Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA

Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Computational Methods for Risk-Averse Undiscounted Transient Markov Models

Özlem Çavuş, Andrzej Ruszczyński

To cite this article:
Özlem Çavuş, Andrzej Ruszczyński (2014) Computational Methods for Risk-Averse Undiscounted Transient Markov Models. Operations Research 62(2):401-417. http://dx.doi.org/10.1287/opre.2013.1251

OPERATIONS RESEARCH
Vol. 62, No. 2, March–April 2014, pp. 401–417
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2013.1251

METHODS

Computational Methods for Risk-Averse Undiscounted Transient Markov Models

Özlem Çavuş
Department of Industrial Engineering, Bilkent University, Ankara 06800, Turkey, ozlem.cavus@bilkent.edu.tr

Andrzej Ruszczyński
Department of Management Science and Information Systems, Rutgers University, Piscataway, New Jersey 08854, rusz@rutgers.edu

The total cost problem for discrete-time controlled transient Markov models is considered. The objective functional is a Markov dynamic risk measure of the total cost. Two solution methods, value and policy iteration, are proposed, and their convergence is analyzed. In the policy iteration method, we propose two algorithms for policy evaluation: the nonsmooth Newton method and convex programming, and we prove their convergence. The results are illustrated on a credit limit control problem.

Subject classifications: dynamic programming; risk measures; transient Markov models; value iteration; policy iteration.
Area of review: Optimization.
History: Received October 2012; revisions received April 2013, September 2013; accepted November 2013. Published online in *Articles in Advance* March 31, 2014.

1. Introduction

Rich literature exists on the optimal control problem for transient Markov processes (see Veinott 1969, Pliska 1979, Hernández-Lerma and Lasserre 1999, and references therein). Specific examples of such models are stochastic shortest path problems (see, e.g., Bertsekas and Tsitsiklis 1991) and optimal stopping problems (cf. Çinlar 1975; Dynkin and Yushkevich 1969, 1979; Puterman 1994). Most of this research has focused on the expected total cost model.

A smaller volume of work has addressed risk aversion in such problems. Four main ideas have been explored. The first one is specific for shortest path problems and uses the arrival probability as the objective function (see, e.g., Nie and Wu 2009; Ohtsubo 2003, 2004; Wu and Lin 1999). The second one is based on the use of a utility function at each stage (see Denardo and Rothblum 1979; Jaquette 1973, 1976; Patek 2001). The third idea is to use mean–variance models, at each stage (see Filar and Lee 1985, Filar et al. 1989; for review, see White 1988). The fourth one, initiated by Howard and Matheson (1972), employs a multiplicative entropic cost function, where the expected value of an exponential of the sum of costs is minimized, rather than the expected sum itself. Finite-horizon and infinite-horizon discounted problems as well as average cost problems have been considered (see Bielecki et al. 1999; Cavazos-Cadena and Fernández-Gaucherand 1999; Coraluppi and Marcus 1999, 2000; Di Masi and Stettner 1999; Fernàndez-Gaucherand and Marcus 1997; Fleming and Hernández-Hernández 1997; Hernández-Hernández and Marcus 1996, 1999; Levitt and Ben-Israel 2001; Mannor and Tsitsiklis 2011).

Our research continues earlier efforts to adapt the recent theory of *dynamic risk measures* (see Scandolo 2003; Ruszczyński and Shapiro 2005, 2006b; Cheridito et al. 2006; Artzner et al. 2007; Pflug and Römisch 2007; and references therein) to the Markov setting. Boda and Filar (2006) proved time consistency of the finite-horizon threshold probability criterion, when decision rules are assumed. In the paper by Ruszczyński (2010), a broad class of *Markov risk measures* was defined, and an infinite-horizon discounted cost problem with such risk measures was solved. Decision rules and dynamic programming equations were derived in this approach. An extension of this approach to undiscounted total risk problems for *risk-transient* models was provided by Çavuş and Ruszczyński (2012).

The main objective of the present work is to propose and analyze numerical methods for solving total risk problems with Markov risk measures. Although their appearance resembles the value iteration and policy iteration methods known from expected value models, their analysis requires specific techniques, exploiting properties of Markov risk measures. Some of our ideas are extensions of the techniques employed by Ruszczyński (2010), but the absence of contraction properties precludes their direct application.

In §2, we briefly introduce the relevant terminology and notation of the theory of discrete-time controlled Markov processes. Section 3 is devoted to the definition of the risk-averse control problem for Markov models with randomized policies. In §4, we introduce the class of *risk-transient* models, and we analyze it in the case of finite

state spaces. In §5, we summarize the main findings of Çavuş and Ruszczyński (2012). In §6, we describe and analyze the value iteration method for risk-averse total cost problems. In §7, we present the policy iteration method and we analyze its convergence. Finally, in §8.2, we illustrate the operation of the methods on an example of controlling credit limits.

2. Controlled Markov Processes

We quickly review the main concepts of controlled Markov models and we introduce relevant notation (for details, see Feinberg and Shwartz 2002; Hernández-Lerma and Lasserre 1996, 1999). Let $\mathcal{X}$ be a state space, and let $\mathcal{U}$ a control space. We assume that $\mathcal{X}$ and $\mathcal{U}$ are finite, but a more general setting with Polish spaces equipped with their Borel σ-algebras is possible as well.

A control set is a multifunction $U\colon \mathcal{X} \rightrightarrows \mathcal{U}$; for each state $x \in \mathcal{X}$, the set $U(x) \subseteq \mathcal{U}$ is a nonempty set of possible controls at x. A controlled transition kernel Q is a mapping from the graph of U to the set $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$. We shall write $Q_{xy}(u)$ to denote the transition probability from state x to state y, when control u is applied.

The cost of transition from x to y, when control u is applied, is represented by $c(x,u,y)$, where $c\colon \mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathbb{R}$. Only $u\in U(x)$ and those $y\in\mathcal{X}$ 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 $\mathcal{X}$, a control space $\mathcal{U}$, 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 $\mathcal{H}_t=\operatorname{graph}(U)^{t-1}\times\mathcal{X}$. Each history is a sequence $h_t=(x_1,u_1,\ldots,x_{t-1},u_{t-1},x_t)\in\mathcal{H}_t$.

