightrightarrows \mathscr{M}$ defined as the composition

$$\mathscr{S}_t(x_t, u_t) = \mathscr{A}_t(x_t, Q_t(x_t, u_t)) \tag{18}$$

the *controlled multikernel* associated with the controlled Markov process $\{x_t\}$ and with the conditional risk mapping ρ_t. Define the function $\Psi_t : \mathscr{V} \times \operatorname{graph}(U_t) \to \mathbb{R}$,

$$\Psi_t(v, x_t, u_t) = \sigma_t(v, x_t, Q_t(x_t, u_t)) = \sup_{\mu \in \mathscr{S}_t(x_t, u_t)} \langle v, \mu \rangle. \tag{19}$$

Obviously, $\rho_t(v(x_{t+1})) = \Psi_t(v, x_t, u_t)$.

In the risk-neutral setting, when $\rho_t(v(x_{t+1})) = \mathbb{E}\left[v(x_{t+1}) | \mathscr{F}_t\right]$ we have a single-valued controlled kernel $\mathscr{S}_t(x_t, u_t) = \{Q_t(x_t, u_t)\}$. Representation (19) means that risk-averse preferences correspond to ambiguity in the transition kernel. We elaborate on this issue in Sect. 10.

Continuity properties of the functions Ψ_t are germane for our analysis. Let us quickly review basic concepts of continuity of multifunctions (see [3] for an extensive presentation). The multifunction $\mathscr{A}$ is *upper semicontinuous* at (x_0, m_0), if for every neighborhood B of $\mathscr{A}(x_0, m_0)$ we can find neighborhoods X_0 of x_0 and B_0 of m_0 such that for all $x \in X_0$ and all $m \in B_0$ we have $\mathscr{A}(x, m) \subset B$. The multifunction $\mathscr{A}$ is *lower semicontinuous* at (x_0, m_0), if for every $\mu \in \mathscr{A}(x_0, m_0)$ and for every sequence $\{x^k, m^k\}$ in the domain of $\mathscr{A}$ converging to (x_0, m_0) we can find a sequence $\mu^k \in \mathscr{A}(x^k, m^k)$ converging to μ. The multifunction $\mathscr{A}$ is *continuous*, if it is both upper and lower semicontinuous at every point.

Proposition 1 *Suppose the kernel Q_t is continuous. If the multifunction $\mathscr{A}_t$ is lower semicontinuous, then for every $v \in \mathscr{V}$ the function $(x_t, u_t) \mapsto \Psi_t(v, x_t, u_t)$ is lower semicontinuous. If $p \in [1, \infty)$ and the multifunction $\mathscr{A}_t$ is upper semicontinuous, then for every $v \in \mathscr{V}$ the function $(x_t, u_t) \mapsto \Psi_t(v, x_t, u_t)$ is upper semicontinuous.*

Proof For a continuous Q_t, the composition $\mathscr{S}_t(x_t, u_t) = \mathscr{A}_t(x_t, Q_t(x_t, u_t))$ inherits the continuity properties of $\mathscr{A}_t$. For every fixed v, the function $\mu \mapsto \langle v, \mu \rangle$ is continuous on $\mathscr{M}$ (in the weak* topology). Moreover, for $p < \infty$ the values of $\mathscr{A}_t$ (and thus of $\mathscr{S}_t$) are weakly* compact in $\mathscr{M}$. The assertion of the theorem follows now from [3, Theorem 1.4.16], whose proof remains valid in our setting as well. □

In our application, lower semicontinuity of $\Psi_t(v, x, \cdot)$ is most important.

Corollary 1 *If $Q_t(x, \cdot)$ is continuous and $\mathscr{A}_t(x, \cdot)$ is lower semicontinuous, then the function $\Psi_t(v, x, \cdot)$ is lower semicontinuous.*

5 Finite horizon problem

We now fix $T \geq 1$ and consider the problem

$$\min_{\Pi} \; J(\Pi, x_1), \tag{20}$$

with $J(\Pi, x_1)$ defined by formula (13),

$$\begin{aligned} J(\Pi, x_1) = c_1(x_1, u_1) + \rho_1 \big(c_2(x_2, u_2) + \rho_2 \big(c_3(x_3, u_3) + \cdots \\ + \rho_{T-1} \big(c_T(x_T, u_T) + \rho_T(c_{T+1}(x_{T+1}))\big) \cdots \big)\big). \end{aligned}$$

Theorem 2 *Assume that the following conditions are satisfied:*

(i) *For every $x \in \mathscr{X}$ the transition kernels $Q_t(x, \cdot)$, $t = 1, \ldots, T$, are continuous;*
(ii) *The conditional risk measures ρ_t, $t = 1, \ldots, T$, are Markov and such that for every $x \in \mathscr{X}$ the multifunctions $\mathscr{A}_t(x, \cdot)$ are lower semicontinuous;*
(iii) *For all measurable selections $u_t(\cdot) \in U_t(\cdot)$, the functions $x \mapsto c_t(x, u_t(x))$, $t = 1, \ldots, T$, and $c_{T+1}(\cdot)$ are elements of $\mathscr{V}$;*
(iv) *For every $x \in \mathscr{X}$ the functions $c_t(x, \cdot)$, $t = 1, \ldots, T$, are lower semicontinuous;*
(v) *For every $x \in \mathscr{X}$ the sets $U_t(x)$, $t = 1, \ldots, T$, are compact.*

Then problem (20) has an optimal solution and its optimal value $v_1(x)$ is the solution of the following dynamic programming equations:

$$v_{T+1}(x) = c_{T+1}(x), \quad x \in \mathscr{X}, \tag{21}$$

$$v_t(x) = \min_{u \in U_t(x)} \{c_t(x, u) + \sigma_t(v_{t+1}, x, Q_t(x, u))\}, \quad x \in \mathscr{X}, \quad t = T, \ldots, 1, \tag{22}$$

where

$$\sigma_t(v, x, Q_t(x,u)) = \sup_{\mu \in \mathscr{A}_t(x, Q_t(x,u))} \langle v, \mu \rangle, \quad t = 1, \ldots, T. \tag{23}$$

Moreover, an optimal Markov policy $\hat{\Pi} = \{\hat{\pi}_1, \ldots, \hat{\pi}_T\}$ *exists and satisfies the equations:*

$$\hat{\pi}_t(x) \in \operatorname*{argmin}_{u \in U_t(x)} \{c_t(x,u) + \sigma_t(v_{t+1}, x, Q_t(x,u))\}, \quad x \in \mathscr{X}, \quad t = T, \ldots, 1. \tag{24}$$

Conversely, every measurable solution of Eqs. (21)–(24) defines an optimal Markov policy $\hat{\Pi}$.

