Risk-Averse Control of Undiscounted Transient Markov Models

Özlem Çavuş*
Andrzej Ruszczyński†

March 27, 2012

Abstract

We use Markov risk measures to formulate a risk-averse version of the undiscounted total cost problem for a transient controlled Markov process. We derive risk-averse dynamic programming equations and we show that a randomized policy may be strictly better than deterministic policies, when risk measures are employed. We illustrate the results on an optimal stopping problem and an organ transplant problem.

Keywords: Dynamic Risk Measures; Markov Risk Measures; Stochastic Shortest Path; Optimal Stopping; Randomized Policy

1 Introduction

The optimal control problem for transient Markov processes is a classical model in Operations Research (see Veinott [44], Pliska [31], Bertsekas and Tsitsiklis [6], Hernandez-Lerma and Lasserre [17], and the references therein). The research is focused on the expected total undiscounted cost model, with increased state and control space generality.

Our objective is to consider a risk-averse model. So far, risk-averse problems for transient Markov models were based on the arrival probability criteria (see, e.g., Nie and Wu [23] and Ohtsubo [25]) and utility functions (see Denardo and Rothblum [10] and Patek [29]). We plan to use the recent theory of dynamic risk measures (see Scandolo [40], Ruszczyński and Shapiro [37, 39], Cheridito, Delbaen and Kupper [7], Artzner et. al. [3], Klöppel and Schweizer [20], Pflug and Römisch [30], and the references therein) to develop and solve new risk-averse formulations of the stochastic optimal control problem for transient Markov models. Specific examples of such models are stochastic shortest path problems (Bertsekas and Tsitsiklis [6]) and optimal stopping problems (*cf.* Çinlar [9], Dynkin and Yushkevich [11, 12], Puterman [32]).

Some applications of stochastic shortest path problems concerned with expected performance criteria are given in the survey paper by White [45] and the references therein. However, in many practical problems, the expected values may not be appropriate to measure performance, because they implicitly assume that the decision maker is risk-neutral. Below, we provide examples of such real-life problems which were modeled before as a discrete-time Markov decision process with expected value as the objective function.

Alagoz et. al. [1] suggest a discounted, infinite horizon, and absorbing Markov decision process model to find the optimal time of liver transplant for a risk-neutral patient under the assumption that the liver is transferred from a living donor. However, referring to Chew and Ho [8], they state that the risk-neutrality of the patient is not a realistic assumption. In that study, the patient can be in one of the states "transplant," "death" and intermediate states corresponding to increasing sickness. The decisions are either to wait or to transplant. The "death" and "transplant" states are absorbing states with zero reward. Therefore, the undiscounted version of the model reduces to a stochastic longest path problem.

*Rutgers University, RUTCOR, Piscataway, NJ 08854, USA
†Rutgers University, Department of Management Science and Information Systems, Piscataway, NJ 08854, USA

A stochastic shortest path problem can be used to find the optimal replacement time of a system. Kurt and Kharoufe [21] propose a discounted, infinite horizon Markov decision process model to solve a similar problem for a system under Markovian deterioration and Markovian environment. They assume that the system returns to the "new" state after it is replaced at a given cost. The state space depends on the environment and deterioration levels of the system. The decisions are either to replace the system at a replacement cost or to maintain it at a maintenance cost. Furthermore, we can consider another control "do nothing," to leave the system in operation without any maintenance or replacement at zero cost. They state that their problem can also be equivalently formulated as a stochastic shortest path problem with some probability of making a transition from each state to a zero-cost absorbing state. However, managers are not risk-neutral in real life and this needs to be considered in such replacement problems (see Tapiero and Venezia [43]).

So and Thomas [42] employ a discrete time Markov decision process to model profitability of credit cards. The objective is to find a policy which maximizes the expected total discounted profit of the creditor. The state space depends on the customer's riskiness and the credit limit bands. Additionally, there are absorbing states which represent account closure and different classes of default. The decisions are either to increase the credit limit or keep it unchanged. If zero reward is collected at some of the absorbing states (e.g. closed account), then the undiscounted version of the model reduces to a stochastic longest path problem. However, creditors are assumed to be risk-neutral in these expected-value models, which may not be a realistic assumption.

Our theory of risk-averse control problems for transient models applies to these and many other models. Our results complement and extend the results of Ruszczyński [36], where infinite-horizon *discounted* models were considered.

In section 2 we quickly review some basic concepts of controlled Markov models. In section 3 we adapt and extend our earlier theory of Markov risk measures. In section 4 we introduce and analyze the concept of a multikernel, which is essential for our theory. Section 5 is devoted to the analysis of a finite horizon model. The main model with infinite horizon and dynamic risk measures is analyzed and solved in sections 6 and 7. Section 8 compares randomized and deterministic polices. Finally, section 9 illustrates our results on risk-averse versions of an optimal stopping problem of Karlin [19] and of the organ transplant problem of Alagoz *et al.* [1].

2 Controlled Markov Processes

We quickly review the main concepts of controlled Markov models and we introduce relevant notation (for details, see [13, 16, 17]). Let $\mathscr{X}$ be a state space, and $\mathscr{U}$ a control space. We assume that $\mathscr{X}$ and $\mathscr{U}$ are Polish spaces, equipped with their Borel σ-algebras. A control set is a measurable multifunction $U : \mathscr{X} \rightrightarrows \mathscr{U}$; for each state $x \in \mathscr{X}$ the set $U(x) \subseteq \mathscr{U}$ is a nonempty set of possible controls at x. A controlled transition kernel Q is a measurable mapping from the graph of U to the set $\mathscr{P}(\mathscr{X})$ of probability measures on $\mathscr{X}$ (equipped with the topology of weak convergence).

The cost of transition from x to y, when control u is applied, is represented by the function $c(x,u,y)$, where $c : \mathscr{X} \times \mathscr{U} \times \mathscr{X} \to \mathbb{R}$. Only $u \in U(x)$ and those $y \in \mathscr{X}$ to which transition is possible matter here, but it is convenient to consider the function $c(\cdot,\cdot,\cdot)$ as defined on the product space.

A *stationary controlled Markov process* is defined by a state space $\mathscr{X}$, a control space $\mathscr{U}$, a control set U, a controlled transition kernel Q, and a cost function c.

For $t = 1,2,\ldots$ we define the space of state and control histories up to time t as $\mathscr{H}_t = \operatorname{graph}(U)^t \times \mathscr{X}$. Each history is a sequence $h_t = (x_1,u_1,\ldots,x_{t-1},u_{t-1},x_t) \in \mathscr{H}_t$.

We denote by $\mathscr{P}(\mathscr{U})$ the set of probability measures on the set $\mathscr{U}$. Likewise, $\mathscr{P}(U(x))$ is the set of probability measures on $U(x)$. A *randomized policy* is a sequence of measurable functions $\pi_t : \mathscr{H}_t \to \mathscr{P}(\mathscr{U})$, $t = 1,2,\ldots$, such that $\pi_t(h_t) \in \mathscr{P}(U(x_t))$ for all $h_t \in \mathscr{H}_t$. In words, the distribution of the control u_t is supported on a subset of the set of feasible controls $U(x_t)$. A *Markov policy* is a sequence of measurable functions $\pi_t : \mathscr{X} \to \mathscr{P}(\mathscr{U})$, $t = 1,2,\ldots$, such that $\pi_t(x) \in \mathscr{P}(U(x))$ for all $x \in \mathscr{X}$. The function $\pi_t(\cdot)$ is called the *decision rule* at time t. A Markov policy is *stationary* if there exists a function $\pi : \mathscr{X} \to \mathscr{P}(\mathscr{U})$

such that $\pi_t(x) = \pi(x)$, for all $t = 1, 2, \ldots$ and all $x \in \mathscr{X}$. Such a policy and the corresponding decision rule are called *deterministic*, if for every $x \in \mathscr{X}$ there exists $u(x) \in U(x)$ such that the measure $\pi(x)$ is supported on $\{u(x)\}$.

For a stationary decision rule π, we write Q^π to denote the corresponding transition kernel.

We focus on *transient* Markov models. We assume that there exists some *absorbing state* $x_{\mathrm{A}} \in \mathscr{X}$, such that $Q\big(\{x_{\mathrm{A}}\}\big|x_{\mathrm{A}}, u\big) = 1$ and $c(x_{\mathrm{A}}, u, x_{\mathrm{A}}) = 0$ for all $u \in U(x_{\mathrm{A}})$. Thus, after the absorbing state is reached, no further costs are incurred.[1] To analyze such Markov models, it is convenient to consider the effective state space $\widetilde{\mathscr{X}} = \mathscr{X} \setminus \{x_{\mathrm{A}}\}$, and the effective controlled substochastic kernel $\widetilde{Q}$ whose arguments are restricted to $\widetilde{\mathscr{X}}$ and whose values are nonnegative measures on $\widetilde{\mathscr{X}}$, so that $\widetilde{Q}\big(B\big|x,u\big) = Q\big(B\big|x,u\big)$, for all Borel sets $B \subset \widetilde{\mathscr{X}}$, all $x \in \widetilde{\mathscr{X}}$, and all $u \in U(x)$. Moreover, we assume that the following Pliska condition [31] is satisfied: a weight function $w : \mathscr{X} \to [1, \infty)$ and a constant K exist, such that for every Markov decision rule π we have

$$\sum_{j=1}^{\infty} \left\| \left(\widetilde{Q}^{\pi}\right)^j \right\|_w \leq K. \tag{1}$$

In the condition above, the norm $\|A\|_w$ of a substochastic kernel A is defined as follows:

$$\|A\|_w = \sup_{x \in \widetilde{\mathscr{X}}} \frac{1}{w(x)} \int_{\widetilde{\mathscr{X}}} w(y)\, A(dy|x). \tag{2}$$

It is the standard operator norm in the space $\mathbb{B}_w(\widetilde{\mathscr{X}}, \mathscr{B}(\widetilde{\mathscr{X}}))$ of measurable functions $v : \widetilde{\mathscr{X}} \to \mathbb{R}$ for which

$$\|v\|_w = \sup_{x \in \widetilde{\mathscr{X}}} \frac{v(x)}{w(x)} < \infty.$$

Hernadez-Lerma and Lasserre [17] extensively discuss the role of weighted norms in dynamic programming models.

Our point of departure is the *expected total cost problem*, which is to find a policy $\Pi = \{\pi_t\}_{t=1}^{\infty}$ so as to minimize the expected cost until absorption:

$$\min_{\Pi} \mathbb{E}\left[\sum_{t=1}^{\infty} c(x_t, u_t, x_{t+1})\right]. \tag{3}$$

Under standard assumptions, the problem has a solution in form of a stationary Markov policy. Moreover, it is sufficient to restrict the considerations to deterministic policies. The optimal policy can be found by solving appropriate dynamic programming equations.

Our intention is to introduce risk aversion to problem (3), and to replace the expected value operator by a dynamic risk measure. We shall show that the Pliska condition (1) is not sufficient in this case, and that properties of risk measures must be taken into account when considering transient models. We shall also show that in the risk-averse case randomized policies can be optimal, and that it is essential to consider general transition cost $c(x_t, u_t, x_{t+1})$, which in problem (3) could easily be reduced to functions depending only on (x_t, u_t). We do not assume that the costs are nonnegative, and thus our approach applies also, among others, to stochastic longest path problems and optimal stopping problems with positive rewards.

3 Markov Risk Measures

Suppose T is a fixed time horizon. Each policy $\Pi = \{\pi_1, \pi_2, \ldots\}$ results in a cost sequence $Z_t = c(x_{t-1}, u_{t-1}, x_t)$, $t = 2, \ldots, T+1$. We define the spaces $\mathscr{Z}_t$ of $\mathscr{F}_t$-measurable random variables on Ω, $t = 1, \ldots, T$. In this paper, we focus on the case when $\mathscr{Z}_t = \mathscr{L}_p(\Omega, \mathscr{F}_t, P)$, for some $p \in [1, \infty]$.

[1] The case of a larger class of absorbing states easily reduces to the case of one absorbing state.

To evaluate risk of this sequence we use a dynamic time-consistent risk measure of the following form:

$$J_T(\Pi, x_1) = \rho_1\Big(c(x_1,u_1,x_2) + \rho_2\Big(c(x_2,u_2,x_3) + \cdots \\ + \rho_{T-1}\big(c(x_{T-1},u_{T-1},x_T) + \rho_T(c(x_T,u_T,x_{T+1}))\big)\cdots\Big)\Big). \tag{4}$$

Here, $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$, $t = 1,\dots,T$, are one-step conditional risk measures satisfying the following conditions:

A1. $\rho_t(\alpha Z + (1-\alpha)W) \le \alpha\rho_t(Z) + (1-\alpha)\rho_t(W)$, $\forall\, \alpha \in (0,1)$, $Z, W \in \mathscr{Z}_{t+1}$;

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. $\rho_t(\beta Z) = \beta\rho_t(Z)$, $\forall\, Z \in \mathscr{Z}_{t+1}$, $\beta \ge 0$.

Ruszczyński [36, sec. 3] derives the nested formulation (4) and conditions (A2) and (A3) from general properties of monotonicity and time-consistency of dynamic measures of risk. Conditions (A1) and (A4) are added to model the diversification effect and scale-invariance of the preferences, similarly to the axioms of coherent measures of risk (see (B1)–(B4) below).