We denote by $\mathcal{P}(\mathcal{U})$ the set of probability measures on the set $\mathcal{U}$. Likewise, $\mathcal{P}(U(x))$ is the set of probability measures on $U(x)$. A *randomized policy* is a sequence of measurable functions $\pi_t\colon \mathcal{H}_t\to\mathcal{P}(\mathcal{U})$, $t=1,2,\ldots,$ such that $\pi_t(h_t)\in\mathcal{P}(U(x_t))$ for all $h_t\in\mathcal{H}_t$. 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 $\pi_t\colon \mathcal{X}\to\mathcal{P}(\mathcal{U})$, $t=1,2,\ldots,$ such that $\pi_t(x)\in\mathcal{P}(U(x))$ for all $x\in\mathcal{X}$. The function $\pi_t(\cdot)$ is called the *decision rule* at time t. A Markov policy is *stationary* if there exists a function $\pi\colon \mathcal{X}\to\mathcal{P}(\mathcal{U})$ such that $\pi_t(x)=\pi(x)$, for all $t=1,2,\ldots,$ and all $x\in\mathcal{X}$. Such a policy and the corresponding decision rule are called *deterministic*, if for every $x\in\mathcal{X}$ 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* $x_A\in\mathcal{X}$ such that $Q_{x_Ax_A}(u)=1$ and $c(x_A,u,x_A)=0$ for all $u\in U(x_A)$. Thus, after the absorbing state is reached, no further costs are incurred. To analyze such Markov models, it is convenient to consider the effective state space $\widetilde{\mathcal{X}}=\mathcal{X}\setminus\{x_A\}$ and the effective controlled substochastic kernel $\tilde{Q}$, whose arguments are restricted to $\widetilde{\mathcal{X}}$ and whose values are nonnegative measures on $\widetilde{\mathcal{X}}$, so that $\tilde{Q}_{xy}(u)=Q_{xy}(u)$, for all $x,y\in\widetilde{\mathcal{X}}$ and all $u\in U(x)$. In other words, $\tilde{Q}(u)$ is the matrix $Q(u)$ with the row and column corresponding to x_A deleted.

3. Risk-Averse Control Problems

To formally introduce the *total risk problem*, we start from the case of a finite horizon T. Each policy $\Pi=\{\pi_1,\ldots,\pi_T\}$ 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 $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random variables on Ω, $t=2,\ldots,T$. For $t=1$, we set $\mathcal{Z}_1=\mathbb{R}$.

For a policy $\Pi=\{\pi_t\}_{t=1}^T$, a *dynamic measure of risk* is defined as follows:

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

In the formula above, $\rho_t\colon \mathcal{Z}_{t+1}\to\mathcal{Z}_t$, $t=1,\ldots,T$, are *one-step conditional risk measures* satisfying the following axioms:

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

(A2) if $Z\leqslant W$, then $\rho_t(Z)\leqslant\rho_t(W)$, $\forall Z,W\in\mathcal{Z}_{t+1}$;

(A3) $\rho_t(Z+W)=Z+\rho_t(W)$, $\forall Z\in\mathcal{Z}_t$, $W\in\mathcal{Z}_{t+1}$;

(A4) $\rho_t(\beta Z)=\beta\rho_t(Z)$, $\forall Z\in\mathcal{Z}_{t+1}$, $\beta\geqslant 0$.

In Ruszczyński (2010, §3), the nested formulation (1) was derived from general properties of monotonicity and time consistency of dynamic measures of risk. Conditions (A1)–(A4) are analogous to the axioms of *coherent measures of risk*, introduced by Artzner et al. (1999); they are extended to the conditional setting, as in Riedel (2004), Ruszczyński and Shapiro (2006b), Scandolo (2003).

The *infinite-horizon total risk problem* is to find a policy $\Pi=\{\pi_t\}_{t=1}^\infty$ that minimizes the *infinite-horizon dynamic measure of risk*:

$$J_\infty(\Pi,x_1)=\lim_{T\to\infty}J_T(\Pi,x_1).\tag{2}$$

At this moment, we do not know whether the limit (2) is well defined and finite; in §5 we provide sufficient conditions.

As indicated in Ruszczyński (2010), the fundamental difficulty of formulation (1) is that at time t the value of $\rho_t(\cdot)$ is $\mathcal{F}_t$-measurable and is allowed to depend on the entire history h_t of the process. Moreover, in Markov decision processes the probability measure depends on the policy Π, whereas the setting with dynamic measures of risk is formulated for a fixed measure P. To overcome these difficulties, in Ruszczyński (2010, §4), a new construction of a

one-step conditional measure of risk was introduced, which was later extended to the case of randomized policies in Çavuş and Ruszczyński (2012). We outline this construction for the case of finite state and control spaces, which is most relevant for applications.

Given a state x and randomized control λ, a probability measure $\lambda\circ Q(x)$ on the product space $\mathcal{U}\times\mathcal{X}$ is defined as follows:

$$[\lambda\circ Q(x)](u,y)=\lambda(u)Q_{xy}(u). \tag{3}$$

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

$$c_x(u,y)=c(x,u,y),\quad u\in\mathcal{U},\ y\in\mathcal{X}. \tag{4}$$

Let $\mathcal{V}$ be the space of all real functions on $\mathcal{U}\times\mathcal{X}$; it is finite-dimensional. It is convenient to think of the dual space $\mathcal{V}'$ as the space of signed measures m on $\mathcal{U}\times\mathcal{X}$. We consider the set of probability measures in $\mathcal{V}'$:

$$\mathcal{M}=\{m\in\mathcal{V}'\colon\ m(\mathcal{U}\times\mathcal{X})=1,\ m\geqslant 0\}.$$

We use the usual symbol $\langle\cdot,\cdot\rangle$ to denote the scalar product:

$$\langle\varphi,m\rangle=\sum_{u\in\mathcal{U},\,y\in\mathcal{X}}\varphi(u,y)m(u,y),\quad \varphi\in\mathcal{V},\ m\in\mathcal{V}'. \tag{5}$$

Definition 1. A measurable function $\sigma\colon\ \mathcal{V}\times\mathcal{X}\times\mathcal{M}\to\mathbb{R}$ is a *risk transition mapping* if for every $x\in\mathcal{X}$ and every $m\in\mathcal{M}$, the function $\varphi\mapsto\sigma(\varphi,x,m)$ is a coherent measure of risk on $\mathcal{V}$.