Proof Due to the monotonicity condition (A2) applied to $\rho_t, t = 1, \ldots, T$, problem (20) can be written as follows:

$$\begin{aligned}
&\min_{\pi_1, \ldots, \pi_{T-1}, \pi_T} \{c_1(x_1, u_1) + \rho_1(c_2(x_2, u_2) \\
&\qquad + \cdots + \rho_{T-1}(c_T(x_T, u_T) + \rho_T(c_{T+1}(x_{T+1}))) \cdots)\} \\
&= \min_{\pi_1, \ldots, \pi_{T-1}} \Big\{ c_1(x_1, u_1) + \rho_1(c_2(x_2, u_2) \\
&\qquad + \cdots + \rho_{T-1}\Big(\min_{\pi_T}\big[c_T(x_T, u_T) + \rho_T(c_{T+1}(x_{T+1}))\big]\Big) \cdots \Big) \Big\}.
\end{aligned}$$

This is the fundamental property of interchangeability, discussed in another setting in [47]. In our case it is connected to the time-consistency property, and can also be derived directly from Theorem 1.

Consider the innermost optimization problem. Owing to the Markov structure of the conditional risk measure ρ_T, this problem can be rewritten as follows:

$$\min_{\pi_T} \{c_T(x_T, \pi_T(h_T)) + \sigma_T(v_{T+1}, x_T, Q_T(x_T, \pi_T(h_T)))\}.$$

For every h_T the optimization can be carried out with respect to $u_T = \pi_T(h_T)$. The problem takes on the form

$$\min_{u_T} \{c_T(x_T, u_T) + \sigma_T(v_{T+1}, x_T, Q_T(x_T, u_T))\}. \tag{25}$$

Moreover, the functions in (25) depend on h_T only via x_T, and thus the optimal solution set $\hat{U}_T$ will be a function of x_T. The problem becomes equivalent to the problem in (22) for $t = T$, and its solution is given by (24) for $t = T$. Due to Corollary 1, the function $\sigma_T(v_{T+1}, x_T, Q_T(x_T, \cdot))$ is lower semicontinuous, and the function $c_T(x, \cdot)$ is lower semicontinuous by assumption. As the set $U_t(x)$ is compact, problem (22) for $t = T$ has for every $x \in \mathscr{X}$ an optimal solution $u_T = \hat{\pi}_T(x)$, which is a measurable function of x (*cf.* [42, Theorem 14.37]). As $c_{T+1} \in \mathscr{V}$, it follows from Definition 5

that the function v_T is an element of $\mathscr{V}$ as well. We conclude that problem (20) is equivalent to the problem

$$\min_{\pi_1,\ldots,\pi_{T-1}} \{c_1(x_1,u_1)+\rho_1\left(c_2(x_2,u_2)+\cdots+\rho_{T-1}\left(v_T(x_T)\right)\cdots\right)\},$$

in which the horizon is decreased by 1, and the terminal cost equals $v_T(x_T)$.

Proceeding in this way for $t = T, T-1, \ldots, 1$ we obtain the assertion of the theorem. □

It follows from our proof that the functions $v_t(\cdot)$ calculated in (22) are the optimal values of tail subproblems formulated for a fixed $x_t = x$ as follows:

$$v_t(x) = \min_{\pi_t,\ldots,\pi_T} \{c_t(x,u_t)+\rho_t\left(c(x_{t+1},u_{t+1})\right.$$
$$\left.+\cdots+\rho_{T-1}\left(c(x_T,u_T)+\rho_T(c_{T+1}(x_{T+1}))\right)\cdots\right)\}.$$

We call them *value functions*, as in risk-neutral dynamic programming.

Equations (21)–(24) provide a computational recipe for solving finite horizon problems, which is easily implementable if the state space $\mathscr{X}$ is finite.

6 Discounted measures of risk

Our next step is to define an infinite horizon risk-averse model. Let $\{\mathscr{F}_t\}$ be a filtration on $(\Omega,\mathscr{F})$, with $\mathscr{F}_1 = \{\Omega,\emptyset\}$, and let Z_t, $t = 1,2\ldots$, be an adapted sequence of random variables. Similarly to the construction in Sect. 4, we consider the spaces $\mathscr{Z}_t = \mathscr{L}_p(\Omega,\mathscr{F}_t,P)$, $t = 1,2,\ldots$, with $p \in [1,\infty]$. We define the space $\mathscr{Z} = \mathscr{Z}_1 \times \mathscr{Z}_2 \times \cdots$. A sequence $Z \in \mathscr{Z}$ is almost surely bounded, if

$$\max_t \operatorname{essup} |Z_t(\omega)| < \infty.$$

Consider a sequence of one-step conditional risk mappings $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$, $t = 1,2\ldots$. Fix the *discount factor* $\alpha \in (0,1)$. For $T = 1,2,\ldots$ we define the functionals $\rho^\alpha_{1,T} : \mathscr{Z}_1 \times \cdots \times \mathscr{Z}_T \to \mathbb{R}$ as follows:

$$\rho^\alpha_{1,T}(Z_1,Z_2,\ldots,Z_T) = \rho_{1,T}(Z_1,\alpha Z_2,\ldots,\alpha^{T-1}Z_T)$$
$$= Z_1+\rho_1\left(\alpha Z_2+\rho_2\left(\alpha^2 Z_3+\cdots+\rho_{T-1}(\alpha^{T-1}Z_T)\cdots\right)\right). \tag{26}$$

They are the same as (10) for $t = 1$, but with discounting applied to the sequence $\{Z_t\}$. Finally, we define the functional $\rho^\alpha : \mathscr{Z} \to \mathbb{R}$ as

$$\rho^\alpha(Z) = \lim_{T\to\infty} \rho^\alpha_{1,T}(Z_1,Z_2,\ldots,Z_T). \tag{27}$$

We call it a *discounted measure of risk*.

Theorem 3 *The discounted measure of risk ρ^α is well defined on the set of almost surely bounded sequences $Z \in \mathscr{Z}$, and has the following properties:*

(i) *It is convex;*
(ii) *For all $Z, W \in \mathscr{Z}$, if $Z_t \le W_t$ for all $t = 1, 2, \ldots$, then $\rho^\alpha(Z) \le \rho^\alpha(W)$;*
(iii) *For all $Z \in \mathscr{Z}$, all $t = 1, 2 \ldots$, and all $W_t \in \mathscr{Z}_t$ we have*

$$\rho^\alpha(Z_1, \ldots, Z_t, Z_{t+1} + W_t, \ldots) = \rho^\alpha(Z_1, \ldots, Z_t + \alpha W_t, Z_{t+1}, \ldots);$$

(iv) $\rho^\alpha(\tau Z) = \tau \rho^\alpha(Z), \forall\, Z \in \mathscr{Z},\ \tau \ge 0.$