It is convenient to introduce vector spaces $\mathscr{Z}_{t,\theta} = \mathscr{Z}_t \times \mathscr{Z}_{t+1} \times \cdots \times \mathscr{Z}_\theta$, where $1 \le t \le \theta \le T+1$ and the conditional risk measures $\rho_{t,\theta} : \mathscr{Z}_{t,\theta} \to \mathscr{Z}_t$ defined as follows:

$$\rho_{t,\theta}(Z_t,\dots,Z_\theta) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{\theta-1}(Z_\theta)\cdots\big)\Big). \tag{5}$$

The operations of addition and multiplication by a scalar are defined in $\mathscr{Z}_{t,\theta}$ in the usual way. We can also define the partial order relation $\preceq$ in a natural way:

$$(Z_t,\dots,Z_\theta) \preceq (W_t,\dots,W_\theta) \iff Z_\tau \le W_\tau, \text{ a.s., } \tau = t,\dots,\theta.$$

Immediately from the definition we obtain the following properties of conditional measures of risk.

Lemma 3.1. *If the one-step conditional risk measures ρ_τ, $\tau = t,\dots,\theta-1$, satisfy conditions A1–A4, then*

(i) $\rho_{t,\theta}(\alpha Z + (1-\alpha)W) \le \alpha\rho_{t,\theta}(Z) + (1-\alpha)\rho_{t,\theta}(W)$, $\forall\, \alpha \in (0,1)$, $Z, W \in \mathscr{Z}_{t,\theta}$;

(ii) *If $Z \preceq W$ then $\rho_{t,\theta}(Z) \le \rho_{t,\theta}(W)$, $\forall\, Z, W \in \mathscr{Z}_{t,\theta}$;*

(iii) $\rho_{t,\theta}(\beta Z) = \beta\rho_{t,\theta}(Z)$, $\forall\, Z \in \mathscr{Z}_{t+1}$, $\beta \ge 0$;

(iv) $\rho_{t,\theta}(Z_t,\dots,Z_{\theta-1},0) = \rho_{t,\theta-1}(Z_t,\dots,Z_{\theta-1})$.

As indicated in [36], the fundamental difficulty of formulation (4) 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 h_t of the process. In order to overcome this difficulty, in [36, sec. 4] a new construction of a one-step conditional measure of risk was introduced. Its arguments are 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 decisions u. We adapt this construction to the case of controlled Markov models with randomized policies. In this case, it is convenient to consider functions on the product space $\mathscr{U} \times \mathscr{X}$ equipped with its product Borel σ-algebra $\mathscr{B}$.

Suppose the current state is x and we use a randomized control λ. This control, together with the transition kernel Q defines a probability measure $\lambda \circ Q_x$ on the product space $\mathscr{U} \times \mathscr{X}$ as follows:

$$[\lambda \circ Q_x](B_u \times B_y) = \int_{B_u} Q(B_y|x,u)\,\lambda(du), \quad B_u \in \mathscr{B}(U), \quad B_y \in \mathscr{B}(\mathscr{X}). \tag{6}$$

The measure is extended to other sets in $\mathscr{B}$ in a usual way. In the case of countable state and control spaces, $\big[\lambda \circ Q_x\big](u,y)$ is the probability that control u will be used at x and the next state will be y.

The cost incurred at the current stage is given by the function c_x on the product space $\mathscr{U} \times \mathscr{X}$ defined as follows:

$$c_x(u,y) = c(x,u,y), \quad u \in \mathscr{U}, \quad y \in \mathscr{X}. \tag{7}$$

Let $\mathscr{V} = \mathscr{L}_p(\mathscr{U} \times \mathscr{X}, \mathscr{B}, P_0)$, where $p \in [1,\infty]$ and P_0 is some reference probability measure on $\mathscr{U} \times \mathscr{X}$. It is convenient to think of the dual space $\mathscr{V}'$ as the space of signed measures m on $(\mathscr{U} \times \mathscr{X}, \mathscr{B})$, which are absolutely continuous with respect to P_0, with densities (Radon–Nikodym derivatives) lying in the space $\mathscr{L}_q(\mathscr{U} \times \mathscr{X}, \mathscr{B}, P_0)$, where $1/p + 1/q = 1$. In the case of finite state and control spaces P_0 may be the uniform measure; in other cases P_0 should be chosen in such a way that the measures $\lambda \circ Q_x$ are elements of $\mathscr{V}'$. The measure P_0 does not play any other role in our considerations. We consider the set of probability measures in $\mathscr{V}'$:

$$\mathscr{M} = \big\{ m \in \mathscr{V}' : m(\mathscr{U} \times \mathscr{X}) = 1,\ m \geq 0 \big\}.$$

We also assume that the spaces $\mathscr{V}$ and $\mathscr{V}'$ are endowed with topologies that make them paired topological vector spaces with the bilinear form

$$\langle \varphi, m \rangle = \int_{\mathscr{U} \times \mathscr{X}} \varphi(u,y)\, m(du \times dy), \quad \varphi \in \mathscr{V}, \quad m \in \mathscr{V}'. \tag{8}$$

The space $\mathscr{V}'$ (and thus $\mathscr{M}$) will be 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 (8), that is, the weak* topology on $\mathscr{L}_\infty(\mathscr{X}, \mathscr{B}, P_0)$.

Definition 3.1. *A measurable function $\sigma : \mathscr{V} \times \mathscr{X} \times \mathscr{M} \to \mathbb{R}$ is a* risk transition mapping *if for every $x \in \mathscr{X}$ and every $m \in \mathscr{M}$, the function $\varphi \mapsto \sigma(\varphi, x, m)$ is a coherent measure of risk on $\mathscr{V}$.*

Recall that $\sigma(\cdot)$ is a coherent measure of risk on $\mathscr{V}$ (we skip the other two arguments for brevity), if

B1. $\sigma(\alpha\varphi + (1-\alpha)\psi) \leq \alpha\sigma(\varphi) + (1-\alpha)\sigma_t(\psi)$, $\forall\, \alpha \in (0,1)$, $\varphi, \psi \in \mathscr{V}$;

B2. If $\varphi \leq \psi$ then $\sigma(\varphi) \leq \sigma(\psi)$, $\forall\, \varphi, \psi \in \mathscr{V}$;

B3. $\sigma(a + \varphi) = a + \sigma(\varphi)$, $\forall\, \varphi \in \mathscr{V}$, $a \in \mathbb{R}$;

B4. $\sigma(\beta\varphi) = \beta\sigma(\varphi)$, $\forall\, \varphi \in \mathscr{V}$, $\beta \geq 0$.

Example 3.1. Consider the first-order mean–semideviation risk measure analyzed by Ogryczak and Ruszczyński [26, 27], and Ruszczyński and Shapiro [38, Example 4.2], [39, Example 6.1]), but with the state and the underlying probability measure as its arguments. We define

$$\sigma(\varphi, x, m) = \langle \varphi, m \rangle + \kappa(x) \big\langle (\varphi - \langle \varphi, m \rangle)_+, m \big\rangle, \tag{9}$$

with some measurable function $\kappa : \mathscr{X} \to [0,1]$. We can verify directly that conditions (B1)–(B4) are satisfied.

Example 3.2. Another important example is the Conditional Average Value at Risk (see, *inter alia*, Ogryczak and Ruszczyński [28, Sec. 4], Pflug and Römisch [30, Sec. 2.2.3, 3.3.4], Rockafellar and Uryasev [34], Ruszczyński and Shapiro [38, Example 4.3], [39, Example 6.2]), which has the following risk transition counterpart:

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

Here $\alpha : \mathscr{X} \to [\alpha_{\min}, \alpha_{\max}] \subset (0,1)$ is measurable. Again, the conditions (B1)–(B4) can be verified directly.

We shall use the property of *law invariance* of a risk transition mapping. For a function $\varphi \in \mathscr{V}$ and a probability measure $\mu \in \mathscr{M}$ we can define the distribution function $F_\varphi^\mu : \mathbb{R} \to [0,1]$ as follows

$$F_\varphi^\mu(\eta) = \mu\big\{(u,y) \in \mathscr{U} \times \mathscr{X} : \varphi(u,y) \le \eta\big\}.$$

Definition 3.2. *A risk transition mapping $\sigma : \mathscr{V} \times \mathscr{X} \times \mathscr{M} \to \mathbb{R}$ is* law invariant*, if for all $\varphi, \psi \in \mathscr{V}$ and all $\mu, \nu \in \mathscr{M}$ such that $F_\varphi^\mu \equiv F_\psi^\nu$, we have $\sigma(\varphi, x, \mu) = \sigma(\psi, x, \nu)$ for all $x \in \mathscr{X}$.*

The concept of law invariance corresponds to a similar concept for coherent measures of risk, but here we additionally need to take into account the variability of the probability measure. The risk transition mappings of Examples 3.1 and 3.2 are law invariant. While we shall not directly use law invariance in our main theoretical considerations, it greatly simplifies the analysis of specific problems, as illustrated in section 9.1.

Risk transition mappings allow for convenient formulation of risk-averse preferences for controlled Markov processes, where the cost is evaluated by formula (4). Consider a controlled Markov process $\{x_t\}$ with some Markov policy $\Pi = \{\pi_1, \pi_2, \dots\}$. For a fixed time t and a measurable function $g : \mathscr{X} \times \mathscr{U} \times \mathscr{X} \to \mathbb{R}$ the value of $Z_{t+1} = g(x_t, u_t, x_{t+1})$ is a random variable. We assume that g is *w-bounded*, that is,

$$\big|g(x,u,y)\big| \le C\big(w(x) + w(y)\big), \quad \forall\, x \in \mathscr{X},\ u \in U(x),\ y \in \mathscr{X},$$

for some constant $C > 0$ and for the weight function $w : \mathscr{X} \to [1, \infty)$, $w \in \mathscr{V}$. Then Z_{t+1} is an element of $\mathscr{Z}_{t+1}$. Let $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$ be a conditional risk measure satisfying (A1)–(A4). By definition, $\rho_t\big(g(x_t, u_t, x_{t+1})\big)$ is an element of $\mathscr{Z}_t$, that is, it is an $\mathscr{F}_t$-measurable function on $(\Omega, \mathscr{F})$. In the definition below, we restrict it to depend on the past only via the current state x_t. We write $g_x : \mathscr{U} \times \mathscr{X} \to \mathbb{R}$ for the function $g_x(u,y) = g(x,u,y)$, π_x for the measure $\pi(\cdot|x)$, and Q_x for the mapping $u \to Q(\cdot|x,u)$.

Definition 3.3. *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 w-bounded measurable functions $g : \mathscr{X} \times \mathscr{U} \times \mathscr{X} \to \mathbb{R}$ and for all feasible decision rules $\pi : \mathscr{X} \to \mathscr{P}(U)$ we have*

$$\rho_t\big(g(x_t, u_t, x_{t+1})\big) = \sigma_t\big(g_{x_t}, x_t, \pi_{x_t} \circ Q_{x_t}\big), \quad a.s. \tag{10}$$

Observe that the right hand side of formula (10) is parametrized by x_t, and thus it defines a special $\mathscr{F}_t$-measurable function of ω, whose dependence on the past is carried only via the state x_t.

Remark 3.1. If $c(x_t, u_t, x_{t+1}) \equiv d(x_t, x_{t+1})$, or if randomized policies are not allowed, then it is sufficient to start from a probability measure P_0 on $\mathscr{X}$ and define $\mathscr{V} = \mathscr{L}_p(\mathscr{X}, \mathscr{B}(\mathscr{X}), P_0)$, $\mathscr{V}'$ - the set of measures on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$ having densities with respect to P_0 in $\mathscr{L}_q(\mathscr{X}, \mathscr{B}(\mathscr{X}), P_0)$, and $\mathscr{M} = \{m \in \mathscr{V}' : m(\mathscr{X}) = 1,\ m \ge 0\}$, exactly as in [36].

Remark 3.2. If, additionally, the stage-wise costs have the form $c(x_t, u_t, x_{t+1}, \xi_t)$, where ξ_t, $t = 1, 2, \dots$, are some random variables distributed a Polish space Ξ according to a measure which is absolutely continuous with respect to some fixed P_ξ, but may depend on x_t and u_t, then we need to consider larger spaces of arguments of a risk transition mapping:

$$\mathscr{V} = \mathscr{L}_p(\mathscr{U} \times \mathscr{X} \times \Xi, \mathscr{B}(\mathscr{U} \times \mathscr{X} \times \Xi), P_0 \times P_\xi),$$
$$\mathscr{V}' = \mathscr{L}_q(\mathscr{U} \times \mathscr{X} \times \Xi, \mathscr{B}(\mathscr{U} \times \mathscr{X} \times \Xi), P_0 \times P_\xi),$$
$$\mathscr{M} = \left\{ m \in \mathscr{V}' : \int\limits_{\mathscr{U} \times \mathscr{X} \times \Xi} m(u,x,\xi)\, P_0(du\,dx\,d\xi) = 1,\ m \ge 0 \right\}.$$

All our considerations remain valid, just the notation complicates.

4 Stochastic Multikernels

In order to analyze Markov measures of risk, we need to introduce the concept of a multikernel.