Risk transition mappings allow for convenient formulation of risk-averse preferences for controlled Markov processes, where the cost is evaluated by formula (1). Consider a controlled Markov process $\{x_t\}$ with some Markov policy $\Pi=\{\pi_1,\pi_2,\ldots\}$. For a fixed time t and a function $g\colon\ \mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathbb{R}$, the value of $Z_{t+1}=g(x_t,u_t,x_{t+1})$ is a random variable, an element of $\mathcal{Z}_{t+1}$. Let $\rho_t\colon\ \mathcal{Z}_{t+1}\to\mathcal{Z}_t$ be a conditional risk measure satisfying (A1)–(A4). By definition, $\rho_t(g(x_t,u_t,x_{t+1}))$ is an element of $\mathcal{Z}_t$, that is, it is an $\mathcal{F}_t$-measurable function on $(\Omega,\mathcal{F})$. In the definition below, we restrict it to depend on the past only via the current state x_t. We write $g_x\colon\ \mathcal{U}\times\mathcal{X}\to\mathbb{R}$ for the function $g_x(u,y)=g(x,u,y)$. The composition $\pi(x)\circ Q(x)$ is defined as in (3).

Definition 2. A one-step conditional risk measure $\rho_t\colon\ \mathcal{Z}_{t+1}\to\mathcal{Z}_t$ is a *Markov risk measure* with respect to the controlled Markov process $\{x_t\}$, if there exists a risk transition mapping $\sigma_t\colon\ \mathcal{V}\times\mathcal{X}\times\mathcal{M}\to\mathbb{R}$ such that for all w-bounded measurable functions $g\colon\ \mathcal{X}\times\mathcal{U}\times\mathcal{X}\to\mathbb{R}$ and for all feasible decision rules $\pi\colon\ \mathcal{X}\to\mathcal{P}(U)$ we have

$$\rho_t(g(x_t,u_t,x_{t+1}))=\sigma_t(g_{x_t},x_t,\pi(x_t)\circ Q(x_t)),\quad \text{a.s.} \tag{6}$$

The right-hand side of formula (6) is parametrized by x_t, and thus it defines an $\mathcal{F}_t$-measurable random variable, whose dependence on the past is carried only via the state x_t.

4. Risk-Transient Models

In this section, we specify to the case of finite state and control spaces the results of Çavuş and Ruszczyński (2012) concerning the existence of the limit in (2) and the optimality conditions.

Since we require the risk transition mapping, as a function of the first argument, to be coherent and finite valued, it follows that it is continuous with respect to this argument. Therefore, it admits the following dual representation:

$$\sigma(\varphi,x,m)=\max_{\mu\in\mathcal{A}(x,m)}\langle\varphi,\mu\rangle, \tag{7}$$

where $\mathcal{A}(x,m)=\partial_\varphi\sigma(0,x,m)\subset\mathcal{M}$ is convex and closed (see Ruszczyński and Shapiro 2006a and references therein).

Example 1. Based on the first-order mean–semideviation risk measure analyzed by Ogryczak and Ruszczyński (1999, 2001) and Ruszczyński and Shapiro (2006a, Example 4.2; 2006b, Example 6.1), we can define the corresponding risk transition mapping

$$\sigma(\varphi,x,m)=\langle\varphi,m\rangle+\kappa\langle(\varphi-\langle\varphi,m\rangle)_+,m\rangle, \tag{8}$$

with $\kappa\in[0,1]$. Following the derivations of Ruszczyński and Shapiro (2006a, Example 4.2), we have

$$\mathcal{A}(x,m)=\big\{\mu\in\mathcal{M}\colon\ \exists\,(h\in\mathcal{V})\,\mu(u,y)=m(u,y)[1+h(u,y)-\langle h,m\rangle]\ \forall(u,y)\in\mathcal{U}\times\mathcal{X},\|h\|_\infty\leqslant\kappa,h\geqslant 0\big\}. \tag{9}$$

Example 2. Another important example is the average value at risk (see, inter alia, Ogryczak and Ruszczyński 2002, §4; Pflug and Römisch 2007, §§2.2.3, 3.3.4; Rockafellar and Uryasev 2002; Ruszczyński and Shapiro 2006a, Example 4.3; 2006b, Example 6.2), which has the following risk transition counterpart:

$$\sigma(\varphi,x,m)=\inf_{\eta\in\mathbb{R}}\left\{\eta+\frac{1}{\alpha}\langle(\varphi-\eta)_+,m\rangle\right\},\quad \alpha\in(0,1).$$

Following the derivations of Ruszczyński and Shapiro (2006a, Example 4.3), we obtain

$$\mathcal{A}(x,m)=\left\{\mu\in\mathcal{M}\colon\ \mu(u,y)\leqslant\frac{1}{\alpha}m(u,y)\ \forall(u,y)\in\mathcal{U}\times\mathcal{X}\right\}. \tag{10}$$

In the formula (7), the bilinear form is sum over $\mathcal{U}\times\mathcal{X}$. If the function φ depends only on the state, it is sufficient to consider the marginal measure

$$\bar\mu(y)=\mu(\mathcal{U}\times\{y\}),\quad y\in\mathcal{X}. \tag{11}$$

Denote by L the linear operator mapping each $\mu\in\mathcal{V}'$ to the corresponding marginal measure $\bar\mu$ on $\mathcal{X}$, as defined

in (11). For every x we can define the set of probability measures

$$\mathfrak{M}_x^\pi = \{L\mu\colon \mu \in \mathscr{A}(x, \pi(x)\circ Q(x))\}, \quad x \in \mathscr{X}. \tag{12}$$

We call the multifunction $\mathfrak{M}^\pi\colon \mathscr{X} \rightrightarrows \mathscr{P}(\mathscr{X})$, assigning to each $x \in \mathscr{X}$ the set $\mathfrak{M}_x^\pi$, the *risk multikernel*, associated with the risk transition mapping $\sigma(\cdot,\cdot,\cdot)$, the controlled kernel Q, and the decision rule π. Its measurable selectors $M^\pi \ll \mathfrak{M}^\pi$ are transition kernels.

The concept of a risk multikernel is crucial for the analysis of the total risk problems.

Definition 3. We call the Markov model with a risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ and with a stationary Markov policy $\{\pi,\pi,\ldots\}$ *risk transient* if a constant K exists such that

$$\|M\|_\infty \leqslant K \quad \text{for all } M \ll \sum_{j=1}^{T} (\tilde{\mathfrak{M}}^\pi)^j \quad \text{and all } T \geqslant 0. \tag{13}$$

If the estimate (13) is uniform for all Markov policies, the model is called *uniformly risk transient*.