Proof Each functional (26) can be regarded as defined on the space $\mathscr{Z}$. We first prove that every $\rho^\alpha_{1,T}(\cdot)$ satisfies conditions (i)–(iv). For $T = 1$ these conditions hold trivially. Supposing they are satisfied for T, we shall prove them for $T + 1$. We have

$$\begin{aligned}
&\rho^\alpha_{1,T+1}(Z_1, Z_2, \ldots, Z_{T+1}) \\
&\quad = Z_1 + \rho_1\left(\alpha Z_2 + \cdots + \rho_{T-1}\left(\alpha^{T-1} Z_T + \rho_T(\alpha^T Z_{T+1})\right)\cdots\right) \\
&\quad = \rho^\alpha_{1,T}(Z_1, Z_2, \ldots, Z_T + \alpha\rho_T(Z_{T+1})).
\end{aligned} \tag{28}$$

Thus $\rho^\alpha_{1,T+1}(Z_1, Z_2, \ldots, Z_{T+1})$ is a composition of $\rho^\alpha_{1,T}(\cdot)$ and the mapping

$$(Z_1, \ldots, Z_T, Z_{T+1}) \mapsto (Z_1, \ldots, Z_T + \alpha\rho_T(Z_{T+1})).$$

The first function is convex and nondecreasing, by virtue of (i) and (ii) for T. The second function is convex and nondecreasing, owing to conditions (A1) and (A2) of a conditional risk mapping $\rho_T(\cdot)$. Therefore, their composition is convex and nondecreasing as well. The positive homogeneity property (iv) follows in the same way. It remains to verify (iii) for $T + 1$. Observe that when $t \le T$ we can apply property (iii) for T to the right hand side of (28), and obtain it for $T + 1$ on the left hand side. It remains to consider the case of $t = T + 1$. By (28), axiom (A3), and condition (iv) for $t = T$ we have

$$\begin{aligned}
\rho^\alpha_{1,T+1}(Z_1, Z_2, \ldots, Z_{T+1} + W_t) &= \rho^\alpha_{1,T}(Z_1, Z_2, \ldots, Z_T + \alpha\rho_T(Z_{T+1} + W_T)) \\
&= \rho^\alpha_{1,T}(Z_1, Z_2, \ldots, Z_T + \alpha W_T + \alpha\rho_T(Z_{T+1})).
\end{aligned}$$

Comparing this to (28) we conclude that (iv) holds true for $T + 1$.

We now prove that for every almost surely bounded $Z \in \mathscr{Z}$ the limit in (27) exists. Fix T. Let $C = \max_t \operatorname{essup} |Z_t(\omega)|$. As $-C \le Z_{T+1} \le C$, from axiom (A2) we obtain

$$-C \le \rho_T(Z_{T+1}) \le C. \tag{29}$$

Using this in (28), we get

$$\rho^\alpha_{1,T}(Z_1, \ldots, Z_T - \alpha C) \le \rho^\alpha_{1,T+1}(Z_1, \ldots, Z_T, Z_{T+1}) \le \rho^\alpha_{1,T}(Z_1, \ldots, Z_T + \alpha C).$$

We can now apply property (iii), which we have just proved for $\rho^{\alpha}_{1,T}$, to both sides of the last inequality. We do it $T-1$ times moving the constants $\pm\alpha C$ to the first argument, and we conclude that

$$\rho^{\alpha}_{1,T}(Z_1,\ldots,Z_T)-\alpha^T C\le\rho^{\alpha}_{1,T+1}(Z_1,\ldots,Z_T,Z_{T+1})\le\rho^{\alpha}_{1,T}(Z_1,\ldots,Z_T)+\alpha^T C.$$

This implies that the sequence in (27) is a Cauchy sequence. As the functions $\rho^{\alpha}_{1,T}(\cdot)$, $T=1,2,\ldots$, satisfy conditions (i)–(iv), the limit function $\rho^{\alpha}(\cdot)$ satisfies these conditions as well. □

7 Discounted infinite horizon problem: value iteration

Consider now application of the discounted risk measure (27) to a controlled Markov process $\{x_t\}$, $t=1,2,\ldots$. We assume that the controlled Markov model is *stationary*, that is, there exist a control set $U:\mathscr{X}\rightrightarrows\mathscr{U}$, a controlled transition kernel $Q:\mathrm{graph}(U)\to\mathscr{P}$, and a cost function $c:\mathrm{graph}(U)\to\mathbb{R}$, such that such that $U_t=U$, $Q_t=Q$, and $c_t=c$, for $t=1,2,\ldots$.

Each policy $\Pi=\{\pi_t\}_{t=1}^{\infty}$ results in a cost sequence $Z_t=c(x_t,u_t)$, $t=1,2,\ldots$. We use a discounted measure of risk to evaluate the cost of the policy Π:

$$\begin{aligned}J(\Pi,x_1)&=\rho^{\alpha}\left(c(x_1,u_1),c(x_2,u_2),\cdots\right)\\&=c(x_1,u_1)+\rho_1\left(\alpha c(x_2,u_2)+\rho_2\left(\alpha^2c(x_3,u_3)+\cdots\right)\right).\end{aligned}\tag{30}$$

The last expression should be understood as the limit (27). We assume that the conditional risk measures ρ_t are Markov, in the sense of Definition 6, and that there exists a risk transition mapping $\sigma:\mathscr{X}\times\mathscr{V}\times\mathscr{M}\to\mathbb{R}$ such that $\sigma_t=\sigma$, for all $t=1,2\ldots$. We call such a sequence of Markov risk mappings *stationary*. For such a sequence, there exists a multifunction $\mathscr{A}:\mathscr{X}\times\mathscr{M}\rightrightarrows\mathscr{M}$ such that representation (17) holds true with $\mathscr{A}_t=\mathscr{A}$, $t=1,2\ldots$, that is,

$$\sigma\left(v,x,Q(x,u)\right)=\sup_{\mu\in\mathscr{A}(x,Q(x,u))}\langle v,\mu\rangle.\tag{31}$$

We are interested in the infinite horizon problem

$$\min_{\Pi}\ J(\Pi,x_1).\tag{32}$$

In order to analyze this problem we need to make several assumptions. It is most convenient to state these assumptions within the formulation of our main result. Its proof will be deferred until after several technical results.

Theorem 4 *Assume that the following conditions are satisfied:*

(i) *For every $x\in\mathscr{X}$ the transition kernel $Q(x,\cdot)$ is continuous;*
(ii) *The conditional risk measures ρ_t, $t=1,\ldots,T$, are Markov, stationary and such that for every $x\in\mathscr{X}$ the multifunction $\mathscr{A}(x,\cdot)$ is lower semicontinuous;*

(iii) *The function $c(\cdot,\cdot)$ is nonnegative and uniformly bounded from above;*
(iv) *For every $x \in \mathscr{X}$ the function $c(x,\cdot)$ is lower semicontinuous;*
(v) *For every $x \in \mathscr{X}$ the set $U(x)$ is compact.*

Then problem (32) has an optimal solution and its optimal value $\hat{v}(x)$, as a function of the initial state $x_1 = x$, satisfies the following dynamic programming equation:

$$v(x) = \min_{u \in U(x)} \{c(x,u) + \alpha\sigma\,(v, x, Q(x,u))\}, \quad x \in \mathscr{X}. \tag{33}$$

Moreover, an optimal stationary Markov policy $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}, \dots\}$ exists and satisfies the equation:

$$\hat{\pi}(x) \in \operatorname*{argmin}_{u \in U(x)} \left\{c(x,u) + \alpha\sigma\left(\hat{v}, x, Q(x,u)\right)\right\}, \quad x \in \mathscr{X}. \tag{34}$$

Conversely, every bounded solution of equation (33) is the optimal value of problem (32), and every measurable solution of (34) defines an optimal stationary Markov policy.