Definition 4.1. *A* multikernel *is a measurable multifunction $\mathfrak{M}$ from $\mathscr{X}$ to the space* $\mathrm{rca}(\mathscr{X},\mathscr{B}(\mathscr{X}))$ *of regular measures on* $(\mathscr{X},\mathscr{B}(X))$. *It is* stochastic*, if its values are sets of probability measures. It is* substochastic*, if* $0 \le M(B|x) \le 1$ *for all* $M \in \mathfrak{M}(x)$, $B \in \mathscr{B}(\mathscr{X})$, *and* $x \in \mathscr{X}$. *It is* convex *(*closed*), if for all* $x \in \mathscr{X}$ *its value* $\mathfrak{M}(x)$ *is a convex (*closed*) set.*

The concept of a multikernel is thus a multivalued generalization of the concept of a kernel. A measurable selector of a stochastic multikernel $\mathfrak{M}$ is a stochastic kernel M such that $M(x) \in \mathfrak{M}(x)$ for all $x \in \mathscr{X}$. We symbolically write $M \lessdot \mathfrak{M}$ to indicate that M is a measurable selector of $\mathfrak{M}$.

Recall that a composition $M_1 \circ M_2$ of (sub-) stochastic kernels M_1 and M_2 is given by the formula:

$$\big[M_1 \circ M_2\big]\big(B\big|x\big) = \int_{\mathscr{X}} M_2(B|y)\, M_1(dy|x), \quad \mathscr{B} \in \mathscr{B}(\mathscr{X}), \quad x \in \mathscr{X}. \tag{11}$$

It is also a (sub-) stochastic kernel. Multikernels, in particular substochastic multikernels, can be composed in a similar fashion.

Definition 4.2. *If* $\mathfrak{M}_i : \mathscr{X} \rightrightarrows \mathrm{rca}(\mathscr{X},\mathscr{B}(\mathscr{X}))$, $i = 1,2$ *are substochastic multikernels, then their* composition $\mathfrak{M}_1 \circ \mathfrak{M}_2$ *is defined as follows:*

$$\big[\mathfrak{M}_1 \circ \mathfrak{M}_2\big]\big(B\big|x\big) = \Big\{ \big[M_1 \circ M_2\big]\big(B\big|x\big) :\ M_i \lessdot \mathfrak{M}_i,\ i = 1,2 \Big\}.$$

It follows from Definition 4.2, that a composition of (sub-) stochastic multikernels is a (sub-) stochastic multikernel. We may compose a substochastic multikernel $\mathfrak{M}$ with itself several times, to obtain its "power":

$$(\mathfrak{M})^k = \underbrace{\mathfrak{M} \circ \mathfrak{M} \ldots \circ \mathfrak{M}}_{k \text{ times}}.$$

The *norm* of a substochastic multikernel $\mathfrak{M} : \widetilde{\mathscr{X}} \rightrightarrows \mathrm{rca}(\widetilde{\mathscr{X}},\mathscr{B}(\widetilde{\mathscr{X}}))$ is defined as follows:

$$\|\mathfrak{M}\|_w = \sup_{M \lessdot \mathfrak{M}} \|M\|_w,$$

where the norm $\|M\|_w$ is given by (2).

The concept of a multikernel and the composition operation arise in a natural way in the context of Markov risk measures. If $\sigma(\cdot,\cdot,\cdot)$ is a Markov risk measure, then the function $\sigma(\cdot,x,m)$ is lower semicontinuous for all $x \in \mathscr{X}$ and $m \in \mathscr{M}$ (see Ruszczyński and Shapiro [38, Proposition 3.1]). Then it follows from [38, Theorem 2.2] that for every $x \in \mathscr{X}$ and $m \in \mathscr{M}$ a closed convex set $\mathscr{A}(x,m) \subset \mathscr{M}$ exists, such that for all $\varphi \in \mathscr{V}$ we have

$$\sigma(\varphi,x,m) = \max_{\mu \in \mathscr{A}(x,m)} \langle \varphi, \mu \rangle. \tag{12}$$

In fact, we also have

$$\mathscr{A}(x,m) = \partial_\varphi \sigma(0,x,m). \tag{13}$$

In many cases, the multifunction $\mathscr{A} : \mathscr{X} \times \mathscr{M} \rightrightarrows \mathscr{M}$ can be described analytically.

Example 4.1. For the mean-semideviation model of Example 3.1, following the derivations of Ruszczyński and Shapiro [38, Example 4.2], we have

$$\mathscr{A}(x,m) = \Big\{ \mu \in \mathscr{M} : \exists \big(h \in \mathscr{L}_\infty(\mathscr{U} \times \mathscr{X}, \mathscr{B}, P_0)\big)\ \frac{d\mu}{dm} = 1 + h - \langle h, m \rangle,\ \|h\|_\infty \le \kappa(x),\ h \ge 0 \Big\}. \tag{14}$$

Similar formulas can be derived for higher order measures.

Example 4.2. For the Conditional Average Value at Risk of Example 3.2, following the derivations of Ruszczyński and Shapiro [38, Example 4.3], we obtain

$$\mathscr{A}(x,m) = \left\{ \mu \in \mathscr{M} : \frac{d\mu}{dm} \le \frac{1}{\alpha(x)} \right\}. \tag{15}$$

Consider the formula (10) and suppose that $g(x_t,u_t,x_{t+1}) = v(x_{t+1})$ for some measurable w-bounded function $v : \mathscr{X} \to \mathbb{R}$. Using the representation (12) we can write it as follows:

$$\rho_t\big(v(x_{t+1})\big) = \max_{\mu \in \mathscr{A}(x_t, \pi_{x_t} \circ Q_{x_t})} \langle v, \mu \rangle, \quad \text{a.s.} \tag{16}$$

In the formula above, the last bilinear form is an integral over $\mathscr{U} \times \mathscr{X}$. The function $v(\cdot)$ depends on x only, and thus it is sufficient to consider the marginal measures

$$\bar{\mu}(B) = \mu(\mathscr{U} \times B), \quad B \in \mathscr{B}(\mathscr{X}). \tag{17}$$

Denote by L the linear operator mapping each $\mu \in \mathscr{V}'$ to the corresponding marginal measure $\bar{\mu}$ on $(\mathscr{X}, \mathscr{B}(\mathscr{X}))$, as defined in (17). For every x we can define the set of probability measures:

$$\mathfrak{M}_x^\pi = \big\{ L\mu : \mu \in \mathscr{A}(x, \pi_x \circ Q_x) \big\}, \quad x \in \mathscr{X}. \tag{18}$$

The multifunction $\mathfrak{M}^\pi : \mathscr{X} \rightrightarrows \mathscr{P}(\mathscr{X})$, assigning to each $x \in \mathscr{X}$ the set $\mathfrak{M}_x^\pi$, is a closed convex stochastic multikernel. We call it a *risk multikernel*, associated with the risk transition mapping $\sigma(\cdot,\cdot,\cdot)$, the controlled kernel Q, and the policy π. Its measurable selectors $M^\pi \lessdot \mathfrak{M}^\pi$ are transition kernels.

It follows that formula (16) can be rewritten as follows:

$$\rho_t\big(v(x_{t+1})\big) = \max_{M \in \mathfrak{M}_{x_t}^\pi} \int_{\mathscr{X}} v(y)\, M(dy). \tag{19}$$

In the risk-neutral case we have

$$\rho_t\big(v(x_{t+1})\big) = \mathbb{E}\big[v(x_{t+1})\big|x_t\big] = \int_{\mathscr{U}} \int_{\mathscr{X}} v(y)\, Q(dy|x_t,u)\, \pi(du|x_t) = \int_{\mathscr{X}} v(y)\, Q_{x_t}^\pi(dy),$$

with the transition kernel Q^π associated with the policy π given by $Q_x^\pi = L[\pi_x \circ Q_x]$. The comparison of the last two displayed equations reveals that in the risk-neutral case we have

$$\mathfrak{M}_x^\pi = \big\{ Q_x^\pi \big\}, \quad x \in \mathscr{X}, \tag{20}$$

that is, the risk multikernel $\mathfrak{M}^\pi$ is single-valued, and its only selector is the kernel Q^π. In the risk-averse case, the risk multikernel $\mathfrak{M}^\pi$ is a closed convex-valued multifunction, whose measurable selectors are transition kernels. It is evident that properties of this multifunction are germane for our analysis. We return to this issue in section 6, where we calculate some examples of transition multikernels.

Remark 4.1. *If $m \in \mathscr{A}(x,m)$ for all $x \in \mathscr{X}$ and $m \in \mathscr{M}$, then it follows from equation* (18) *that Q^π is a measurable selector of $\mathfrak{M}^\pi$. Moreover, it follows from* (12) *that for any function $\varphi \in \mathscr{V}$ we have*

$$\rho_t\big(\varphi(u_t,x_{t+1})\big) \ge \int_{\mathscr{U} \times \mathscr{X}} \varphi(u,y)\, \big[Q_{x_t} \circ \pi_{x_t}\big]\, (du \times dy) = \mathbb{E}\big[\varphi(u_t,x_{t+1})\big|x_t\big].$$

It follows that the dynamic risk measure (4) *is bounded from below by the expected value of the total cost.*

The condition $m \in \mathscr{A}(x,m)$ is satisfied by the measures of risk in Examples 4.1 and 4.2.

Interestingly, uncertain transition matrices were used by Nilim and El Ghaoui in [24] to increase robustness of control rules for Markov models. In our theory, controlled multikernels (generalization of such matrices), arise in a natural way in the analysis of risk-averse preferences.

Let us quickly recall continuity properties of the multifunctions involved in the construction of a Markov risk measure.

Proposition 4.1. *Suppose $\varphi \in \mathscr{V}$ and $x \in \mathscr{X}$. If the controlled kernel $u \mapsto Q(\cdot|x,u)$ is continuous, and the multifunction $m \mapsto \mathscr{A}(\varphi,x,m)$ is lower semicontinuous, then the function $\lambda \mapsto \sigma(\varphi,x,\lambda \circ Q_x)$ is weakly* lower semicontinuous on $\mathscr{P}\big(U(x)\big)$.*

Proof. For a continuous Q, the multifunction $\lambda \mapsto \mathscr{A}(x,\lambda \circ Q_x)$ inherits the continuity properties of $\mathscr{A}$. The function $\mu \mapsto \langle \varphi, \mu \rangle$ is continuous on $\mathscr{M}$ (in the weak* topology). The assertion of the theorem follows now from the dual representation (12) by [4, Theorem 1.4.16], whose proof remains valid in our setting as well. □

Some comments on the assumptions of Proposition 4.1 are in order. The continuity of the kernel Q is a standard condition in the theory of risk-neutral Markov control processes (see, e.g., [16]). If the risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ is continuous, then its subdifferential (13) is upper semicontinuous. However, in Proposition 4.1 we assume *lower* semicontinuity of the mapping $m \mapsto \partial_\varphi \sigma(0,x,m)$, which is not trivial and should be verified for each case. For example, the subdifferentials derived in Examples 4.1 and 4.2 are continuous with respect to m.

5 Finite Horizon Problem

We consider the Markov model at times $1,2,\ldots,T+1$ under general policies $\Pi = \{\pi_1,\pi_2,\ldots,\pi_T\}$. The cost at the last stage is given by a function $v_{T+1}(x_{T+1})$. Consider the problem

$$\min_{\Pi} J_T(\Pi,x_1), \tag{21}$$

with $J_T(\Pi,x_1)$ defined by formula (4), with Markov conditional risk measures ρ_t, $t=1,\ldots,T$, with risk transition mappings $\sigma_t(\cdot,\cdot,\cdot)$:

$$J_T(\Pi,x_1) = \rho_1\Big(c(x_1,u_1,x_2) + \rho_2\Big(c(x_2,u_2,x_3) + \cdots + \rho_T\big(c(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big). \tag{22}$$

Theorem 5.1. *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 and such that for every $x \in \mathscr{X}$ the multifunction $\mathscr{A}_t(x,\cdot)$ is lower semicontinuous;*

(iii) *The function $c(\cdot,\cdot,\cdot)$ is w-bounded, measurable, and lower semicontinuous with respect to the second argument;*

(iv) *For every $x \in \mathscr{X}$ the set $U(x)$ is compact;*

(v) *The function $v_{T+1}(\cdot)$ is w-bounded and measurable.*

Then problem (21) *has an optimal solution and its optimal value $v_1(x)$ is the solution of the following dynamic programming equations:*

$$v_t(x) = \min_{\lambda \in \mathscr{P}(U(x))} \sigma_t\big(c_x + v_{t+1},x,\lambda \circ Q_x\big), \quad x \in \mathscr{X}, \quad t = T,\ldots,1. \tag{23}$$

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

$$\hat{\pi}_t(x) \in \operatorname*{argmin}_{\lambda \in \mathscr{P}(U(x))} \sigma_t\big(c_x + v_{t+1},x,\lambda \circ Q_x\big), \quad x \in \mathscr{X}, \quad t = T,\ldots,1. \tag{24}$$

Conversely, every solution of equations (23)–(24) *defines an optimal Markov policy $\hat{\Pi}$.*

Proof. Our proof is similar to the proof of Ruszczyński [36, Thm. 2], but with adjustments due to the use of randomized strategies. We provide its short outline.