The above property is essential for the finite risk evaluation in an infinite-horizon problem. The following theorem is a special case of Çavuş and Ruszczyński (2012, Theorem 7.1).

Theorem 1. *Suppose a stationary policy $\Pi = \{\pi,\pi,\ldots\}$ is applied to a controlled Markov model with a Markov risk transition mapping $\sigma(\cdot,\cdot,\cdot)$. If the model is risk transient for the policy Π, then the limit* (2) *is finite, and $\|J_\infty(\Pi,\cdot)\|_\infty < \infty$. If the model is uniformly risk transient, then $\|J_\infty(\Pi,\cdot)\|_\infty$ is uniformly bounded. Moreover, for all $x_1 \in \widetilde{\mathscr{X}}$ and any function $f\colon \mathscr{X} \to \mathbb{R}$, we have*

$$J_\infty(\Pi, x_1) = \lim_{T\to\infty} \rho_1\big(c(x_1,u_1,x_2) + \rho_2\big(c(x_2,u_2,x_3) + \cdots + \rho_{T-1}\big(c(x_{T-1},u_{T-1},x_T) + \rho_T(c(x_T,u_T,x_{T+1}) + f(x_{T+1}))\big)\cdots\big)\big).$$

The condition that the model is risk transient is essential, as the following example demonstrates.

Example 3. Consider a transient Markov chain with two states and with the following transition probabilities: $Q_{11} = 1-p$, $Q_{12} = p$, and $Q_{22} = 1$, with $p \in (0,1)$. 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 p. Let $x_1 = 1$. If the limit (2) is finite, then (skipping the dependence on Π) we have

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

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 2, with $0 < \alpha \leqslant 1-p$, to define $\rho_1(\cdot)$. From standard identities for the average value at risk (see, e.g., Shapiro et al. 2009, Theorem 6.2), we deduce that

$$\begin{aligned} J_\infty(1) &= 1 + \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{\alpha}\mathbb{E}[(J_\infty(x_2) - \eta)_+] \right\} \\ &= 1 + \frac{1}{\alpha}\int_{1-\alpha}^{1} F^{-1}(\beta)\, d\beta, \end{aligned} \tag{14}$$

where $F(\cdot)$ is the distribution function of $J_\infty(x_2)$. If $\beta \geqslant p$, all β-quantiles of $J_\infty(x_2)$ are equal to $J_\infty(1)$. Then a contradiction results from the last equation: $J_\infty(1) = 1 + J_\infty(1)$. It follows that a composition of average values at risk has no finite limit, if $0 < \alpha \leqslant 1-p$. On the other hand, if $1-p < \alpha < 1$, then

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

Let us verify condition (13). From (14) we obtain $J_\infty(1) = 1 + ((1-p)/\alpha)J_\infty(1)$, and thus $J_\infty(1) = \alpha/(\alpha-(1-p))$.

From (10) we obtain

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

As only one control is possible, formula (12) simplifies to

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

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

$$\tilde{\mathfrak{M}} = \left[0, \min\left(1, \frac{1-p}{\alpha}\right)\right].$$

For $0 < \alpha \leqslant 1-p$ we can select $\tilde{M} = 1 \in \tilde{\mathfrak{M}}$ to show that $1 \in (\tilde{\mathfrak{M}})^j$ for all j, and thus condition (13) is not satisfied. On the other hand, if $1-p < \alpha \leqslant 1$, then for every $\tilde{M} \in \tilde{\mathfrak{M}}$ we have $0 \leqslant \tilde{M} < 1$, and condition (13) is satisfied.

The next example verifies Definition 3 for the mean–semideviation model of Example 1.

Example 4. For the risk transition mapping of Example 1, we obtain

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

We conclude that $J_\infty(1) = 1/(p - \kappa p(1-p))$ for all $\kappa \in [0,1]$.

Let us verify condition (13). From (9) we obtain

$$\mathscr{A}(i,m)=\{(\mu_1,\mu_2)\colon \mu_j=m_j(1+h_j-(h_1m_1+h_2m_2)),\ 0\leqslant h_j\leqslant\kappa, j=1,2\},$$

$$\mathfrak{M}(i)=\{(\mu_1,\mu_2)\colon \mu_j=Q_{ij}(1+h_j-(h_1Q_{i1}+h_2Q_{i2})),\ 0\leqslant h_j\leqslant\kappa, j=1,2\},\quad i=1,2.$$

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

$$\tilde{\mathfrak{M}}=[(1-p)(1-\kappa p),(1-p)(1+\kappa p)].$$

Definition 3 is satisfied for every $\kappa\in[0,1]$.

A question arises as to whether we can easily verify Definition 3 for a specific transition kernel Q and risk transition mapping $\sigma(\cdot,\cdot,\cdot)$. It is reasonable to assume that in the dual representation (7) we have $m\in\mathscr{A}(x,m)$ for all $m\in\mathscr{M}$ and all $x\in\mathscr{X}$, which is equivalent to

$$\sigma(\varphi,x,m)\geqslant\langle\varphi,m\rangle\quad\forall\varphi\in\mathscr{V},x\in\mathscr{X},m\in\mathscr{M}.$$

Although this property is not implied by the axioms of a coherent measure of risk, it is true for all practically relevant measures of risk, including those of Examples 1 and 2. Then it follows from (12) that $Q\ll\mathfrak{M}$, and thus $\tilde{Q}\ll\tilde{\mathfrak{M}}$ (for simplicity, we skip the superscript π representing the decision rule). Choosing $M=\sum_{j=1}^{T}(\tilde{Q})^j$ in condition (13), we see that a necessary condition for a model to be risk transient is that the series $\sum_{j=1}^{\infty}(\tilde{Q})^j$ is convergent. This holds true if and only if for some finite n we have

$$\|(\tilde{Q})^n\|_\infty<1,\tag{15}$$

that is, if for every state $x\in\widetilde{\mathscr{X}}$ a path to x_A exists in the graph of Q (clearly, the path length n is then smaller than the number of states). The reader may consult, for example, Çinlar (1975, Chapters 5 and 6) for these basic properties of Markov chains. The condition (15), however, is not sufficient, as shown in Example 3. We need to have it satisfied for every selection of $\tilde{\mathfrak{M}}$.