The assumption that the function $c(\cdot,\cdot)$ is nonnegative can be replaced by the assumption of uniform boundedness of this function from below, because we can always add a sufficiently large constant to the function $c(\cdot,\cdot)$ to make it nonnegative. Indeed, adding a constant C to each $Z_t = c(x_t, u_t)$, due to Theorem 3(iii), has the following effect on the discounted measure of risk:

$$\rho^\alpha(Z_1 + C, Z_2 + C, Z_3 + C, \dots) = \rho^\alpha(Z_1, Z_2, Z_3, \dots) + C + \alpha C + \alpha^2 C + \cdots.$$

Thus the problem of minimizing $\rho^\alpha(Z_1, Z_2, Z_3, \dots)$ is equivalent to the problem of minimizing $\rho^\alpha(Z_1 + C, Z_2 + C, Z_3 + C, \dots)$.

If C is the uniform upper bound on $c(\cdot,\cdot)$, then from Theorem 3(ii)–(iii) we obtain the following uniform bound on $\hat{v}(x)$:

$$0 \le J(\Pi, x) = \rho^\alpha(Z_1, Z_2, Z_3, \dots) \le \rho^\alpha(C, C, C, \dots) = \frac{C}{1-\alpha}.$$

Thus the optimal value $\hat{v}(x)$ is well defined and is uniformly bounded.

In order to prove Theorem 4, we need to establish several auxiliary results. For a measurable selection $\pi : \mathscr{X} \to \mathscr{U}$ of $U(\cdot)$ we define the operator $\mathfrak{D}_\pi : \mathscr{V} \to \mathscr{V}$ as follows:

$$\mathfrak{D}_\pi v(x) = c(x, \pi(x)) + \alpha\sigma\,(v, x, Q(x, \pi(x))), \quad x \in \mathscr{X}.$$

We also define the operator $\mathfrak{D} : \mathscr{V} \to \mathscr{V}$ by the following relation:

$$\mathfrak{D} v(x) = \min_{u \in U(x)} \{c(x,u) + \alpha\sigma\,(v, x, Q(x,u))\}, \quad x \in \mathscr{X}.$$

With this notation, Eq. (33) can be compactly written as

$$v = \mathfrak{D} v, \tag{35}$$

and thus properties of the operator $\mathfrak{D}$ are germane for our problem.

Lemma 1 *The operators $\mathfrak{D}_\pi$ and $\mathfrak{D}$ are nondecreasing in the sense that $v \geq w$ implies $\mathfrak{D}_\pi v \geq \mathfrak{D}_\pi w$ and $\mathfrak{D} v \geq \mathfrak{D} w$.*

Proof Both results follow from (31) and the fact that the sets $\mathscr{A}(x, Q(x,u))$ contain only probability measures. □

Lemma 2 *The operators $\mathfrak{D}_\pi$ and $\mathfrak{D}$ are contraction mappings with modulus α on the space $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$, that is, for all $v, w \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ we have*

$$\|\mathfrak{D}_\pi v - \mathfrak{D}_\pi w\|_\infty \leq \alpha \|v - w\|_\infty, \tag{36}$$

$$\|\mathfrak{D} v - \mathfrak{D} w\|_\infty \leq \alpha \|v - w\|_\infty. \tag{37}$$

Proof For any probability measure μ we have

$$\langle v, \mu \rangle \leq \langle w, \mu \rangle + \|v - w\|_\infty. \tag{38}$$

Taking the supremum over $\mu \in \mathscr{A}(x, Q(x,u))$ preserves this inequality, and thus, for all $\pi(x) \in U(x)$ we have

$$\mathfrak{D}_\pi v(x) \leq \mathfrak{D}_\pi w(x) + \alpha \|v - w\|_\infty, \quad x \in \mathscr{X}. \tag{39}$$

Reversing the roles of v and w we obtain (36). Inequality (37) follows from taking for every $x \in \mathscr{X}$ the infimum of both sides of (39) with respect to $\pi(x)$. □

Lemma 3 (i) *If $w \in \mathscr{V}_+$ and $w \geq \mathfrak{D} w$, then $w \geq \hat{v}$;*
(ii) *If $w \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ and $w \leq \mathfrak{D} w$, then $w \leq \hat{v}$.*

Proof (i) Let π be the measurable selection of U for which

$$w \geq \mathfrak{D}_\pi w. \tag{40}$$

It exists, due to assumptions (i)–(v) of Theorem 4, exactly as in the proof of Theorem 2. Applying the operator $\mathfrak{D}_\pi$ to both sides of inequality (40) and using Lemma 1, we obtain

$$w \geq [\mathfrak{D}_\pi]^T w, \quad T = 1, 2, \ldots. \tag{41}$$

The right hand side of this inequality represents the cost of a finite horizon problem with the stationary Markov policy $\Pi = \{\pi, \pi, \ldots\}$ and with the final cost $c_{T+1}(\cdot) = w(\cdot)$.

Denoting by $Z_t = c(x_t, \pi(x_t)), t = 1, 2, \dots$, the cost sequence in this system, we get

$$[\mathfrak{D}_\pi]^T w = \rho^\alpha_{1,T+1}(Z_1, Z_2, \dots, Z_T, w) \geq \rho^\alpha_{1,T}(Z_1, Z_2, \dots, Z_T).$$

In the last inequality we used the fact that $w \geq 0$ and Theorem 3(ii). Combining the last two inequalities and passing to the limit with $T \to \infty$ we conclude that

$$w(x) \geq \rho^\alpha(Z_1, Z_2, \dots) = J(\Pi, x), \quad x \in \mathscr{X}.$$

This proves (i).