Owing to the monotonicity condition (B2) applied to ρ_t, $t = 1,\dots,T$, problem (21) can be written as follows:

$$\min_{\pi_1,\dots,\pi_T} \Big\{ \rho_1\Big(c(x_1,u_1,x_2) + \cdots + \rho_T\big(c(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big\} =$$
$$\min_{\pi_1,\dots,\pi_{T-1}} \Big\{ \rho_1\Big(c(x_1,u_1,x_2) + \cdots + \min_{\pi_T}\rho_T\big(c(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big\}.$$

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

$$\min_{\lambda\in\mathscr{P}(U(x_T))} \sigma_T\big(c_{x_T} + v_{T+1}, x_T, \lambda\circ Q_{x_T}\big). \tag{25}$$

The problem becomes equivalent to (23) for $t = T$, and its solution is given by (24) for $t = T$. By Proposition 4.1, the function $\lambda \mapsto \sigma_T\big(c_{x_T} + v_{T+1}, x_T, \lambda\circ Q_{x_T}\big)$ is lower semicontinuous. As the set of $\lambda\in\mathscr{P}(\mathscr{U})$ such that $\lambda(U(x_T)) = 1$ is weakly* compact, the optimal randomized policy $\pi_T(x)$, which is the minimizer in (25), exists.

After that, the horizon $T+1$ is decreased to T, and the final cost becomes $v_T(x_T)$. Proceeding in this way for $T, T-1,\dots,1$ we obtain the assertion of the theorem. □

It follows from our proof that the functions $v_t(\cdot)$ calculated in (23) are the optimal values of tail subproblems formulated for a fixed $x_t = x$ as follows:

$$v_t(x) = \min_{\pi_t,\dots,\pi_T} \rho_t\Big(c(x_t,u_t,x_{t+1}) + \rho_{t+1}\big(c(x_{t+1},u_{t+1},x_{t+2}) + \cdots + \rho_T\big(c(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\big)\Big).$$

We call them *value functions*, as in risk-neutral dynamic programming. It is clear that we may also have non-stationary costs and transition kernels in this case. Also, the assumption that the process is transient is not needed.

Equations (23)–(24) provide a computational recipe for solving finite horizon problems.

6 Evaluation of Stationary Markov Policies in Infinite Horizon Problems

Consider a policy $\Pi = \{\pi_1,\pi_2,\dots\}$ and define the cost until absorption as follows:

$$J_\infty(\Pi,x_1) = \lim_{T\to\infty} J_T(\Pi,x_1), \tag{26}$$

where each $J_T(\Pi,x_1)$ is defined by the formula

$$\begin{aligned} J_T(\Pi,x_1) &= \rho_1\Big(c(x_1,u_1,x_2) + \rho_2\Big(c(x_2,u_2,x_3) + \cdots + \rho_T\big(c(x_T,u_T,x_{T+1})\big)\cdots\Big)\Big) \\ &= \rho_{1,T+1}\big(0, c(x_1,u_1,x_2), c(x_2,u_2,x_3),\dots,c(x_T,u_T,x_{T+1})\big), \end{aligned} \tag{27}$$

with Markov conditional risk measures ρ_t, $t = 1,\dots,T$, sharing the same risk transition mapping $\sigma(\cdot,\cdot,\cdot)$. We assume all conditions of Theorem 5.1. We still have to index each conditional risk measure by time, because by definition it acts from the space $\mathscr{Z}_{t+1}$ to the space $\mathscr{Z}_t$.

The first question to answer is when this cost is finite. This question is nontrivial, because even for uniformly bounded costs $Z_t = c(x_{t-1},u_{t-1},x_t)$, $t = 2,3,\dots$, and for a transient finite-state Markov chain, the limit in (26) may be infinite, as the following example demonstrates.

Example 6.1. Consider a transient Markov chain with two states and with the following transition probabilities: $Q_{11} = Q_{12} = \frac{1}{2}$, $Q_{22} = 1$. Only one control is possible in each state, the cost of each transition from state 1 is equal to 1, and the cost of the transition from 2 to 2 is 0. Clearly, the time until absorption is a geometric random variable with parameter $\frac{1}{2}$. Let $x_1 = 1$. If the limit (26) is finite, then (skipping the dependence on Π) we have

$$J_\infty(1) = \lim_{T\to\infty} J_T(1) = \lim_{T\to\infty} \rho_1\big(1 + J_{T-1}(x_2)\big) = \rho_1\big(1 + J_\infty(x_2)\big).$$

In the last equation we used the continuity of $\rho_1(\cdot)$. Clearly, $J_\infty(2) = 0$.

Suppose that we are using the Average Value at Risk from Example 3.2, with $0 < \alpha \le \frac{1}{2}$, to define $\rho_1(\cdot)$. Using standard identities for the Average Value at Risk (see, e.g., [41, Thm. 6.2]), we obtain

$$\begin{aligned} J_\infty(1) &= \inf_{\eta\in\mathbb{R}} \Big\{ \eta + \frac{1}{\alpha}\mathbb{E}\big[\big(1 + J_\infty(x_2) - \eta\big)_+\big] \Big\} \\ &= 1 + \inf_{\eta\in\mathbb{R}} \Big\{ \eta + \frac{1}{\alpha}\mathbb{E}\big[\big(J_\infty(x_2) - \eta\big)_+\big] \Big\} = 1 + \frac{1}{\alpha}\int_{1-\alpha}^{1} F^{-1}(\beta)\, d\beta, \end{aligned} \tag{28}$$

where $F(\cdot)$ is the distribution function of $J_\infty(x_2)$. As all β-quantiles of $J_\infty(x_2)$ for $\beta \ge \frac{1}{2}$ are equal to $J_\infty(1)$, the last equation yields

$$J_\infty(1) = 1 + J_\infty(1),$$

a contradiction. It follows that a composition of average values at risk has no finite limit, if $0 < \alpha \le \frac{1}{2}$.

On the other hand, if $\frac{1}{2} < \alpha < 1$, then

$$F^{-1}(\beta) = \begin{cases} J_\infty(2) = 0 & \text{if } 1-\alpha \le \beta < \frac{1}{2}, \\ J_\infty(1) & \text{if } \frac{1}{2} \le \beta \le 1. \end{cases}$$

Formula (28) then yields

$$J_\infty(1) = 1 + \frac{1}{2\alpha} J_\infty(1).$$

This equation has a solution $J_\infty(1) = 2\alpha/(2\alpha - 1)$.

If we use the mean-semideviation model of Example 3.1, we obtain

$$\begin{aligned} J_\infty(1) &= \mathbb{E}\big[1 + J_\infty(x_2)\big] + \kappa\mathbb{E}\Big[\Big(1 + J_\infty(x_2) - \mathbb{E}\big[1 + J_\infty(x_2)\big]\Big)_+\Big] \\ &= 1 + \frac{1}{2}J_\infty(1) + \kappa\frac{1}{2}\Big(J_\infty(1) - \frac{1}{2}J_\infty(1)\Big) = 1 + \frac{2+\kappa}{4}J_\infty(1). \end{aligned}$$

Thus $J_\infty(1) = 4/(2-\kappa)$, which is finite for all $\kappa \in [0,1]$, which are all values of κ for which the model defines a coherent measure of risk.

It follows that deeper properties of the measures of risk and their interplay with the transition kernel need to be investigated to answer the question about finiteness of the dynamic measure of risk in this case. We propose a condition that generalizes the Pliska condition (1) to the risk-averse case.

Recall that with every risk transition mapping $\sigma(\cdot,\cdot,\cdot)$, every controlled kernel Q, and every decision rule π, a multikernel $\mathfrak{M}^\pi$ is associated, as defined in (18). Similarly to the expected value case, it is convenient to consider the effective state space $\widetilde{\mathscr{X}} = \mathscr{X} \setminus \{x_{\mathrm{A}}\}$, and the *effective substochastic multikernel* $\widetilde{\mathfrak{M}}^\pi$ whose arguments are restricted to $\widetilde{\mathscr{X}}$ and whose values are convex sets of nonnegative measures on $\widetilde{\mathscr{X}}$, so that $\widetilde{M}\big(B\big|x,u\big) = M\big(B\big|x,u\big)$, for all Borel sets $B \subset \widetilde{\mathscr{X}}$, and all $M \in \widetilde{\mathfrak{M}}^\pi$.

Definition 6.1. *We call the Markov model with a risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ and with a stationary Markov policy $\{\pi,\pi,\dots\}$* risk-transient *if a weight function $w:\mathscr{X}\to[1,\infty)$, $w\in\mathscr{V}$, and a constant K exist such*

$$\sum_{j=1}^{\infty}\left\|\left(\widetilde{\mathfrak{M}}^{\pi}\right)^{j}\right\|_{w}\le K. \tag{29}$$

If the estimate (29) *is uniform for all Markov policies, the model is called* uniformly risk-transient.

In the special case of a risk-neutral model, Definition 6.1 reduces to the Pliska condition (1), owing to the equation (20).

Example 6.2. Consider the simple transient chain of Example 6.1 with the Average Value at Risk from Examples 3.2 and 4.2, where $0<\alpha\le 1$. From (15) we obtain

$$\mathscr{A}(x,m)=\left\{(\mu_1,\mu_2):0\le\mu_j\le\frac{m_j}{\alpha},\ j=1,2;\ \mu_1+\mu_2=1\right\}.$$

As only one control is possible, formula (18) simplifies to

$$\mathfrak{M}_i=\left\{(\mu_1,\mu_2):0\le\mu_j\le\frac{Q_{ij}}{\alpha},\ j=1,2;\ \mu_1+\mu_2=1\right\},\quad i=1,2.$$

The effective state space is just $\widetilde{\mathscr{X}}=\{1\}$, and we conclude that the effective multikernel has the form

$$\widetilde{\mathfrak{M}}_1=\left[0,\min\left(1,\frac{1}{2\alpha}\right)\right].$$

For $0<\alpha\le\frac{1}{2}$ we can select $\widetilde{M}=1\in\widetilde{\mathfrak{M}}_1$ to show that $1\in\left(\widetilde{\mathfrak{M}}_1\right)^j$ for all j, and thus condition (29) is not satisfied. On the other hand, if $\frac{1}{2}<\alpha\le 1$, then for every $\widetilde{M}\in\widetilde{\mathfrak{M}}_1$ we have $0\le\widetilde{M}<1$, and condition (29) is satisfied.

Consider now the mean-semideviation model of Examples 3.1 and 4.1. From (14) we obtain

$$\begin{aligned}\mathscr{A}(x,m)&=\left\{(\mu_1,\mu_2):\mu_j=m_j\left(1+h_j-(h_1m_1+h_2m_2)\right),\ 0\le h_j\le\kappa,\ j=1,2\right\},\\ \mathfrak{M}_i&=\left\{(\mu_1,\mu_2):\mu_j=Q_{ij}\left(1+h_j-(h_1Q_{i1}+h_2Q_{i2})\right),\ 0\le h_j\le\kappa,\ j=1,2\right\},\quad i=1,2.\end{aligned}$$

Calculating the lowest and the largest possible values of μ_1 we conclude that

$$\widetilde{\mathfrak{M}}_1=\left[\frac{1}{2}\left(1-\frac{\kappa}{2}\right),\frac{1}{2}\left(1+\frac{\kappa}{2}\right)\right].$$

For every $\kappa\in[0,1]$, Definition 6.1 is satisfied.

We start our analysis from an estimate of the risk in a finite horizon model of a final cost given by a certain function $v(x_T)$, where T is the horizon, and $v:\mathscr{X}\to\mathbb{R}$ is a measurable function with $\|v\|_w<\infty$ for the weight function $w:\mathscr{X}\to[1,\infty)$, $w\in\mathscr{V}$, and with $v(x_{\mathrm{A}})=0$. In the lemma below, we consider $x_1\in\mathscr{X}$ as a parameter of the problem, and thus $\rho_{1,T}\left(0,\dots,0,v(x_T)\right)$ is a function of x_1.

Lemma 6.1. *Suppose a stationary policy $\Pi=\{\pi,\pi,\dots\}$ is applied to a controlled Markov model with a Markov risk transition mapping $\sigma(\cdot,\cdot,\cdot)$. If the model is risk-transient, then there exists a function $\bar{v}_1:\mathscr{X}\to\mathbb{R}$, $\|\bar{v}_1\|_w<\infty$, such that for all $x_1\in\mathscr{X}$, and all $T\ge 1$*

$$\rho_{1,T}\left(0,\dots,0,v(x_T)\right)\le\bar{v}_1(x_1), \tag{30}$$

and

$$\left\|\bar{v}_1\right\|_w\le\left\|\left(\widetilde{\mathfrak{M}}^{\pi}\right)^{T-1}\right\|_w\cdot\left\|v\right\|_w, \tag{31}$$

where $\widetilde{\mathfrak{M}}^{\pi}$ is the substochastic risk multikernel implied by π and σ.