The theorem below provides an easily verifiable sufficient condition for Definition 3. The notation $m\ll\mu$ means that a measure m is absolutely continuous with respect to a measure μ.

THEOREM 2. *Suppose the set of states $\widetilde{\mathscr{X}}$ is transient for a policy $\{\pi,\pi,\ldots\}$. If $m\ll\mu$ for all $\mu\in\mathscr{A}(x,m)$, all $m\in\mathscr{M}$, and all $x\in\widetilde{\mathscr{X}}$, then the model is risk transient.*

PROOF. Let n be such that condition (15) is satisfied. Consider a selector $S\ll(\mathfrak{M}^\pi)^n$. By the definition of the composition of multifunctions, $S=S_1S_2,\ldots,S_n$, with $S_j\ll\mathfrak{M}^\pi$, $j=1,\ldots,n$. Then $S_j=LM_j$, with $M_j(x)\in\mathscr{A}(x,\pi(x)\circ Q(x))$ for all $x\in\mathscr{X}$. By assumption, $\pi(x)\circ Q(x)\ll M_j(x)$ for all j. Therefore,

$$Q^\pi(x)=L(\pi(x)\circ Q(x))\ll L(M_j(x))=S_j(x),\quad j=1,\ldots,n.$$

It follows that the graph of S_j contains all edges of the graph of Q^π, for all $j=1,\ldots,n$. Consequently, the graph representing S contains all edges of the graph of $(Q^\pi)^n$. In particular, for every state x, we have $S_{x,x_A}>0$.

If $x=x_A$, then $\pi(x_A)\circ Q(x_A)$ is a Dirac measure supported at (x_A,u_A). As $\sigma(x,\cdot)$ is a coherent measure of risk, $\mathscr{A}(x_A,\pi(x_A))$ is also a Dirac measure supported at (x_A,u_A). Thus,

$$\mathfrak{M}^\pi(x_A)=L\mathscr{A}(x_A,\pi(x_A)\circ Q(x_A))=\{\delta_{x_A}\}.$$

It follows that every selector S_j has value 1 at the position corresponding to (x_A,x_A). By deleting from S_j the row and column corresponding to x_A, we obtain a selector $\tilde{S}_j\ll\tilde{\mathfrak{M}}^\pi$. Conversely, every selector $\tilde{S}_j\ll\tilde{\mathfrak{M}}^\pi$ can be extended to a selector $S_j\ll\mathfrak{M}^\pi$ by completing every row to 1 and adding a unit row corresponding to x_A. Similar correspondence exists between the products $\tilde{S}=\tilde{S}_1\tilde{S}_2,\ldots,\tilde{S}_n$ and $S=S_1S_2,\ldots,S_n$.

Since $S_{x,x_A}>0$ for all x, we have $\|\tilde{S}\|_\infty<1$. The multikernel $\tilde{\mathfrak{M}}^\pi$ is closed, and thus $\alpha\in[0,1)$ exists such that $\|\tilde{S}\|_\infty<\alpha$ for all $\tilde{S}\ll(\tilde{\mathfrak{M}}^\pi)^n$. We can now apply the last estimate to (13). Every selector

$$M\ll\sum_{j=1}^{T}(\tilde{\mathfrak{M}}^\pi)^j$$

can be written as a sum of selectors:

$$M=\sum_{j=1}^{T}M_j,\quad\text{with } M_j\ll(\tilde{\mathfrak{M}}^\pi)^j.$$

Because $\|M_j\|_\infty\leqslant\alpha^{\lfloor j/n\rfloor}$, we obtain the following uniform bound:

$$\|M\|_\infty\leqslant\sum_{j=1}^{\infty}\alpha^{\lfloor j/n\rfloor}=\frac{n}{1-\alpha}.$$

In the formulas above, $\lfloor c\rfloor$ denotes the integer round down of a real number c. □

The examples below illustrate application of Theorem 2.

EXAMPLE 5. Let us consider the average value at risk from Example 2, but this time combined with the expected value with a coefficient $\kappa\in[0,1)$ as follows:

$$\sigma(\varphi,x,m)=(1-\kappa)\langle\varphi,m\rangle+\kappa\inf_{\eta\in\mathbb{R}}\left\{\eta+\frac{1}{\alpha}\langle(\varphi-\eta)_+,m\rangle\right\},\quad\alpha\in(0,1).\tag{16}$$

Using (10), we can write the subdifferential:

$$\mathscr{A}(x,m)=\partial_\varphi\sigma(0,x,m)=(1-\kappa)m+\kappa\left\{\nu\in\mathscr{M}\colon \nu(u,y)\leqslant\frac{1}{\alpha}m(u,y)\ \forall(u,y)\in\mathscr{U}\times\mathscr{X}\right\}.\tag{17}$$

We immediately see that every $\mu \in \mathscr{A}(x,m)$ satisfies the inequality $\mu \geqslant (1-\kappa)m$ and thus $m \ll \mu$. The sufficient condition of Theorem 2 is satisfied. In particular, for the model discussed in Example 3 with $0 < \alpha \leqslant 1-p$, proceeding similarly to (14), we obtain

$$\begin{aligned} J_\infty(1) &= 1+(1-\kappa)(1-p)J_\infty(1)+\kappa J_\infty(1) \\ &= 1+[1-(1-\kappa)p]J_\infty(1). \end{aligned}$$

If $\kappa \in [0,1)$, this equation has a solution for all $p \in (0,1]$.

Example 6. For the mean–semideviation model of Example 1, we see that every $\mu \in \mathscr{A}(x,m)$ satisfies the relation

$$\mu(u,y) = m(u,y)[1+h(u,y)-\langle h,m\rangle] \quad \forall (u,y) \in \mathscr{U}\times\mathscr{X},$$

with $0 \leqslant h(\cdot,\cdot) \leqslant \kappa$. For any $\kappa \in [0,1]$, the expression in brackets is strictly positive for all (u,y), and thus $m \ll \mu$. The model is risk transient for every transient Markov chain.

5. Dynamic Programming Equations

The main findings of Çavuş and Ruszczyński (2012) substantially simplify in the case of finite state and control spaces. The following theorem is a special case of Çavuş and Ruszczyński (2012, Thorem 7.2).