(ii) Consider an arbitrary (not necessarily Markov) policy $\Pi = \{\pi_1, \pi_2, \dots\}$ and the resulting cost sequence $Z_t = c(x_t, u_t), t = 1, \dots, T$, in a finite horizon problem with the terminal cost $w(x_{T+1})$, We write $u_t = \pi_t(h_t)$, where h_t is the history of the process up to time t. We have

$$\begin{aligned}
&\rho^\alpha_{1,T+1}(Z_1, \dots, Z_T, w(x_{T+1})) \\
&\quad = Z_1 + \rho_1\left(\alpha Z_2 + \dots + \rho_{T-1}\left(\alpha^{T-1} Z_T + \rho_T(\alpha^T w(x_{T+1}))\right)\cdots\right) \\
&\quad = c_1 + \rho_1\left(\alpha Z_2 + \dots + \rho_{T-1}\left(\alpha^{T-1}[Z_T + \alpha\rho_T(w(x_{T+1}))]\right)\cdots\right).
\end{aligned}$$

By assumption (ii), the expression in brackets can be estimated as follows:

$$Z_T + \alpha\rho_T(w(x_{T+1})) \geq \mathfrak{D}w(x_T) \geq w(x_T).$$

Therefore, $\rho^\alpha_{1,T+1}(Z_1, \dots, Z_T, w(x_{T+1})) \geq \rho^\alpha_{1,T}(Z_1, \dots, Z_{T-1}, w(x_T))$. Proceeding in this way, we conclude that

$$\rho^\alpha_{1,T+1}(Z_1, \dots, Z_T, w(x_{T+1})) \geq w(x_1).$$

Let C be an upper bound on $|w(x)|, x \in \mathscr{X}$. By property (ii) and (iii) of Theorem 3, the cost of policy Π can be estimated as follows:

$$\rho^\alpha_{1,T}(Z_1, \dots, Z_T) \geq \rho^\alpha_{1,T+1}(Z_1, \dots, Z_T, w(x_{T+1})) - C\alpha^T.$$

Combining the last two inequalities and passing to the limit with $T \to \infty$ we conclude that for every policy Π

$$J(\Pi, x) \geq w(x).$$

Therefore the infimum over all Π is bounded from below by $w(x)$, as claimed. □

Lemma 4 *The value v of a stationary policy $\Pi = \{\pi, \pi, \dots\}$ is the unique bounded solution of the equation*

$$v = \mathfrak{D}_\pi v. \tag{42}$$

Proof By Lemma 2, the operator $\mathfrak{D}_\pi$ is a contraction in $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ and thus Eq. (42) has a unique bounded solution v. Observe that for every $T = 1, 2, \ldots$ the expression $[\mathfrak{D}_\pi]^T v$ represents the cost of a T-period problem having $v(\cdot)$ as its terminal cost. Due to (42) this is the same as v. Passing to the limit we obtain the result. □

We are now ready to provide a simple proof of the main theorem.

Proof of Theorem 4 By Lemma 2, the operator $\mathfrak{D}$ is a contraction in $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$. By Banach's contraction mapping principle, Eq. (35), which is a compact form of (33), has a unique bounded solution v. Owing to Lemma 3, $v = \hat{v}$.

Consider the problem on the right hand side of (34). Owing to the assumptions of Theorem 4, an optimal solution of this problem exists, exactly in the same way as in the proof of Theorem 2. Applying the operator $\mathfrak{D}_{\hat{\pi}}$ to both sides of the equation $\hat{v} = \mathfrak{D}_{\hat{\pi}} \hat{v}$, we obtain

$$\hat{v} = \left[\mathfrak{D}_{\hat{\pi}}\right]^T \hat{v}, \quad T = 1, 2, \ldots.$$

The right hand side of this inequality represents the cost of a finite horizon problem with the stationary Markov policy $\{\hat{\pi}, \hat{\pi}, \ldots\}$ and with the final cost $c_{T+1}(\cdot) = \hat{v}(\cdot)$. Passing to the limit with $T \to \infty$ we conclude that $\hat{v}$ is the cost of the policy $\{\hat{\pi}, \hat{\pi}, \ldots\}$ in the infinite horizon problem.

Suppose another optimal Markov policy $\Pi = \{\pi, \pi, \ldots\}$ exists, but $\hat{v} \neq \mathfrak{D}_\pi \hat{v}$. The value $v(\cdot)$ of policy Π satisfies Eq. (42). Thus $v \neq \hat{v}$, a contradiction. □

Our analysis of the dynamic programming Eq. (33) suggests the following iterative method, corresponding to the classical *value iteration* method in risk-neutral dynamic programming. We start from an arbitrary function $v^1 \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ and generate a sequence

$$v^{k+1} = \mathfrak{D} v^k, \quad k = 1, 2, \ldots. \tag{43}$$

Theorem 5 *If conditions of Theorem 4 are satisfied, then the sequence $\{v^k\}$ generated by the value iteration method is convergent linearly in $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ to the optimal value function $\hat{v}$, with the quotient α. Moreover, If $v^1 = 0$, then the sequence $\{v^k\}$ is nondecreasing, while for $v^1 \geq \sup\{c(x, u) : x \in \mathscr{X}, u \in U(x)\}$ the sequence $\{v^k\}$ is nonincreasing.*

Proof Linear convergence of the sequence $\{v^k\}$ to $\hat{v}$ follows from the contraction property of $\mathfrak{D}$ in $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$, established in Lemma 2, by virtue of Banach's contraction mapping principle. Suppose $v^1 = 0$. Then $v^2 \geq v^1$, and Lemma 1 implies that $v^{k+1} \geq v^k$, $k = 1, 2, \ldots$. The case of v^1 larger than the maximum cost per stage is similar. □

8 Policy iteration

Consider now another approach to solving the dynamic programming equations, extending the risk-neutral *policy iteration* method (see [5,22]). In the new method, for

$k = 0, 1, 2, \ldots$, given a stationary Markov policy $\Pi^k = \{\pi^k, \pi^k, \ldots\}$, we find the corresponding value function $v^k \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$ by solving the equation

$$v(x) = c(x, \pi^k(x)) + \alpha\sigma\left(v, x, Q(x, \pi^k(x))\right), \quad x \in \mathscr{X}. \tag{44}$$

Then we find the next policy $\pi^{k+1}(\cdot)$ as a measurable function satisfying the relation

$$\pi^{k+1}(x) \in \operatorname*{argmin}_{u \in U(x)} \left\{c(x, u) + \alpha\sigma\left(v^k, x, Q(x, u)\right)\right\}, \quad x \in \mathscr{X}. \tag{45}$$

After that, k is increased by 1, and the iteration continues.
Clearly, Eqs. (44)–(45) can be compactly written as follows:

$$v^k = \mathfrak{D}_{\pi^k} v^k, \tag{46}$$

$$\mathfrak{D}_{\pi^{k+1}} v^k = \mathfrak{D} v^k. \tag{47}$$

Theorem 6 *Assume that the assumptions of Theorem 4 are satisfied. Then the sequence of functions $v^k, k = 0, 1, 2, \ldots$, is nonincreasing and convergent to the unique bounded solution $\hat{v}(\cdot)$ of the dynamic programming equation (33).*

Proof It follows from (47) and (46) that

$$\mathfrak{D}_{\pi^{k+1}} v^k = \mathfrak{D} v^k \le \mathfrak{D}_{\pi^k} v^k = v^k.$$

Using the monotonicity property of the operator $\mathfrak{D}_{\pi^{k+1}}$ (Lemma 1), we obtain that

$$\left[\mathfrak{D}_{\pi^{k+1}}\right]^T v^k \le \mathfrak{D} v^k \le v^k, \quad T = 0, 1, 2, \ldots.$$