Proof. By construction,

$$\rho_{1,T}\big(0,\dots,0,v(x_T)\big) = \rho_1\Big(\rho_2\big(\cdots\rho_{T-1}\big(v(x_T)\big)\cdots\big)\Big).$$

Applying (19), we obtain

$$\rho_{T-1}\big(v(x_T)\big) = \max_{m_{T-1}\in\mathfrak{M}^\pi_{x_{T-1}}} \int\limits_{\mathscr{X}} v(y)\, m_{T-1}(dy). \tag{32}$$

It is a function of x_{T-1}, which we denote as $v_{T-1}(x_{T-1})$. Since $\|v\|_w < \infty$ and $w \in \mathscr{V}$, then $v \in \mathscr{V}$. As the sets $\mathfrak{M}^\pi_x$ are weakly* compact, the maximum in (32) is achieved. Moreover,

$$\big\|v_{T-1}\big\|_w \le \big\|\widetilde{\mathfrak{M}}^\pi\big\|_w \cdot \big\|v\big\|_w < \infty.$$

One step earlier, in a similar way we obtain

$$\begin{aligned}\rho_{T-2}\big(\rho_{T-1}\big(v(x_T)\big)\big) &= \max_{m_{T-2}\in\mathfrak{M}^\pi_{x_{T-2}}} \int\limits_{\mathscr{X}} v_{T-1}(y)\, m_{T-2}(dy)\\ &= \max_{m_{T-2}\in\mathfrak{M}^\pi_{x_{T-2}}} \int\limits_{\mathscr{X}} \max_{m_{T-1}\in\mathfrak{M}^\pi_y} \int\limits_{\mathscr{X}} v(z)\, m_{T-1}(dz)\, m_{T-2}(dy).\end{aligned}$$

The maximizers $\hat{m}_{T-1} \in \mathfrak{M}^\pi_y$ under the integral can be chosen in such a way that they form a measurable selector $M_{T-1} \ll \mathfrak{M}^\pi$ (see, e.g., [35, Thm. 14.37]. On the other hand, no measurable selector can do better than the pointwise maximizers. We can, therefore, interchange the operations of maximization and integration and conclude that

$$\rho_{T-2}\big(\rho_{T-1}\big(v(x_T)\big)\big) = \max_{m_{T-2}\in\mathfrak{M}^\pi_{x_{T-2}}} \max_{M_{T-1}\ll\mathfrak{M}^\pi} \int\limits_{\mathscr{X}}\int\limits_{\mathscr{X}} v(z)\, M_{T-1}(dz|y)\, m_{T-2}(dy).$$

Similarly, the outer maximizer may be represented as a value of a certain measurable selector of $\mathfrak{M}^\pi$ at x_{T-2}. Denoting the value of the above expression by $v_{T-2}(x_{T-2})$, we obtain

$$v_{T-2}(x) = \max_{M_{T-2}\ll\mathfrak{M}^\pi} \max_{M_{T-1}\ll\mathfrak{M}^\pi} \int\limits_{\mathscr{X}}\int\limits_{\mathscr{X}} v(z)\, M_{T-1}(dz|y)\, M_{T-2}(dy|x).$$

Changing the order of integration we observe that the double integral above can be represented as an integral with respect to a composition of the kernels M_{T-2} and M_{T-1} (*cf.* formula (11)). We obtain

$$v_{T-2}(x) = \max_{M_{T-2}\ll\mathfrak{M}^\pi} \max_{M_{T-1}\ll\mathfrak{M}^\pi} \int\limits_{\mathscr{X}} v(z)\, \big[M_{T-2}\circ M_{T-1}\big]\big(dz\big|x\big) \le \max_{M\ll(\mathfrak{M}^\pi)^2} \int\limits_{\mathscr{X}} v(z)\, M\big(dz\big|x\big) = \bar{v}_{T-2}(x).$$

The last inequality follows from the fact that $M_{T-2}\circ M_{T-1} \ll \big(\mathfrak{M}^\pi\big)^2$. Therefore, $v_{T-2} \le \bar{v}_{T-2}$, where

$$\big\|\bar{v}_{T-2}\big\|_w \le \Big\|\big(\widetilde{\mathfrak{M}}^\pi\big)^2\Big\|_w \cdot \big\|v\big\|_w < \infty.$$

Continuing in this way, we conclude that

$$\rho_1\Big(\rho_2\Big(\cdots\rho_{T-1}\big(v(x_T)\big)\cdots\Big)\Big) \le \max_{M\ll(\mathfrak{M}^\pi)^{T-1}} \int\limits_{\mathscr{X}} v(z)\, M\big(dz\big|x_1\big) = \max_{\widetilde{M}\ll(\widetilde{\mathfrak{M}}^\pi)^{T-1}} \int\limits_{\widetilde{\mathscr{X}}} v(z)\, \widetilde{M}\big(dz\big|x_1\big).$$

Denoting the right-hand side by $\bar{v}_1(x_1)$, we obtain the estimates (30)–(31). □

We can now provide sufficient conditions for the finiteness of the limit (26).

Theorem 6.1. *Suppose a stationary policy $\Pi = \{\pi, \pi, \dots\}$ is applied to a the controlled Markov model with a Markov risk transition mapping $\sigma(\cdot,\cdot,\cdot)$. If the model is risk-transient for the policy Π and the cost function $c(\cdot,\cdot,\cdot)$ is w-bounded, then the limit* (26) *is finite and $\left\|J_\infty(\Pi,\cdot)\right\|_w < \infty$. If the model is uniformly risk-transient, then $\left\|J_\infty(\Pi,\cdot)\right\|_w$ is uniformly bounded.*

Proof. By Lemma 3.1, each conditional risk measure $\rho_{1,T}(\cdot)$ is convex and positively homogeneous, and thus subadditive. For any $1 < T_1 < T_2$ we obtain the following estimate of (27):

$$\begin{aligned} J_{T_2-1}(\Pi,x_1) &= \rho_{1,T_2}(0,Z_2,\dots,Z_{T_2}) \\ &\le \rho_{1,T_2}(0,Z_2,\dots,Z_{T_1},0,\dots,0) + \sum_{j=T_1}^{T_2-1} \rho_{1,T_2}(0,\dots,0,Z_{j+1},0,\dots,0) \\ &= \rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) + \sum_{j=T_1}^{T_2-1} \rho_{1,j+1}(0,\dots,0,Z_{j+1}). \end{aligned} \tag{33}$$

By assumption, $Z_{j+1} \le C\big(\bar w(x_j) + \bar w(x_{j+1})\big)$, where $\bar w(x) = w(x)$ if $x \in \widetilde{\mathscr{X}}$, and $\bar w(x_{\mathrm{A}}) = 0$. Owing to the monotonicity and positive homogeneity of the conditional risk mappings

$$\begin{aligned} \rho_{1,j+1}(0,\dots,0,Z_{j+1}) &\le C\rho_1\Big(\rho_2\Big(\cdots\rho_{j-1}\Big(\rho_j\big(\bar w(x_j)+\bar w(x_{j+1})\big)\Big)\cdots\Big)\Big) \\ &= C\rho_1\Big(\rho_2\Big(\cdots\rho_{j-1}\big(\bar w(x_j)+\rho_j\big(\bar w(x_{j+1})\big)\big)\cdots\Big)\Big) \\ &\le C\rho_1\big(\rho_2\big(\cdots\rho_{j-1}\big(\bar w(x_j)\big)\cdots\big)\big) + C\rho_1\big(\rho_2\big(\cdots\rho_j\big(\bar w(x_{j+1})\big)\cdots\big)\big). \end{aligned}$$

In the middle equation we used the fact that $\bar w(x_j)$ is $\mathscr{F}_j$-measurable, and in the last inequality – the subadditivity of the risk measures. Since $\left\|\bar w\right\|_w = 1$, Lemma 6.1 implies that

$$\rho_1\big(\rho_2\big(\cdots\rho_j\big(\bar w(x_{j+1})\big)\cdots\big)\big) \le \bar v_j(x_1)$$

with

$$\left\|\bar v_j\right\|_w \le \left\|\big(\widetilde{\mathfrak{M}}^\pi\big)^j\right\|_w. \tag{34}$$

Substitution to (33) yields the estimate

$$\rho_{1,T_2}(0,Z_2,\dots,Z_{T_2}) \le \rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) + 2C\sum_{j=T_1+1}^{T_2} \bar v_j(x_1). \tag{35}$$

Consider now the sequence of costs $Z_1,\dots,Z_{T_1},-Z_{T_1+1},\dots,-Z_{T_2}$, in which we flip the sign of the costs $Z_{t+1} = c(x_t,u_t,x_{t+1})$ for $t \ge T_1$. As $|Z_{t+1}|$ are bounded by $C\big(\bar w(x_t)+\bar w(x_{t+1})\big)$, the estimate (35) applies to the new sequence. We obtain

$$\rho_{1,T_2}(0,Z_2,\dots,Z_{T_1},-Z_{T_1+1},\dots,-Z_{T_2}) \le \rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) + 2C\sum_{j=T_1+1}^{T_2} \bar v_j(x_1). \tag{36}$$

By convexity and positive homogeneity of $\rho_{1,T_2}(\cdot)$,

$$2\rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) \le \rho_{1,T_2}(0,Z_2,\dots,Z_{T_1},Z_{T_1+1},\dots,Z_{T_2}) + \rho_{1,T_2}(0,Z_2,\dots,Z_{T_1},-Z_{T_1+1},\dots,-Z_{T_2}).$$

Substituting the estimate (36), we deduce that

$$\rho_{1,T_2}(0,Z_2,\dots,Z_{T_2}) \ge \rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) - 2C\sum_{j=T_1+1}^{T_2} \bar v_j(x_1).$$

This combined with (35) yields

$$\left|J_{T_2-1}(\Pi,x_1)-J_{T_1-1}(\Pi,x_1)\right| \le 2C \sum_{j=T_1+1}^{T_2} \bar{v}_j(x_1).$$

In view of (34), we conclude that

$$\left\|J_{T_2-1}(\Pi,\cdot)-J_{T_1-1}(\Pi,\cdot)\right\|_w \le 2C \sum_{j=T_1+1}^{T_2} \left\| \left(\widetilde{\mathfrak{M}}^{\pi}\right)^j \right\|_w. \tag{37}$$

By Definition 6.1, the right hand side of the last displayed inequality converges to 0, when $T_1, T_2 \to \infty$, $T_1 < T_2$. Hence, the sequence of functions $J_T(\Pi,\cdot)$, $T = 1,2,\dots$ is convergent to some limit $J_\infty(\Pi,\cdot)$. Moreover, $\left\|J_\infty(\Pi,\cdot)\right\|_w < \infty$. If the model is uniformly risk-transient, then the estimate (37) is the same for all Markov policies Π, and thus $\left\|J_\infty(\Pi,\cdot)\right\|_w$ is uniformly bounded. □

Remark 6.1. *It is clear from the proof of Theorem 6.1, that*

$$J_\infty(\Pi,x_1) = \lim_{T\to\infty} \rho_{1,T}\big(0,Z_2,\dots,Z_T+f(x_T)\big), \tag{38}$$

for any measurable function $f : \mathscr{X} \to \mathbb{R}$, with $\left\|f\right\|_w < \infty$, because $c(x_{T-1},u_t,x_T)+f(x_T)$ is still w-bounded.

This analysis allows us to derive dynamic programming equations for the infinite horizon problem, in the case of a fixed Markov policy.

Theorem 6.2. *Suppose a controlled Markov model with a Markov risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ is risk-transient for the stationary Markov policy $\Pi = \{\pi,\pi,\dots\}$, with some weight function $w(\cdot)$. Then a measurable function $v : \mathscr{X} \to \mathbb{R}$, with $\|v\|_w < \infty$, satisfies the equations*

$$v(x) = \sigma\big(c_x + v, x, \pi(x)\circ Q_x\big), \quad x \in \widetilde{\mathscr{X}}, \tag{39}$$

$$v(x_{\mathrm{A}}) = 0, \tag{40}$$

if and only if $v(x) = J_\infty(\Pi,x)$ for all $x \in \mathscr{X}$.

Proof. Denote $Z_t = c(x_{t-1},u_{t-1},x_t)$. Suppose a measurable function $v(\cdot)$ satisfies the dynamic programming equations (39)–(40). Since $\|v\|_w < \infty$ and $w \in \mathscr{V}$, then also $v \in \mathscr{V}$. By assumption, $c(\cdot,\cdot,\cdot)$ is w-bounded, and thus $c_x(\cdot,\cdot) \in \mathscr{V}$. Consequently, the right-hand side of (39) is well-defined. By iteration of (39), we obtain for all $x_1 \in \mathscr{X}$ the following equation:

$$v(x_1) = \rho_1\Big(c(x_1,u_1,x_2) + \rho_2\Big(c(x_2,u_2,x_3) + \cdots + \rho_T\big(c(x_T,u_T,x_{T+1}) + v(x_{T+1})\big)\cdots\Big)\Big).$$

Denote $Z_t = c(x_{t-1},u_{t-1},x_t)$. Using monotonicity and subadditivity of the conditional risk measures we deduce that:

$$\rho_{1,T+1}\big(0,Z_2,\dots,Z_{T+1}+v(x_{T+1})\big) \le \rho_{1,T+1}\big(0,Z_2,\dots,Z_{T+1}\big) + \rho_{1,T+1}\big(0,0,\dots,v(x_{T+1})\big). \tag{41}$$

By Lemma 6.1,

$$v(x_1) = \rho_{1,T+1}\big(0,Z_2,\dots,Z_{T+1}+v(x_{T+1})\big) \le \rho_{1,T+1}\big(0,Z_2,\dots,Z_{T+1})\big) + d_T(x_1), \tag{42}$$

with

$$\left\|d_T\right\|_w \le \left\| \left(\widetilde{\mathfrak{M}}^{\pi}\right)^{T-1} \right\|_w \cdot \left\|v\right\|_w. \tag{43}$$

By convexity of $\rho_{1,T+1}(\cdot)$,

$$\begin{aligned} 2\rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}\big) &\le \rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}+v(x_{T+1})\big)+\rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}-v(x_{T+1})\big) \\ &= v(x_1)+\rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}-v(x_{T+1})\big). \end{aligned} \tag{44}$$

Similar to (41)–(42),

$$\rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}-v(x_{T+1})\big) \le \rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1}\big)+d_T(x_1).$$

Substituting into (44) we obtain

$$v(x_1) \ge \rho_{1,T+1}\big(0,Z_2,\ldots,Z_{T+1})\big)-d_T(x_1).$$

Combining this estimate with (42) and using (43) we conclude that

$$\big\|v(\cdot)-J_T(\Pi,\cdot)\big\|_w \le \big\|d_T\big\|_w \to 0, \quad \text{as} \quad T\to\infty.$$

Thus $v(\cdot)\equiv J_\infty(\Pi,\cdot)$, as postulated.