Theorem 3. *Suppose a controlled Markov model with a Markov risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ is risk transient for the stationary Markov policy $\Pi = \{\pi,\pi,\ldots\}$. Then a function $v\colon \mathscr{X} \to \mathbb{R}$ satisfies the equations*

$$v(x) = \sigma(c_x + v, x, \pi(x) \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}, \tag{18}$$

$$v(x_A) = 0, \tag{19}$$

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

Let $\mathbf{\Pi}$ be the set of all policies. Define the *optimal value function*

$$J^*(x) = \inf_{\Pi \in \mathbf{\Pi}} J_\infty(\Pi, x). \tag{20}$$

The following theorem follows from Çavuş and Ruszczyński (2012, Theorems 8.1, 8.2].

Theorem 4. *Assume that the conditional risk measures ρ_t, $t = 1,\ldots,T$, are Markov and the model is uniformly risk transient. Then a function $v\colon \mathscr{X} \to \mathbb{R}$ satisfies the equations*

$$v(x) = \inf_{\lambda \in \mathscr{P}(U(x))} \sigma(c_x + v, x, \lambda \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}, \tag{21}$$

$$v(x_A) = 0, \tag{22}$$

if and only if $v(x) = J^*(x)$ for all $x \in \mathscr{X}$. Moreover, the minimizer $\pi^*(x)$, $x \in \widetilde{\mathscr{X}}$, on the right-hand side of (21) exists and defines an optimal stationary Markov policy $\Pi^* = \{\pi^*,\pi^*,\ldots\}$ in problem (20).

In the risk-averse case, randomized policies may be strictly superior to deterministic policies. In some cases, however, it is possible to prove that deterministic policies are among the optimal policies. It turns out that we can prove this for the combination of the average value at risk and the expected value from Example 5. Interchanging the calculation of the expected value and the infimum in (16), we obtain the following lower bound:

$$\begin{aligned} &\sigma(\varphi, x, \lambda \circ Q(x)) \\ &\quad = (1-\kappa) \sum_{u \in U(x)} \sum_{y \in \mathscr{X}} \lambda(u) Q_{xy}(u) \varphi(u,y) \\ &\qquad + \kappa \inf_{\eta \in \mathbb{R}} \sum_{u \in U(x)} \sum_{y \in \mathscr{X}} \lambda(u) Q_{xy}(u) \left\{ \eta + \frac{1}{\alpha}(\varphi(u,y)-\eta)_+ \right\} \\ &\quad \geqslant (1-\kappa) \sum_{u \in U(x)} \lambda(u) \sum_{y \in \mathscr{X}} Q_{xy}(u) \varphi(u,y) \\ &\qquad + \kappa \sum_{u \in U(x)} \lambda(u) \inf_{\eta \in \mathbb{R}} \sum_{y \in \mathscr{X}} Q_{xy}(u) \left\{ \eta + \frac{1}{\alpha}(\varphi(u,y)-\eta)_+ \right\}. \end{aligned}$$

The above inequality becomes an equation for every Dirac measure λ. Substituting this expression into the right-hand side of (21) we obtain the following inequality:

$$\begin{aligned} &\inf_{\lambda \in \mathscr{P}(U(x))} \sigma(c_x + v, x, \lambda \circ Q(x)) \\ &\quad \geqslant \inf_{\lambda \in \mathscr{P}(U(x))} \sum_{u \in U(x)} \lambda(u) \inf_{\eta \in \mathbb{R}} \sum_{y \in \mathscr{X}} Q_{xy}(u) \Bigg[(1-\kappa)(c(x,u,y) \\ &\qquad + v(y)) + \kappa \left\{ \eta + \frac{1}{\alpha}(c(x,u,y)+v(y)-\eta)_+ \right\} \Bigg]. \end{aligned}$$

Because the right-hand side achieves its minimum over $\lambda \in \mathscr{P}(U(x))$ at a Dirac measure concentrated at one 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 (16), deterministic Markov policies are optimal.

6. Risk-Averse Value Iteration Method

To find the unique solution J^* of the dynamic programming equations (21) and (22), we adopt and extend the classical value iteration method of Bellman (1957). A similar method has been suggested in Ruszczyński (2010) for risk-averse infinite-horizon discounted models with deterministic policies. We extend it to undiscounted models with randomized policies. This requires different techniques, because the dynamic programming operators do not have the contraction property.

The value iteration method uses Equations (21) and (22) to construct as sequence $\{v^k\}$ of approximations of J^* in the following iterative way:

$$\begin{aligned} v^{k+1}(x) &= \min_{\lambda \in \mathscr{P}(U(x))} \sigma(c_x + v^k, x, \lambda \circ Q(x)), \\ &\qquad x \in \widetilde{\mathscr{X}}, k = 0,1,2,\ldots, \end{aligned} \tag{23}$$

$$v^{k+1}(x_A) = 0, \quad k = 0,1,2,\ldots.$$

We provide the steps of this method in Algorithm 1. The algorithm stops when the successive value functions do not change. However, in practice, an approximate satisfaction of this stopping condition is required.

Algorithm 1 (Risk-averse value iteration)

procedure VALUEITERATION(v^0)
  $k \leftarrow 0$
  repeat
    $k \leftarrow k+1$
    $v^k(x) \leftarrow \min_{\lambda \in \mathcal{P}(U(x))} \sigma(c_x + v^{k-1}, x, \lambda \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}$
    $v^k(x_A) \leftarrow 0$
  until $v^k = v^{k-1}$
  $\pi^*(x) \leftarrow \operatorname{argmin}_{\lambda \in \mathcal{P}(U(x))} \sigma(c_x + v^k, x, \lambda \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}$
  return v^k, π^*
end procedure

We now focus on the convergence of the method. Let us define the operators $\mathfrak{D}\colon \mathscr{V} \to \mathscr{V}$ and $\mathfrak{D}_\pi\colon \mathscr{V} \to \mathscr{V}$ as follows:

$$[\mathfrak{D}v](x) = \min_{\lambda \in \mathcal{P}(U(x))} \sigma(c_x + v, x, \lambda \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}, \tag{24}$$

$$[\mathfrak{D}_\pi v](x) = \sigma(c_x + v, x, \pi(x) \circ Q(x)), \quad x \in \widetilde{\mathscr{X}}, \tag{25}$$

where $\pi(x) \in \mathcal{P}(U(x))$. To prove the convergence, we first provide the following two lemmas similar to Lemmas 1 and 3 in Ruszczyński (2010).