By Lemma 2 and Banach's contraction mapping principle, the left hand side of this inequality converges to v^{k+1}, as $t \to \infty$. Thus

$$v^{k+1} \le \mathfrak{D} v^k \le v^k. \tag{48}$$

Therefore, the sequence $\{v^k\}$ is nonincreasing. Each v^k, as the value of not necessarily optimal policy Π^k, satisfies the inequality $v^k \ge \hat{v}$. From the first part of (48) we deduce that

$$0 \le v^{k+1} - \hat{v} \le \mathfrak{D} v^k - \hat{v} = \mathfrak{D} v^k - \mathfrak{D}\hat{v}.$$

Owing to the contraction property of the operator $\mathfrak{D}$ established in Lemma 2, we conclude that

$$\|v^{k+1} - \hat{v}\|_\infty \le \alpha \|v^k - \hat{v}\|_\infty.$$

Hence, the sequence $\{v^k\}$ is convergent to $\hat{v}$. □

If $\mathfrak{D}_{\pi^{k+1}} v^k = v^k$, then it follows from (47) that $v^k = \mathfrak{D} v^k$. By virtue of Theorem 4, π^k and π^{k+1} are optimal policies, and v^k is the optimal value function.

Consider the special case of finitely many possible policies. Only finitely many different value functions v^k can be generated by the policy iteration method, and they form a nondecreasing sequence. Thus, a value function can be repeated only if it occurs in two consecutive steps, but then we know that it is optimal. Consequently, the policy iteration method is convergent in finitely many steps.

9 Specialized nonsmooth Newton method

At every step of the policy iteration method, in order to evaluate the current policy π^k, we have to solve the nonlinear Eq. (44). Observe that the mapping $v \mapsto \sigma\left(v, x, Q(x, \pi^k(x))\right)$ is *nonsmooth*, in general. Because of that, Eq. (44) is much harder than the corresponding *linear* equation in the risk-neutral case.

Our aim is to propose a version of the nonsmooth Newton method for solving this equation. The general theory of nonsmooth Newton methods originated in [31,44]; it is discussed extensively in [26, Chapter 10]. In our case, owing to the special form of Eq. (44), we can use the simplest algorithm with linear auxiliary problems, as presented in [31] and in [26, Sect. 10.1]. We can also provide a direct proof of its global and monotonic convergence.

To simplify notation, we suppress the index k of the current iteration of the policy method, and we use ℓ to denote iterations of the Newton method.

Consider Eq. (44), which we write as

$$v(x) = \bar{c}(x) + \alpha \sup_{\mu \in \bar{\mathscr{A}}(x)} \langle v, \mu \rangle, \quad x \in \mathscr{X}, \tag{49}$$

with $\bar{c}(x) = c(x, \pi^k(x))$ and $\bar{\mathscr{A}}(x) = \mathscr{A}(x, Q(x, \pi^k(x)))$. Suppose at iteration ℓ of the Newton method we have a certain approximation v_ℓ of the solution of (49). We calculate a kernel μ_ℓ by

$$\mu_\ell(x) \in \operatorname*{argmax}_{\mu \in \bar{\mathscr{A}}(x)} \langle v_\ell, \mu \rangle, \quad x \in \mathscr{X}.$$

Under the assumptions of Theorem 6 with $p \in [1, \infty)$, the set $\bar{\mathscr{A}}(x)$ is weakly* compact, and thus an optimal kernel μ_ℓ exists. It may not be unique, but this will not play any role in our method. Then we solve the *linear* equation

$$v(x) = \bar{c}(x) + \alpha \langle v, \mu_\ell(x) \rangle, \quad x \in \mathscr{X}. \tag{50}$$

We denote its solution by $v_{\ell+1}$, increase ℓ by one, and continue.

Observe that the linear Eq. (50) is the evaluation of the expected cost in the process with transition kernel μ_ℓ, which can be done in the same way as in the policy iteration method for a risk-neutral model. The essence of our method in this case is to construct a sequence of risk-neutral models, for $\ell = 1, 2, \ldots$, to evaluate discounted risk.

Theorem 7 *Assume conditions of Theorem 4 and let* $\mathscr{V} = \mathscr{L}_p(\mathscr{X}, \mathscr{B}, P_0)$ *with* $p \in [1, \infty)$. *Then for every initial function* v_1 *the sequence* $\{v_\ell\}$ *generated by the Newton method is convergent to the unique solution* v^* *of Eq. (49). Moreover, the sequence is monotone:* $v_{\ell+1} \geq v_\ell$, $\ell = 2, 3, \ldots$.

Proof Let us denote by $\mathfrak{M}_\ell$ the affine operator appearing on the right hand side of Eq. (50):

$$[\mathfrak{M}_\ell v](x) = \bar{c}(x) + \alpha \langle \mu_\ell(x), v \rangle.$$

The operator $\mathfrak{M}_\ell$ is a contraction in $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$. By virtue of Banach's contraction mapping principle, Eq. (50) has a unique solution $v_{\ell+1} \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$.

By construction, for every $v \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$, we have

$$\mathfrak{M}_\ell v \leq \mathfrak{D}_{\pi^k} v. \tag{51}$$

As $\mu_\ell(x)$ is a probability measure for every x, the operator $\mathfrak{M}_\ell$ is nondecreasing. Applying $\mathfrak{M}_\ell$ to both sides of (51) we get

$$[\mathfrak{M}_\ell]^2 v \leq \mathfrak{M}_\ell \mathfrak{D}_{\pi^k} v \leq [\mathfrak{D}_{\pi^k}]^2 v.$$

Continuing in this way, we conclude that for every $T = 1, 2, \ldots$

$$[\mathfrak{M}_\ell]^T v \leq [\mathfrak{D}_{\pi^k}]^T v.$$

Let us pass to the limit on both sides of this inequality with $T \to \infty$. Owing to the contraction properties of $\mathfrak{M}_\ell$ and $\mathfrak{D}_{\pi^k}$, the left hand side converges to $v_{\ell+1}$, the unique solution of (50), while the right hand side converges to the value function $\bar{v}$, the unique solution of the dynamic programming Eq. (49). Consequently, $v_{\ell+1} \leq \bar{v}$.