To prove the converse implication we can use the fact that all conditional risk measures $\rho_t(\cdot)$ share the same risk transition mapping to rewrite (27) as follows:

$$J_T(\Pi,x_1)=\rho_1\big(c(x_1,u_1,x_2)+J_{T-1}(\Pi,x_2)\big).$$

The function $\rho_1(\cdot)$, as a finite-valued coherent measure of risk on a Banach lattice, is continuous (see [38, Prop. 3.1]). Since $\big\|J_T(\Pi,\cdot)-J_\infty(\Pi,\cdot)\big\|_w\to 0$, as $T\to\infty$, then the sequence $\big\{J_T(\Pi,\cdot)\big\}$ is also convergent in the space $\mathscr{V}$. Therefore,

$$\lim_{T\to\infty} J_T(\Pi,x_1)=\rho_1\Big(c(x_1,u_1,x_2)+\lim_{T\to\infty}J_{T-1}(\Pi,x_2)\Big).$$

This is identical with equation (39) with $v(\cdot)\equiv J_\infty(\Pi,\cdot)$. Equation (40) is obvious. □

7 Dynamic Programming Equations for Infinite Horizon Problems

We shall now focus on the optimal value function

$$J^*(x)=\inf_{\Pi\in\Pi^{\mathrm{RM}}} J_\infty(\Pi,x), \quad x\in\mathscr{X}, \tag{45}$$

where Π^{RM} is the set of all stationary Markov policies.

Theorem 7.1. *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 and such that for every $x\in\mathscr{X}$ the multifunction $\mathscr{A}(x,\cdot)$ is lower semicontinuous;*

(iii) *The function $c(\cdot,\cdot,\cdot)$ is w-bounded and lower semicontinuous with respect to the second argument;*

(iv) *For every $x\in\mathscr{X}$ the set $U(x)$ is compact;*

(v) *The model is uniformly risk-transient.*

Then a function $v : \mathscr{X} \to \mathbb{R}$, with $\|v\|_w < \infty$, satisfies the equations

$$v(x) = \inf_{\lambda \in \mathscr{P}(U(x))} \sigma\big(c_x + v, x, \lambda \circ Q_x\big), \quad x \in \mathscr{X}, \tag{46}$$

$$v(x_{\mathrm{A}}) = 0, \tag{47}$$

if and only if $v(x) = J^*(x)$ for all $x \in \mathscr{X}$. Moreover, the minimizer $\pi^*(x)$, $x \in \mathscr{X}$, on the right hand side of (46) *exists and defines an optimal randomized Markov policy $\Pi^* = \{\pi^*, \pi^*, \dots\}$.*

Proof. Suppose $J^*(\cdot)$ is given by (45). The set of policies of the form $\{\lambda, \pi, \pi, \dots\}$ is larger than Π^{RM}, and thus

$$J^*(x_1) \geq \inf_{\substack{\lambda \in \mathscr{P}(U(x_1)) \\ \Pi \in \Pi^{\mathrm{RM}}}} \rho_1\big(c(x_1, u_1, x_2) + J_\infty(\Pi, x_2)\big).$$

By the monotonicity of $\rho_1(\cdot)$ we can move the infimum operator inside:

$$\begin{aligned} J^*(x_1) &\geq \inf_{\lambda \in \mathscr{P}(U(x_1))} \rho_1\Big(c(x_1, u_1, x_2) + \inf_{\Pi \in \Pi^{\mathrm{RM}}} J_\infty(\Pi, x_2)\Big) \\ &= \inf_{\lambda \in \mathscr{P}(U(x_1))} \rho_1\big(c(x_1, u_1, x_2) + J^*(x_2)\big). \end{aligned}$$

As the model is uniformly risk-transient, $\big\|J^*\big\|_w < \infty$, and the right-hand side is well-defined. Thus $J^*(\cdot)$ satisfies the inequality

$$J^*(x) \geq \inf_{\lambda \in \mathscr{P}(U(x))} \sigma\big(c_x + J^*, x, \lambda \circ Q_x\big), \quad x \in \mathscr{X}. \tag{48}$$

The mapping $\lambda \mapsto \sigma\big(c_x + J^*, x, \lambda \circ Q_x\big)$ is continuous for all x, and the set of $\lambda \in \mathscr{P}(\mathscr{U})$ such that $\lambda(U(x)) = 1$ is weakly* compact. Therefore, there exists a minimizer $\pi^*(x)$ on the right hand side of (48). Hence,

$$J^*(x) \geq \sigma\big(c_x + J^*, x, \pi^*(x) \circ Q_x\big), \quad x \in \mathscr{X}.$$

Iterating this inequality we conclude that $J^*(x_1)$ is bounded below by

$$J^*(x_1) \geq \rho_{1,T}\big(0, Z_2, \dots, Z_T + J^*(x_T)\big), \tag{49}$$

with the sequence of controls and states resulting from the stationary Markov policy $\Pi^* = \{\pi^*, \pi^*, \dots\}$. Owing to Remark 6.1, we can pass to the limit on the right-hand side and obtain the inequality:

$$J^*(x_1) \geq J_\infty(\Pi^*, x_1), \quad x_1 \in \mathscr{X}.$$

It follows that Π^* is the optimal stationary Markov policy, and thus $J^*(\cdot) = J_\infty(\Pi^*, \cdot)$. By Theorem 6.2, relation (48) is an equation, which proves (46)–(47).

To prove the converse implication, suppose $v(\cdot)$ satisfies (46)–(47) and $\|v\|_w < \infty$. By the continuity of the mapping $\lambda \mapsto \sigma\big(c_x + v, x, \lambda \circ Q_x\big)$ and weak* compactness of the set of $\lambda \in \mathscr{P}(\mathscr{U})$ such that $\lambda(U(x)) = 1$, there exists a randomized control $\hat{\pi}(\cdot)$, which is a minimizer on the right hand side of (46). We obtain the equation

$$v(x) = \sigma\big(c_x + v, x, \hat{\pi}(x) \circ Q_x\big), \quad x \in \mathscr{X}.$$

By Theorem 6.2,

$$v(x) = J_\infty(\hat{\Pi}, x) \geq J^*(x), \quad x \in \mathscr{X}, \tag{50}$$

where $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}, \dots\}$. On the other hand, it follows from (46) that for the optimal policy $\Pi^* = \{\pi^*, \pi^*, \dots\}$ we have

$$v(x) \leq \sigma\big(c_x + v, x, \pi^*(x) \circ Q_x\big), \quad x \in \mathscr{X}. \tag{51}$$

The risk transition mapping σ is nondecreasing with respect to the first argument. Therefore, iterating inequality (51) we obtain an inequality corresponding to (49):

$$v(x_1) \le \rho_{1,T}\big(0, Z_2, \dots, Z_T + v(x_T)\big),$$

Passing to the limit with $T \to \infty$ and applying Remark 6.1, we conclude that

$$v(x) \le J_\infty(\Pi^*, x) = J^*(x), \quad x \in \mathscr{X}.$$

The last estimate together with (50) implies that $v(\cdot) \equiv J^*(\cdot)$ and that both stationary policies Π^* and $\hat{\Pi}$ are optimal. □

We can now address the case of general non-stationary policies. For a policy $\Lambda = \{\lambda_1, \lambda_2, \dots\}$ we define

$$J_\infty(\Lambda, x_1) = \liminf_{T\to\infty} J_T(\Lambda, x_1)$$

and

$$\hat{J}(x_1) = \inf_{\Lambda} J_\infty(\Lambda, x_1).$$

Theorem 7.2. *Assume that the conditions of Theorem 7.1 are satisfied, together with the following assumption: there exists a constant C such that $J_\infty(\Lambda, x) \ge -Cw(x)$ for all $x \in \mathscr{X}$ and for all policies Λ. Then a function $v : \mathscr{X} \to \mathbb{R}$, with $\|v\|_w < \infty$, satisfies the equations* (46)–(47) *if and only if $v(x) = \hat{J}(x)$ for all $x \in \mathscr{X}$. Moreover, the minimizer $\pi^*(x)$, $x \in \mathscr{X}$, on the right hand side of* (46) *exists and defines an optimal policy $\Pi^* = \{\pi^*, \pi^*, \dots\}$.*

Proof. As for stationary Markov policies Π we have $\big\|J_\infty(\Pi, \cdot)\big\|_w < \infty$, in view of the additional assumption we have $\big\|\hat{J}\big\|_w < \infty$. Denote $\Lambda^1 = \{\lambda_2, \lambda_3, \dots\}$. Due to the monotonicity and continuity of $\rho_1(\cdot)$, we have the chain of relations

$$\begin{aligned}
\hat{J}(x_1) &= \inf_{\lambda_1, \lambda_2, \dots} \liminf_{T\to\infty} \rho_1\big(c(x_1, u_1, x_2) + J_{T-1}(\Lambda^1, x_2)\big) \\
&\ge \inf_{\lambda_1, \lambda_2, \dots} \liminf_{T\to\infty} \rho_1\big(c(x_1, u_1, x_2) + \inf_{\tau \ge T-1} J_\tau(\Lambda^1, x_2)\big) \\
&= \inf_{\lambda_1, \lambda_2, \dots} \lim_{T\to\infty} \rho_1\big(c(x_1, u_1, x_2) + \inf_{\tau \ge T-1} J_\tau(\Lambda^1, x_2)\big) \\
&= \inf_{\lambda_1, \lambda_2, \dots} \rho_1\Big(c(x_1, u_1, x_2) + \liminf_{T\to\infty} J_{T-1}(\Lambda^1, x_2)\Big) = \inf_{\lambda_1, \lambda_2, \dots} \rho_1\big(c(x_1, u_1, x_2) + J_\infty(\Lambda^1, x_2)\big),
\end{aligned}$$

Owing to the monotonicity of $\rho_1(\cdot)$, we can move the minimization with respect to Λ^1 inside the argument, to obtain

$$\hat{J}(x_1) \ge \inf_{\lambda_1} \rho_1\Big(c(x_1, u_1, x_2) + \inf_{\Lambda^1} J_\infty(\Lambda^1, x_2)\Big) = \inf_{\lambda_1} \rho_1\big(c(x_1, u_1, x_2) + \hat{J}(x_2)\big).$$

Thus $\hat{J}(\cdot)$ satisfies an inequality analogous to (48):

$$\hat{J}(x) \ge \inf_{\lambda \in \mathscr{P}(U(x))} \sigma\big(c_x + \hat{J}, x, \lambda \circ Q_x\big), \quad x \in \mathscr{X}. \tag{52}$$

We can now repeat the argument from the proof of Theorem 7.1. Denoting by $\hat{\lambda}$ the minimizer above, iterating inequality (52), and passing to the limit we conclude that

$$\hat{J}(x) \ge J_\infty(\hat{\Lambda}, x), \quad x \in \mathscr{X},$$

where $\hat{\Lambda} = \{\hat{\lambda}, \hat{\lambda}, \dots\}$ is a stationary Markov policy. Therefore, optimization with respect to stationary Markov policies is sufficient, and the result follows from Theorem 7.1. □

Our additional technical assumption that $J_\infty(\Lambda,x) \geq -Cw(x)$ is obviously true for nonnegative costs $c(\cdot,\cdot,\cdot)$. More generally, it is true in the case when the cost function is w-bounded, the model is transient, and $\mu \in \mathscr{A}(x,\mu)$, for all $x \in \mathscr{X}$ and $\mu \in \mathscr{M}$. Indeed, by virtue of Remark 4.1, the dynamic risk measure is bounded from below by the expected value of the cost, which is finite in this case.