LEMMA 1. *For any φ and ψ in $\mathscr{V}$ such that $\varphi \geqslant \psi$, we have the relations $\mathfrak{D}_\pi \varphi \geqslant \mathfrak{D}_\pi \psi$ and $\mathfrak{D}\varphi \geqslant \mathfrak{D}\psi$.*

PROOF. The proof is similar to the proof of Lemma 1 in Ruszczyński (2010), which we will provide here for completeness. From the dual representation (7), we have

$$[\mathfrak{D}_\pi v](x) = \max_{\mu \in \mathscr{A}(x, \pi(x) \circ Q(x))} \langle c_x + v, \mu \rangle. \tag{26}$$

Since the elements of sets $\mathscr{A}(x, \pi(x) \circ Q(x))$ are just probability measures, $\mathfrak{D}_\pi \varphi \geqslant \mathfrak{D}_\pi \psi$ for $\varphi \geqslant \psi$. Taking the minimum of both sides with respect to π, we also obtain $\mathfrak{D}\varphi \geqslant \mathfrak{D}\psi$. □

LEMMA 2. *Suppose the controlled Markov model is uniformly risk transient. Then, for any function $\varphi\colon \mathscr{X} \to \mathbb{R}$, with $\varphi(x_A) = 0$, the following implications are true*:
(i) *if $\varphi \leqslant \mathfrak{D}\varphi$, then $\varphi \leqslant J^*$;*
(ii) *if $\varphi \geqslant \mathfrak{D}\varphi$, then $\varphi \geqslant J^*$.*

PROOF. (i) If $\varphi \leqslant \mathfrak{D}\varphi$, then for any $\pi \in \mathcal{P}(U)$, we have

$$\varphi \leqslant \mathfrak{D}\varphi \leqslant \mathfrak{D}_\pi \varphi. \tag{27}$$

If we apply the operator $\mathfrak{D}_\pi$ to relation (27), then from the monotonicity property stated in Lemma 1, we obtain the following chain of inequalities:

$$\varphi \leqslant \mathfrak{D}\varphi \leqslant \mathfrak{D}_\pi \varphi \leqslant \mathfrak{D}_\pi \mathfrak{D}\varphi \leqslant [\mathfrak{D}_\pi]^2 \varphi.$$

Proceeding in this way, we get

$$\varphi \leqslant [\mathfrak{D}_\pi]^T \varphi, \quad T = 1, 2, \ldots. \tag{28}$$

Let the Markov policy $\Pi = \{\pi, \pi, \ldots\}$ result in the cost sequence $Z_t = c(x_{t-1}, u_{t-1}, x_t)$, $t = 2, 3, \ldots$. It is clear from Equation (25) that the right-hand side of (28) is equal to the total risk in a finite-horizon problem with the final state cost $v_{T+1} \equiv \varphi$ and with policy $\{\pi, \ldots, \pi\}$. Thus, for every $x_1 \in \widetilde{\mathscr{X}}$, the following inequality is satisfied:

$$\begin{aligned}\varphi(x_1) &\leqslant [[\mathfrak{D}_\pi]^T \varphi](x_1) \\ &= \rho_1\big(c(x_1, u_1, x_2) + \rho_2(c(x_2, u_2, x_3) + \cdots \\ &\qquad + \rho_{T-1}(c(x_{T-1}, u_{T-1}, x_T) + \rho_T(c(x_T, u_T, x_{T+1}) \\ &\qquad\qquad + \varphi(x_{T+1})))\cdots)\big).\end{aligned}$$

Passing to the limit with $T \to \infty$ and using Theorem 1, we conclude that

$$\varphi(x) \leqslant J_\infty(\Pi, x), \quad x \in \mathscr{X}.$$

Since the above inequality holds true for any stationary Markov policy $\Pi = \{\pi, \pi, \ldots\}$, then $\varphi \leqslant J^*$.

(ii) If $\varphi \geqslant \mathfrak{D}\varphi$, then $\pi \in \mathcal{P}(U)$ exists such that

$$\varphi \geqslant \mathfrak{D}_\pi \varphi = \mathfrak{D}\varphi. \tag{29}$$

If we apply the operator $\mathfrak{D}_\pi$ to both sides of the above relation, then from the monotonicity property of the operator $\mathfrak{D}_\pi$ we get

$$\varphi \geqslant [\mathfrak{D}_\pi]^T \varphi, \quad T = 1, 2, \ldots.$$

Similar to the proof of part (i),

$$\begin{aligned}\varphi(x_1) &\geqslant [[\mathfrak{D}_\pi]^T \varphi](x_1) \\ &= \rho_1\big(c(x_1, u_1, x_2) + \rho_2\big(c(x_2, u_2, x_3) + \cdots \\ &\qquad + \rho_{T-1}\big(c(x_{T-1}, u_{T-1}, x_T) + \rho_T(c(x_T, u_T, x_{T+1}) \\ &\qquad\qquad + \varphi(x_{T+1}))\big)\cdots\big)\big).\end{aligned} \tag{30}$$

If we pass to the limit with $T \to \infty$ in (30), again from Theorem 1 we obtain

$$\varphi(x) \geqslant J_\infty(\Pi, x) \geqslant J^*(x), \quad x \in \mathscr{X},$$

as postulated. □

We are now ready to prove the main convergence theorem of this section.

THEOREM 5. *Suppose the assumptions of Theorem 4 are satisfied, and let $v^0 \equiv 0$.*

(i) *If $c(x, u, y) \leqslant 0$ for all $x, y \in \mathscr{X}$ and $u \in U(x)$, then the sequence $\{v^k\}$ obtained by the value iteration method is nonincreasing and convergent to the unique solution J^* of (21) and (22).*

(ii) *If $c(x,u,y)\geqslant 0$ for all $x,y\in\mathscr{X}$ and $u\in U(x)$, and the multifunction $\mathscr{A}(x,\cdot)$ is continuous for all $x\in\mathscr{X}$, then the sequence $\{v^k\}$ is nondecreasing and convergent to J^*.*