In view of the argument in the preceding paragraph, we have $v_\ell \leq \bar{v}$ for all $\ell \geq 2$. Then

$$v_\ell \leq \mathfrak{D}_{\pi^k} v_\ell = \mathfrak{M}_\ell v_\ell.$$

Applying the monotone operator $\mathfrak{M}_\ell$ to both sides we conclude that

$$v_\ell \leq \mathfrak{D}_{\pi^k} v_\ell \leq [\mathfrak{M}_\ell]^T v_\ell \overset{T\to\infty}{\longrightarrow} v_{\ell+1}. \tag{52}$$

It follows that the sequence $\{v_\ell\}$ is nondecreasing and bounded from above. Consequently, it has a limit $v^* \in \mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$. Passing to the limit with $\ell \to \infty$ in the relations (52), we conclude that

$$v^* \leq \mathfrak{D}_{\pi^k} v^* \leq v^*.$$

This means that the limit v^* is the unique solution of Eq. (49). □

10 Relation to min-max Markov control models

Close and intriguing relations exist between our risk-averse control models and min-max Markov control problems, as discussed in [20,33] and the references therein. We already mentioned in Sect. 4 that risk-averse preferences may be interpreted as ambiguity in the transition kernel, where the distribution of the next state is chosen from the set $\mathscr{S}_t(x_t, u_t) = \mathscr{A}_t(x_t, Q_t(x_t, u_t))$. We can, therefore, envisage a *min–max control model*, in which there are two players. Player 1 chooses at each time t and at each state x_t a control $u_t \in U_t(x_t)$. Player 2, the opponent, given x_t and u_t, selects a probability measure $\mu_t \in \mathscr{S}_t(x_t, u_t)$ which describes the distribution of the next state x_{t+1}.

Let us start from the finite horizon problem. We denote, as before, by $\Pi = (\pi_1, \dots, \pi_T)$ a policy of Player 1, where $\pi_t : \mathscr{H}_t \to \mathscr{U}$, with $\pi_t(x_1, \dots, x_t) \in U_t(x_t)$. Similarly, $\Gamma = (\gamma_1, \dots, \gamma_T)$ is a policy of Player 2, where $\gamma_t : \mathscr{H}_t \times \mathscr{U} \to \mathscr{M}$. The feasible set of Player 2 depends on the policy of Player 1: $\mathbb{G}(\Pi) = \{\Gamma : \gamma_t(x_1, \dots, x_t, u_t) \in \mathscr{S}_t(x_t, u_t), \ t = 1, 2, \dots\}$. The min–max control problem is defined as follows:

$$\min_{\Pi} \max_{\Gamma \in \mathbb{G}(\Pi)} \mathbb{E}\left[\sum_{t=1}^{T} c_t(x_t, u_t) + c_{T+1}(x_{T+1})\right], \tag{53}$$

where for each t we have $u_t = \pi_t(x_1, \dots, x_t)$, and the conditional distribution of x_{t+1}, given $\{x_1, \dots, x_t, u_t\}$, is $\mu_t = \gamma_t(x_1, \dots, x_t, u_t)$.

Corollary 2 *Suppose all assumptions of Theorem 2 are satisfied. Then every optimal solution of problem (20) is also an optimal solution of problem (53). Furthermore, for every Markov optimal policy $\hat{\Pi} = \{\hat{\pi}_1, \dots, \hat{\pi}_T\}$ of Player 1, the sequence of transition kernels $\{\hat{\mu}_1, \dots, \hat{\mu}_T\}$ such that*

$$\hat{\mu}_t(x) \in \operatorname*{argmax}_{\mu \in \mathscr{S}_t(x, \hat{\pi}_t(x))} \langle \hat{v}_{t+1}, \mu \rangle, \quad x \in \mathscr{X}, \quad t = 1, \dots, T,$$

is an optimal Markov policy of Player 2.

Proof The min-max control problem (53) is a special case of the problem analyzed in [20, Sect. 3]. Assumptions of Theorem 2 imply assumptions of [20, Thm. 3.1], and the dynamic programming equations (21)–(22) coincide with Eq. (3.11) in [20]. Therefore solutions of both problems are identical. □

Similar observation hold true for the *stationary infinite horizon discounted problem*. A policy of Player 1 is a sequence $\Pi = (\pi_1, \pi_2, \dots)$ with each $\pi_t : \mathscr{H}_t \to \mathscr{U}$ such that $\pi_t(x_1, \dots, x_t) \in U(x_t)$. A policy of Player 2 is a sequence $\Gamma = (\gamma_1, \gamma_2, \dots)$ with each $\gamma_t : \mathscr{H}_t \times \mathscr{U} \to \mathscr{M}$. The feasible set of Player 2, $\mathbb{G}(\Pi)$, is defined by the conditions $\gamma_t(x_1, \dots, x_t, u_t) \in \mathscr{S}(x_t, u_t) = \mathscr{A}(x_t, Q(x_t, u_t))$, $t = 1, 2, \dots$, and depends on the policy of Player 1. For a given $\alpha \in (0, 1)$, our aim is to find a policy $\Pi = \{\pi_t\}_{t=1}^{\infty}$ so as to minimize the worst expected discounted cost:

$$\min_{\Pi} \max_{\Gamma \in \mathbb{G}(\Pi)} \mathbb{E}\left[\sum_{t=1}^{\infty} \alpha^{t-1} c(x_t, u_t)\right]. \tag{54}$$

In problem (54), for each t the control of Player 1 equals $u_t = \pi_t(x_1, \ldots, x_t)$, and the conditional distribution of x_{t+1}, given $x_1, \ldots, x_t, u_t$, is the control of Player 2: $\mu_t = \gamma_t(x_1, \ldots, x_t, u_t)$.

Corollary 3 *Suppose all assumptions of Theorem 4 are satisfied. Then every optimal solution of problem (32) is also an optimal solution of problem (54). Furthermore, for every stationary Markov optimal policy $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}, \ldots\}$ of Player 1, the sequence of transition kernels $\{\hat{\mu}, \hat{\mu}, \ldots\}$ such that*

$$\hat{\mu}(x) \in \operatorname*{argmax}_{\mu \in \mathscr{S}(x, \hat{\pi}(x))} \langle \hat{v}, \mu \rangle, \quad x \in \mathscr{X},$$

is an optimal stationary Markov policy of Player 2.

Proof Problem (54) is a special case of the problem analyzed in [20, Sect. 4]. Assumptions of Theorem 4 imply assumptions of [20, Thm. 4.2], and the dynamic programming Eq. (33) is the same as Eq. (4.4) in [20]. This implies that the solutions of both problems are identical. □

It may be worth stressing that the min-max problems (53) and (54) are specific in the sense that the feasible sets of the actions of Player 2 depend on the actions of Player 1. Therefore the "min" and the "max" operators cannot be, in general, interchanged, even if we allow mixed strategies of the players.

Our nonsmooth Newton method within the policy iteration exploits the specific structure of the risk-averse model. However, we hope that our ideas may be useful for developing policy iteration for other min–max Markov decision models.

Acknowledgments This research was supported by the NSF award CMMI-0965689. The author is very grateful to the Department of Operations Research and Financial Engineering of Princeton University for providing him with excellent conditions to carry out this research during his sabbatical leave. Thanks are also offered to two anonymous Referees for their observations that helped to improve the manuscript in a substantial way.