8 Randomized versus Deterministic Control

Observe that the mapping $\lambda \mapsto \sigma\big(c_x + v, x, \lambda \circ Q_x\big)$, which plays the key role in the dynamic programming equation (46), is nonlinear, in general, as opposed to the expected value model, where

$$\sigma\big(c_x + v, x, \lambda \circ Q_x\big) = \int_{U(x)} \int_{\mathscr{X}} \big(c(x,u,y) + v(y)\big)\, Q(dy|x,u)\, \lambda(du|x).$$

In the expected value case, it is sufficient to consider only the extreme points of the set $\mathscr{P}\big(U(x)\big)$, which are the measures assigning unit mass to one of the controls $u \in U(x)$:

$$\inf_{\lambda \in \mathscr{P}(U(x))} \int_{U(x)} \int_{\mathscr{X}} \big(c(x,u,y) + v(y)\big)\, Q(dy|x,u)\, \lambda(du|x) = \inf_{u \in U(x)} \int_{\mathscr{X}} \big(c(x,u,y) + v(y)\big)\, Q(dy|x,u).$$

In the risk averse case this simplification is not justified and a randomized policy may be strictly better than any deterministic policy. Of course, we may always restrict the set of possible decision rules to deterministic rules, and solve the corresponding version of the dynamic equation (46):

$$v(x) = \min_{\lambda \in \mathscr{P}^\delta(U(x))} \sigma\big(c_x + v, x, \lambda \circ Q_x\big), \quad x \in \mathscr{X}, \tag{53}$$

where $\mathscr{P}^\delta(U(x))$ denotes the set of Dirac measures supported at $U(x)$. For a fixed $x \in \mathscr{X}$ and a Dirac measure $\lambda = \delta_u$, the function $c_x + v = c(x,u) + v(y)$ is only a function of the next state $y \in \mathscr{X}$, and the measure $\lambda \circ Q_x$ is the measure $Q(\cdot|x,u)$ on the state space $\mathscr{X}$. We can, therefore, rewrite (53) in a simpler form

$$v(x) = \min_{u \in U(x)} \Big\{ c(x,u) + \sigma\big(v, x, Q(\cdot|x,u)\big) \Big\}, \quad x \in \mathscr{X}, \tag{54}$$

where (with a slight abuse of notation) $\sigma : \mathscr{L}_p(\mathscr{X}, \mathscr{B}(\mathscr{X}), P_x) \times \mathscr{X} \times \mathscr{L}_q(\mathscr{X}, \mathscr{B}(\mathscr{X}), P_x) \to \mathbb{R}$, and $\sigma(\cdot,\cdot,\cdot)$ is a coherent measure of risk with respect to its first argument. In equation (54) we also used the translation property of coherent measures of risk. This is almost exactly the form of the dynamic programming equation which we derived in [36] for discounted problems, but with the discount factor $\alpha = 1$.

A question arises whether it is possible to identify cases in which deterministic policies are sufficient. It turns out that we can prove this for a class of measures of risk which are called *optimized certainty equivalents* [5]:

$$\sigma(\varphi, x, \mu) = \inf_{\xi \in \mathbb{R}} \int_{U(x) \times \mathscr{X}} \big\{ \xi + G\big(\varphi(u,y) - \xi; x\big) \big\}\, \mu(du \times dy). \tag{55}$$

In formula (55), the function $G : \mathbb{R} \to \mathbb{R}$ is nondecreasing and convex, with $G(0) = 0$ and $1 \in \partial G(0)$. We assume that $|G(z)| \leq c(1 + z^p)$ for all $z \in \mathbb{R}$, with some $c > 0$ and $p \geq 1$, and we define $\mathscr{V}$ using the same p, so that the integral above is finite for $\varphi \in \mathscr{V}$.

Lemma 8.1. *If the risk transition mapping has the form* (55) *then the dynamic programming equations* (46) *have a solution in deterministic decision rules.*

Proof. Interchanging the integration and the infimum in the definition of an optimized certainty equivalent, we obtain a lower bound

$$\begin{aligned}\sigma(\varphi,x,\lambda\circ Q_x) &= \inf_{\xi\in\mathbb{R}} \int_{U(x)} \int_{\mathscr{X}} \big\{\xi + G\big(\varphi(u,y)-\xi;x\big)\big\}\; Q(dy|x,u)\;\lambda(du|x)\\ &\geq \int_{U(x)} \inf_{\xi\in\mathbb{R}} \int_{\mathscr{X}} \big\{\xi + G\big(\varphi(u,y)-\xi;x\big)\big\}\; Q(dy|x,u)\;\lambda(du|x).\end{aligned}$$

The above inequality becomes an equation for every Dirac measure λ. On the right-hand side of (46) we have

$$\inf_{\lambda\in\mathscr{P}(U(x))} \sigma(c_x+v,x,\lambda\circ Q_x) \geq \inf_{\lambda\in\mathscr{P}(U(x))} \int_{U(x)} \inf_{\xi\in\mathbb{R}} \int_{\mathscr{X}} \big\{\xi + G\big(c(x,u,y)+v(y)-\xi;x\big)\big\}\; Q(dy|x,u)\;\lambda(du|x).$$

As the right hand side achieves its minimum over $\lambda\in\mathscr{P}(U(x))$ at a Dirac measure concentrated at an extreme point of $U(x)$, and both sides coincide in this case, the minimum of the left hand side is also achieved at such measure. Consequently, for risk transition mappings of form (55) deterministic Markov policies are optimal. □

9 Illustrative Examples

We illustrate our models and results on two simple examples.

9.1 Asset Selling

Let us at first consider the classical example of asset selling originating from Karlin [19]. Offers Y_t arriving in time periods $t=1,2,\dots$ are independent integer-valued integrable random variables. At each time we may accept the highest offer received so far, or we may wait, in which case a waiting cost c is incurred. Denoting the random stopping time by τ we see that the total "cost" equals $Z=c\tau-\max_{0\leq j\leq\tau}Y_j$. The problem is an example of an *optimal stopping problem*, a structure of considerable theoretical and practical relevance (see, e.g., Çinlar [9], Dynkin and Yushkevich [11, 12], and Puterman [32]).

Formally, we introduce the state space $\mathscr{X}=\{x_{\mathrm{A}}\}\cup\{0,1,2,\dots\}$, where x_{A} is the absorbing state reached after the transaction, and the other states represent the highest offer received so far. The control space is $\mathscr{U}=\{0,1\}$, with 0 representing "wait" and 1 representing "sell." The state evolves according to the equation

$$x_{t+1}=\begin{cases}\max(x_t,Y_{t+1}) & \text{if } u_t=0,\\ x_{\mathrm{A}} & \text{if } u_t=1.\end{cases}$$

This formula defines the transition kernel Q. The cost is

$$c(x_t,u_t)=\begin{cases}c & \text{if } u_t=0,\\ -x_t & \text{if } u_t=1.\end{cases}$$

The expected value version of this problem has a known solution: accept the first offer greater than or equal to the solution $\hat{x}$ of the equation

$$c=\mathbb{E}\big[(Y-\hat{x})_+\big]. \tag{56}$$

We shall solve the risk-averse version of the problem in deterministic policies. For a stationary risk transition mapping σ, equation (54) has the form:

$$v(x)=\min\Big\{-x,c+\sigma\big(v,x,Q_x\big)\Big\},\quad x\in\widetilde{\mathscr{X}}. \tag{57}$$

Suppose σ is law invariant (Definition 3.2). As the distribution of v with respect to the measure Q_x is the same as the distribution of $v\big(\max(x,Y)\big)$ under the measure P_Y of Y, we obtain

$$\sigma\big(v,x,Q_x\big) = \sigma\big(v\big(\max(x,Y)\big),x,P_Y\big).$$

Suppose our attitude to risk does not depend on the current state, that is, σ does not depend on its second argument. Using (12), we may rewrite the last equation as follows:

$$\sigma\big(v,x,Q_x\big) = \max_{\mu\in\mathscr{A}} \mathbb{E}_\mu\big[v\big(\max(x,Y)\big)\big].$$

The convex closed set of probability measures $\mathscr{A}$ is fixed. Equation (57) takes on the form

$$v(x) = \min\Big\{-x, c + \max_{\mu\in\mathscr{A}} \mathbb{E}_\mu\big[v\big(\max(x,Y)\big)\big]\Big\}, \quad x\in\widetilde{\mathscr{X}}. \tag{58}$$

Observe that $v(x)\le -x$ and thus $v\big(\max(x,Y)\big)\le -\max(x,Y)$. The last displayed inequality implies that

$$v(x) \le \min\Big\{-x, c + \max_{\mu\in\mathscr{A}} \mathbb{E}_\mu\big[-\max(x,Y)\big]\Big\} = \min\Big\{-x, c - \min_{\mu\in\mathscr{A}} \mathbb{E}_\mu\big[\max(x,Y)\big]\Big\}, \quad x\in\widetilde{\mathscr{X}}.$$

If the offer at level x is accepted, then $v(x) = -x$. We obtain the inequality:

$$\min_{\mu\in\mathscr{A}} \mathbb{E}\big[(Y-x)_+\big] \le c.$$

This suggests the solution: *accept any offer that is greater or equal to the solution x^* of the equation*

$$\min_{\mu\in\mathscr{A}} \mathbb{E}\big[(Y-x^*)_+\big] = c; \tag{59}$$

if $x < x^*$, then wait. The corresponding value function equals:

$$v^*(x) = -\max(x,x^*).$$

Equation (58) can be verified by direct substitution.

Observe that the solution (59) of the risk-averse problem is closely related to the solution (56) of the expected value problem. The only difference is that we have to account for the least favorable distribution of the offers. If $P_Y\in\mathscr{A}$, then the critical level $x^*\le\hat{x}$.

9.2 Organ Transplant

We illustrate our results on a risk-averse version of a simplified organ transplant problem discussed in Alagoz et. al. [1]. We consider the discrete-time absorbing Markov chain depicted in Figure 1. State S, which is the initial state, represents a patient in need of an organ transplant. State L represents life after a successful transplant. State D (absorbing state) represents death. Two control values are possible in state S: W (for "Wait"), in which case transition to state D or back to state S may occur, and T (for "Transplant"), which results in a transition to states L or D. The probability of death is lower for W than for T, but successful transplant may result in a longer life, as explained below. In other two states only one (formal) control value is possible: "Continue". The rewards collected at each time step are months of life. In state S a reward equal to 1 is collected, if the control is W; otherwise, the immediate reward is 0. In state L the reward $r(\mathrm{L})$ is collected, representing the sure equivalent of the random length of life after transplant. In state D the reward is 0.

Generally, in a cost minimization problem, the value of a dynamic measure of risk (4) is the "fair" sure charge one would be willing to incur, instead of a random sequence of costs. In our case, which will be a

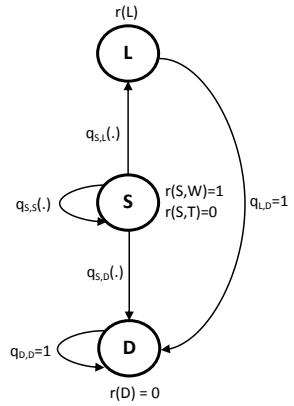

Figure 1: The organ transplant model.

maximization problem, we shall work with the negatives of the months of life as our "costs." The value of the measure of risk, therefore, can be interpreted as the negative of a sure life length which we consider to be equivalent to the random life duration faced by the patient.

Let us start from describing the way the deterministic equivalent length of life $r(\mathrm{L})$ at state L is calculated. The state L is in fact an aggregation of n states in a survival model representing months of life after transplant, as depicted in Figure 2.

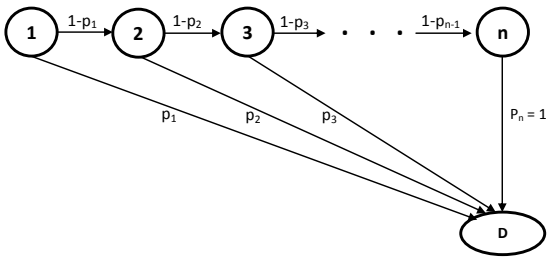

Figure 2: The survival model.

In state $i = 1,\ldots,n$, the patient dies with probability p_i and survives with probability $1-p_i$. The probability $p_n = 1$. The reward collected at each state $i = 1,\ldots,n$ is equal to 1. In order to follow the notation of our paper, we define the cost $c(\cdot) = -r(\cdot)$. For illustration, we apply the mean–semideviation model of Example 3.1 with $\kappa = 1$.

The risk transition mapping has the form:

$$\sigma(\varphi,i,\nu) = \underbrace{\mathbb{E}_\nu[\varphi]}_{\text{expected value}} + \kappa \underbrace{\mathbb{E}_\nu\big[\big(\varphi - \mathbb{E}_\nu[\varphi]\big)_+\big]}_{\text{semideviation}}. \tag{60}$$

Owing to the monotonicity property B2, $\sigma(\varphi,i,\nu) \le 0$, whenever $\varphi(\cdot) \le 0$.

In (60), the measure ν is the transition kernel at the current state i, and the function $\varphi(\cdot)$ is the cost incurred at the current state and control plus the value function at the next state. At each state $i = 1,\ldots,n-1$ two transitions are possible: to D with probability p_i and $\varphi = -1$, and to $i+1$ with probability $1-p_i$ and $\varphi = -1 + v_{i+1}(i+1)$. At state $i = n$ the transition to D occurs with probability 1, and $\varphi = -1$. Therefore, $v_n(n) = -1$.

The survival problem is a finite horizon problem, and thus we apply equation (23). As there is no control to choose, the minimization operation in is eliminated. The equation has the form:

$$v_i(i) = \sigma(\varphi, i, Q_i), \quad i = 1, \ldots, n-1,$$

with φ and Q_i as explained above. By induction, $v_i(i) \leq 0$, for $i = n-1, n-2, \ldots, 1$.