References

1. Artzner, P., Delbaen, F., Eber, J.M., Heath, D.: Coherent measures of risk. Math. Finance **9**, 203–228 (1999)
2. Artzner, P., Delbaen, F., Eber, J.-M., Heath, D., Ku, H.: Coherent multiperiod risk adjusted values and Bellmans principle. Ann. Oper. Res. **152**, 5–22 (2007)
3. Aubin, J.-P., Frankowska, H.: Set-Valued Analysis. Birkhäuser, Boston (1990)
4. Bellman, R.: On the theory of dynamic programming. Proc. Natl. Acad. Sci **38**, 716 (1952)
5. Bellman, R.: Applied Dynamic Programming. Princeton University Press, Princeton (1957)
6. Bertsekas, D., Shreve, S.E.: Stochastic Optimal Control. The Discrete Time Case. Academic Press, New York (1978)
7. Boda, K., Filar, J.A.: Time consistent dynamic risk measures. Math. Methods Oper. Res. **63**, 169–186 (2006)

8. Cheridito, P., Delbaen, F., Kupper, M.: Dynamic monetary risk measures for bounded discrete-time processes. Electron. J. Probab. **11**, 57–106 (2006)
9. Chung, K.-J., Sobel, M.J.: Discounted MDPs: distribution functions and exponential utility maximization. SIAM J. Control Optim. **25**, 49–62 (1987)
10. Delbaen, F.: Coherent risk measures on general probability spaces, In essays in honour of Dieter Sondermann. Springer, Berlin (2002)
11. Eichhorn, A., Römisch, W.: Polyhedral risk measures in stochastic programming. SIAM J. Optim. **16**, 69–95 (2005)
12. Fleming, W.H., Sheu, S.J.: Optimal long term growth rate of expected utility of wealth. Ann. Appl. Probab. **9**, 871–903 (1999)
13. Fleming, W.H., Sheu, S.J.: Risk-sensitive control and an optimal investment model. Math. Finance **10**, 197–213 (2000)
14. Föllmer, H., Penner, I.: Convex risk measures and the dynamics of their penalty functions. Stat. Decis. **24**, 61–96 (2006)
15. Föllmer, H., Schied, A.: Convex measures of risk and trading constraints. Finance Stoch. **6**, 429–447 (2002)
16. Föllmer, H., Schied, A.: Stochastic Finance. An Introduction in Discrete Time. de Gruyter, Berlin (2004)
17. Fritelli, M., Rosazza Gianin, E.: Putting order in risk measures. J. Bank. Finance **26**, 1473–1486 (2002)
18. Frittelli, M., Rosazza Gianin, E. : Dynamic convex risk measures. In: Szegö, G. (ed.) Risk Measures for the 21st Century, pp. 227–248. Wiley, Chichester (2005)
19. Fritelli, M., Scandolo, G.: Risk measures and capital requirements for processes. Math. Finance **16**, 589–612 (2006)
20. González-Trejo, J.I., Hernández-Lerma, O., Hoyos-Reyes, L.F.: Minimax control of discrete-time stochastic systems. SIAM J. Control Optim. **41**, 1626–1659 (2003)
21. Hernández-Lerma, O., Lasserre, J.B.: Discrete-time Markov Control Processes. Basic Optimality Criteria. Springer, New York (1996)
22. Howard, R.A.: Dynamic Programming and Markov Processes. Wiley, New York (1960)
23. Jaquette, S.C.: Markov decision processes with a new optimality criterion: Discrete time. Ann. Stat. **1**, 496–505 (1973)
24. Jaquette, S.C.: A utility criterion for Markov decision processes. Manag. Sci. **23**, 43–49 (1976)
25. Jobert, L., Rogers, L.C.G.: Valuations and dynamic convex risk measures. Math. Finance **18**, 1–22 (2008)
26. Klatte, D., Kummer, B.: Nonsmooth Equations in Optimization. Kluwer, Dordrecht (2002)
27. Klein Haneveld, W.: *Duality in stochastic linear and dynamic programming*. Lecture notes economics and mathematical systems 274. Springer, Berlin (1986)
28. Klöppel, S., Schweizer, M.: Dynamic indifference valuation via convex risk measures. Math. Finance **17**, 599–627 (2007)
29. Koopmans, T.C.: Stationary ordinal utility and impatience. Econometrica **28**, 287–309 (1960)
30. Kreps, M.K., Porteus, E.L.: Temporal resolution of uncertainty and dynamic choice theory. Econometrica **46**, 185–200 (1978)
31. Kummer, B.: Newton's method for non-differentiable functions. In: Guddat, J. et al. (eds.) Advances in Mathematical Optimization, pp. 114–125. Academie Verlag, Berlin (1988)
32. Kushner, H.J.: Introduction to Stochastic Control. Holt, Rhinehart, and Winston, New York (1971)
33. Küenle, H.-U.: Stochastiche Spiele und Entscheidungsmodelle. B. G. Teubner, Leipzig (1986)
34. Leitner, J.: A short note on second-order stochastic dominance preserving coherent risk measures. Math. Finance **15**, 649–651 (2005)
35. Ogryczak, W., Ruszczyński, A.: From stochastic dominance to mean-risk models: Semideviations as risk measures. Eur. J. Oper. Res. **116**, 33–50 (1999)
36. Ogryczak, W., Ruszczyński, A.: On consistency of stochastic dominance and mean-semideviation models. Math. Program. **89**, 217–232 (2001)
37. Ogryczak, W., Ruszczyński, A.: Dual stochastic dominance and related mean-risk models. SIAM J. Optim. **13**(1), 60–78 (2002)
38. Pflug, G.Ch., Römisch, W.: Modeling, Measuring and Managing Risk. World Scientific, Singapore (2007)
39. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)

40. Riedel, F.: Dynamic coherent risk measures. Stoch. Process. Appl. **112**, 185–200 (2004)
41. Rockafellar, R.T., Uryasev, S.P.: Conditional value-at-risk for general loss distributions. J. Bank. Finance **26**, 1443–1471 (2002)
42. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis. Springer, Berlin (1998)
43. Rockafellar, R.T., Uryasev, S., Zabarankin, M.: Deviation measures in risk analysis and optimization. Finance Stoch. **10**, 51–74 (2006)
44. Robinson, S.M.: Newton's method for a class of nonsmooth functions. Set-Valued Anal. **2**, 291–305 (1994)
45. Ruszczyński, A., Shapiro, A.: Optimization of risk measures. In: Calafiore, G., Dabbene, F. (eds.) Probabilistic and Randomized Methods for Design Under Uncertainty, Springer, London (2005)
46. Ruszczyński, A., Shapiro, A.: Optimization of convex risk functions. Math. Oper. Res. **31**, 433–452 (2006)
47. Ruszczyński, A., Shapiro, A.: Conditional risk mappings. Math. Oper. Res. **31**, 544–561 (2006)
48. Scandolo, G.: *Risk measures in a dynamic setting*. PhD Thesis, Università degli Studi di Milano, Milan (2003)
49. Shapiro, A.: On a time consistency concept in risk averse multistage stochastic programming. Oper. Res. Lett. **37**, 143–147 (2009)
50. White, D.J.: Markov Decision Processes. Wiley, New York (1993)