Let us calculate the mean and semideviation components of (60) at states $i = 1, \ldots, n-1$:

$$\begin{aligned}
&\mathbb{E}_{Q_i}[\varphi] = -p_i + (1-p_i)\big(-1 + v_{i+1}(i+1)\big) = -1 + (1-p_i)v_{i+1}(i+1),\\
&\mathbb{E}_{Q_i}\big[\big(\varphi - \mathbb{E}_{Q_i}[\varphi]\big)_+\big] = \mathbb{E}_{Q_i}\big[\big(\varphi + 1 - (1-p_i)v_{i+1}(i+1)\big)_+\big]\\
&\quad = p_i\big(-1+1-(1-p_i)v_{i+1}(i+1)\big)_+ + (1-p_i)\big(-1+v_{i+1}(i+1)+1-(1-p_i)v_{i+1}(i+1)\big)_+\\
&\quad = p_i\big(-(1-p_i)v_{i+1}(i+1)\big)_+ + (1-p_i)\big(p_i v_{i+1}(i+1)\big)_+\\
&\quad = -p_i(1-p_i)v_{i+1}(i+1).
\end{aligned}$$

In the last equation we used the fact that $v_{i+1}(i+1) \leq 0$. For $i = 1, \ldots, n-1$, the dynamic programming equations (23) take on the form:

$$v_i(i) = \underbrace{-1 + (1-p_i)v_{i+1}(i+1)}_{\text{expected value}} - \kappa \underbrace{p_i(1-p_i)v_{i+1}(i+1)}_{\text{semideviation}}, \quad i = n-1, n-2, \ldots, 1.$$

The value $v(1)$ is the negative of the risk-adjusted length of life with new organ. For $\kappa = 0$ the above formulas give the negative of the expected length of life with new organ.

In our calculations we used the transition data provided in Table 1. They have been chosen for purely illustrative purposes and do not correspond to any real medical situation.

Control	S	L	D
W	0.99882	0	0.00118
T	0	0.90782	0.09218

Table 1: Transition probabilities from state S.

For the survival model, we used the distribution function, $F(x)$, of lifetime of the American population from Jasiulewicz [18]. It is a mixture of Weibull, lognormal, and Gompertz distributions:

$$F(x) = w_1\Big(1 - \exp\Big(-\Big(\frac{x}{\delta}\Big)^{\beta}\Big)\Big) + w_2\Phi\Big(\frac{\log x - m}{\sigma}\Big) + w_3\Big(1 - \exp\Big(-\frac{b}{\alpha}(e^{\alpha x} - 1)\Big)\Big), \quad x \geq 0.$$

The values of the parameters and weights, provided by Jasiulewicz [18], are given in Table 2.

Distribution	Parameters	Weights
Weibull	$\delta = 0.297$, $\beta = 0.225$	$w_1 = 0.0170$
Lognormal	$m = 3.11$, $\sigma = 0.218$	$w_2 = 0.0092$
Gompertz	$b = 0.0000812$, $\alpha = 0.0844$	$w_3 = 0.9737$

Table 2: Values of parameters for $F(x)$.

Then, we calculated the probability of dying at age k (in months) as follows:

$$p_k = \frac{F(k/12 + 1/24) - F(k/12 - 1/24)}{1 - F(k/12 - 1/24)}, \quad k = 1, 2, \ldots.$$

The maximum lifetime of the patient was taken to be 1200 months, and that the patient after transplant has survival probabilities starting from $k = 300$. Therefore, $n = 900$ in the survival model used for calculating $r(\mathrm{L})$.

Let $\lambda = (\lambda_{\mathrm{W}}, \lambda_{\mathrm{T}})$ be the randomized policy in the state S and let $\Lambda = \{\lambda \in \mathbb{R}^2 : \lambda_{\mathrm{W}} + \lambda_{\mathrm{T}} = 1,\ \lambda \geq 0\}$. The dynamic programming equation (46) at S takes on the form

$$v(\mathrm{S}) = \min_{\lambda \in \Lambda} \Bigg\{ \underbrace{\lambda_{\mathrm{W}} \big[q_{\mathrm{S,S}}(\mathrm{W})\big(v(\mathrm{S}) - 1\big)\big) + q_{\mathrm{S,D}}(\mathrm{W})\big(v(\mathrm{D}) - 1\big)\big] + \lambda_{\mathrm{T}} \big[q_{\mathrm{S,L}}(\mathrm{T})v(\mathrm{L}) + q_{\mathrm{S,D}}(\mathrm{T})v(\mathrm{D})\big]}_{\text{expected value } \mu}$$
$$+ \kappa \Big(\underbrace{\lambda_{\mathrm{W}} \big[q_{\mathrm{S,S}}(\mathrm{W})\big(v(\mathrm{S}) - 1 - \mu\big)_+ + q_{\mathrm{S,D}}(\mathrm{W})\big(v(\mathrm{D}) - 1 - \mu\big)_+\big]}_{\text{semideviation} \ldots}$$
$$+ \underbrace{\lambda_{\mathrm{T}} \big[q_{\mathrm{S,L}}(\mathrm{T})\big(v(\mathrm{L}) - \mu\big)_+ + q_{\mathrm{S,D}}(\mathrm{T})\big(v(\mathrm{D}) - \mu\big)_+\big]}_{\ldots \text{semideviation}} \Big) \Bigg\}.$$

In the semideviation parts, we wrote μ for the expectation of the value function in the next state, which is given by the first underbraced expression, and which is also dependent on λ. Of course, the above expression can be simplified, by using the fact that $v(\mathrm{L}) < v(\mathrm{S}) < v(\mathrm{D}) = 0$, but we prefer to leave it in the above form to illustrate the way it has been developed.

We compared two optimal control models for this problem. The first one was the expected value model ($\kappa = 0$), which corresponds to the expected reward $r(\mathrm{L}) = 610.46$ in the survival model. Standard dynamic programming equations were solved, and the optimal decision in state S turned out to be W.

The second model was the risk-averse model using the mean–semideviation risk transition mapping with $\kappa = 1$. This changed the reward at state L to 515.35. We considered two versions of this model. In the first version, we restricted the feasible policies to be deterministic. In this case, the optimal action in state S was T. In the second version, we allowed randomized policies, as in our general model. Then the optimal policy in state S was W with probability $\lambda_{\mathrm{W}} = 0.9873$ and T with probability $\lambda_{\mathrm{T}} = 0.0127$.

How can we interpret these results? The optimal randomized policy results in a random waiting time before transplanting the organ. This is due to the fact that immediate transplant entails a significant probability of death, and a less risky policy is to "dilute" this probability in a long waiting time. This cannot be derived from an expected value model, no matter what the data, because deterministic policies are optimal in such a model: either transplant immediately or never.

References

[1] Alagoz, O., L. M. Maillart, A. J. Schaefer and M. S. Roberts, The optimal timing of living-donor liver transplantation, *Management Science*, 50, 1420–1430, 2004.

[2] Artzner, P., F. Delbaen, J. M. Eber and D. Heath, Coherent measures of risk, *Mathematical Finance*, 9, 203–228, 1999.

[3] Artzner, P., F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, Coherent multiperiod risk adjusted values and Bellmans principle, *Annals of Operations Research* 152, 5–22, 2007.

[4] Aubin, J.-P., and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.

[5] Ben Tal, A., and M. Teboulle, An old-new concept of convex risk measures: the optimized certainty equivalent, *Mathematical Finance* 17, 449-476, July 2007.

[6] Bertsekas, D. P., and Tsitsiklis J. N., An analysis of stochastic shortest-path problems, *Mathematics Of Operations Research*, 16 (3), 580–595, 1991.

[7] Cheridito, P., F. Delbaen, and M. Kupper, Dynamic monetary risk measures for bounded discrete-time processes. *Electronic Journal of Probability* 11, 57–106, 2006.

[8] Chew, S. H. and J. L. Ho, Hope: An empirical study of attitude toward the timing of uncertainty resolution, *Joumal of Risk and Uncertainty* 8 (3), 267–288, 1994.

[9] Çinlar, E., *Introduction to Stochastic Processes*, Prentice-Hall, Englewood Ciffs, 1975.

[10] Denardo, E. V. and U. G. Rothblum, Optimal stopping, exponential utility, and linear programming, *Mathematical Programming* 16, 228–244, 1979.

[11] Dynkin, E.B., and A.A. Yushkevich, *Markov Processes: Theory and Problems*, Plenum, New York, 1969.

[12] Dynkin, E.B., and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.

[13] Feinberg, E. A., and A. Shwartz, *Handbook of Markov Decision Processes: Methods and Applications*, Kluwer, Dordrecht, 2002.

[14] Fritelli, M., and G. Scandolo, Risk measures and capital requirements for processes. *Mathematical Finance*, 16, 589–612, 2006.

[15] González-Trejo, J. I., O. Hernández-Lerma, and L. F. Hoyos-Reyes, Minimax control of discrete-time stochastic systems, *SIAM J. Control Optim.*, 41, 1626-1659, 2003.

[16] Hernández-Lerma, O., and J. B. Lasserre, *Discrete-Time Markov Control Processes. Basic Optimality Criteria*, Springer, New York, 1996.

[17] Hernández-Lerma, O., and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer, New York, 1999.

[18] Jasiulewicz, H., Application of mixture models to approximation of age-at-death distribution, *Mathematics and Economics*, 19, 237–241, 1997.

[19] Karlin, S., Stochastic models and optimal policies for selling an asset, in: *Studies in Applied Probability and Management Science*, K. J. Arrow, S. Karlin, and S. Scarf (Eds.), Stanford University Press, Palo Alto, 1962, pp. 148–158.

[20] Klöppel, S., and M. Schweizer, Dynamic indifference valuation via convex risk measures, *Math. Finance*, 17, 599–627, 2007.

[21] Kurt, M., and J. P. Kharoufeh, Monotone Optimal Replacement Policies for a Markovian Deteriorating System in a Controllable Environment, *Operations Research Letters*, 38, 273–279, 2010.

[22] Leitner, J., A short note on second-order stochastic dominance preserving coherent risk measures, *Mathematical Finance*, 15, 649–651, 2005.

[23] Nie, Y., and Wu, X., Shortest path problem considering on-time arrival probability, *Transportation Research B*, 43 (6), 597–613, 2009.

[24] Nilim, A., and El Ghaoui, L., Robust control of Markov decision processes with uncertain transition matrices, *Operations Research* 53 (5), 780798, 2005.

[25] Ohtsubo, Y., Minimizing risk models in stochastic shortest path problems, *Mathematical Methods of Operations Research*, 57 (1), 79–88, 2003.

[26] Ogryczak, W., and A. Ruszczyński, From stochastic dominance to mean-risk models: Semideviations as risk measures, *European Journal of Operational Research*, 116, 33–50, 1999.

[27] Ogryczak, W., and A. Ruszczyński, On consistency of stochastic dominance and mean-semideviation models, *Mathematical Programming*, 89, 217–232, 2001.

[28] Ogryczak, W., and A. Ruszczyński, Dual stochastic dominance and related mean-risk models, *SIAM Journal on Optimization*, 13(1), 60–78, 2002.

[29] Patek, S. D., On terminating Markov decision processes with a risk averse objective function, *Automatica*, 37(9), 1379–1386, 2001.

[30] Pflug, G. Ch., and W. Römisch. *Modeling, Measuring and Managing Risk*. World Scientific, Singapore, 2007.

[31] Pliska, S. R., On the transient case for Markov decision chains with general state spaces, in: *Dynamic Programming and Its Applications*, M. L. Puterman (ed.), Academic Press, New York, 1979, pp. 335–349.

[32] Puterman, M. L., *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York, 1994.

[33] Riedel, F., Dynamic coherent risk measures, *Stochastic Processes and Their Applications*, 112, 185–200, 2004.

[34] R. T. Rockafellar and S. P. Uryasev, Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance*, 26, 1443–1471, 2002.

[35] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.

[36] Ruszczyński, A., Risk-averse dynamic programming for Markov decision processes, *Mathematical Programming*, Series B, 125, 235–261, 2010.

[37] Ruszczyński, A. and A. Shapiro, Optimization of risk measures, In *Probabilistic and Randomized Methods for Design under Uncertainty*, G. Calafiore and F. Dabbene (Eds.), Springer, London, 2005.

[38] Ruszczyński, A. and A. Shapiro, Optimization of convex risk functions, *Mathematics of Operations Research*, 31, 433–452, 2006.

[39] Ruszczyński, A. and A. Shapiro, Conditional risk mappings, *Mathematics of Operations Research*, 31, 544–561, 2006.

[40] Scandolo, G., *Risk Measures in a Dynamic Setting*, PhD Thesis, Università degli Studi di Milano, Milan, 2003.

[41] Shapiro, A., D. Dentcheva and A. Ruszczyński, *Lectures on Stochastic Programming*, SIAM Publications, Philadelphia 2009.

[42] So, M.M.C. and L.C.. Thomas, Modelling the profitability of credit cards by Markov decision processes, *European Journal of Operational Research*, 212, 123–130, 2011.

[43] Tapiero, C. S. and I. Venezia, A Mean Variance Approach to the Optimal Machine Maintenance and Replacement Problem, *The Journal of the Operational Research Society*, 30, 457–466, 1979.

[44] Veinott, A. F., Discrete dynamic programming with sensitive discount optimality criteria, *Annals of Mathematical Statistics*, 40, 1635–1660, 1969.

[45] White, D. J., A survey of applications of Markov decision processes, *Journal of Operational Research Society*, 44, 1073–1096, 1993.