Proof. (i) Owing to the monotonicity axiom (A2) and the fact that $c(x,u,y)\leqslant 0$, we obtain $v^0\geqslant\mathfrak{D}v^0$. By virtue of Lemmas 1 and 2,

$$0\geqslant v^k\geqslant v^{k+1}\geqslant J^*,\quad k=0,1,2,\ldots. \tag{31}$$

We have a nonincreasing and bounded sequence that is thus pointwise convergent to some limit $v^\infty\geqslant J^*$. For all $x\in\mathscr{X}$ and all $\lambda\in\mathscr{P}(U(x))$, the function $\sigma(\cdot,x,\lambda\circ Q(x))$, as a finite-valued convex function, is continuous. Let us fix an arbitrary $x\in\mathscr{X}$. Since the function $\sigma(\cdot,x,\lambda\circ Q(x))$ is nondecreasing, we conclude that

$$\sigma(c_x+v^k,x,\lambda\circ Q(x))\downarrow\sigma(c_x+v^\infty,x,\lambda\circ Q(x)),\quad \text{as } k\to\infty,\ \forall\lambda\in\mathscr{P}(U(x)). \tag{32}$$

By the value iteration (23),

$$v^{k+1}(x)\leqslant\sigma(c_x+v^k,x,\lambda\circ Q(x)),\quad\forall\lambda\in\mathscr{P}(U(x)). \tag{33}$$

Passing to the limit with $k\to\infty$ on the left- and right-hand sides of (33) and using (32), we conclude that

$$v^\infty(x)\leqslant\sigma(c_x+v^\infty,x,\lambda\circ Q(x)),\quad\forall\lambda\in\mathscr{P}(U(x)).$$

Because this is true for all $x\in\widetilde{\mathscr{X}}$ and all $\lambda\in\mathscr{P}(U(x))$, it follows that

$$v^\infty\leqslant\mathfrak{D}v^\infty.$$

By Lemma 2, $v^\infty\leqslant J^*$, and thus $v^\infty=J^*$, which completes the proof in this case.

(ii) Owing to the monotonicity axiom (A2) and the fact that $c(x,u,y)\geqslant 0$, proceeding similarly to case (i), we conclude that

$$v^k\uparrow v^\infty\leqslant J^*,\quad\text{as } k\to\infty. \tag{34}$$

Since the multifunction $\mathscr{A}(x,\cdot)$ is continuous, the mapping $(v,\lambda)\mapsto\sigma(c_x+v,x,\lambda\circ Q(x))$ is also continuous (see, e.g., Aubin and Frankowska 1990, Theorem 1.4.16). By the same token, the mapping

$$v\mapsto\min_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v,x,\lambda\circ Q(x))$$

is continuous as well. It follows that for all $x\in\mathscr{X}$,

$$\begin{aligned}v^\infty(x)&=\lim_{k\to\infty}v^{k+1}(x)=\lim_{k\to\infty}\min_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v^k,x,\lambda\circ Q(x))\\&=\min_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v^\infty,x,\lambda\circ Q(x)).\end{aligned}$$

Thus $v^\infty=\mathfrak{D}v^\infty$, as postulated. □

The assumption of all nonnegative or all nonpositive costs corresponds to similar conditions in risk-neutral models (see, e.g., Puterman 1994, Chapter 7). In our case, however, due to the nonlinearity of the risk mappings, stronger assumptions are required in case (ii).

7. Risk-Averse Policy Iteration Method

7.1. The Method

As an alternative way to solve the dynamic programming equations (21) and (22), we suggest a risk-averse policy iteration method that is analogous to the classical policy iteration method of Howard (1960). A similar approach was proposed in Ruszczyński (2010) for risk-averse *discounted* infinite-horizon problems with the feasible set being restricted to deterministic policies.

At iteration k of the method, for a stationary policy $\Pi^k=\{\pi^k,\pi^k,\ldots\}$, the *policy evaluation step* solves the following system of equations to find $J_\infty(\Pi^k,x)=v^k(x)$, $x\in\mathscr{X}$:

$$v(x)=\sigma(c_x+v,x,\pi^k(x)\circ Q(x)),\quad x\in\widetilde{\mathscr{X}}, \tag{35}$$

$$v(x_A)=0. \tag{36}$$

Then the *policy improvement step* finds a new decision rule π^{k+1} if it gives an improved value function:

$$\pi^{k+1}(x)\leftarrow\operatorname*{argmin}_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v^k,x,\lambda\circ Q(x)),\quad x\in\widetilde{\mathscr{X}}. \tag{37}$$

These steps are repeated until the value function does not change. The operation of the method is presented in Algorithm 2.

Algorithm 2 (Risk-averse policy iteration)

procedure PolicyIteration(π^0)
  $k\leftarrow 0$
  repeat
    Policy Evaluation Step:
    $v(x_A)\leftarrow 0$
    Solve the equation $v(x)=\sigma(c_x+v,x,\pi^k(x)\circ Q(x))$, $x\in\widetilde{\mathscr{X}}$
    $v^k\leftarrow v$
    Policy Improvement Step:
    $\bar{v}(x_A)\leftarrow 0$
    $\bar{v}(x)\leftarrow\min_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v^k,x,\lambda\circ Q(x))$, $x\in\widetilde{\mathscr{X}}$
    for $x\in\widetilde{\mathscr{X}}$ **do**
      if $\bar{v}(x)<v^k(x)$ **then**
        $\pi^{k+1}(x)\leftarrow\operatorname*{argmin}_{\lambda\in\mathscr{P}(U(x))}\sigma(c_x+v^k,x,\lambda\circ Q(x))$
      else
        $\pi^{k+1}(x)\leftarrow\pi^k(x)$
      end if
    end for
    $k\leftarrow k+1$
  until $\bar{v}=v^{k-1}$
  return $\bar{v}$, π^k
end procedure

7.2. Convergence

Let the operators $\mathfrak{D}$ and $\mathfrak{D}_\pi$ be defined as (24) and (25), respectively. Then (35) can be equivalently written as follows:

$$v^k=\mathfrak{D}_{\pi^k}v^k. \tag{38}$$

Similarly, (37) is equivalent to the equation

$$\mathfrak{D}_{\pi^{k+1}}v^k=\mathfrak{D}v^k. \tag{39}$$

THEOREM 6. *Suppose the assumptions of Theorem* 4 *are satisfied. Then for any* π^0 *such that* $\pi^0(x)\in\mathscr{P}(U(x))$, $x\in\mathscr{X}$, *the sequence* $\{v^k\}$ *obtained by the policy iteration method is nonincreasing and pointwise convergent to the unique solution* J^* *of* (21) *and* (22).

PROOF. Using Equations (38) and (39), we obtain

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

Applying the operator $\mathfrak{D}_{\pi^{k+1}}$ to above relation, from the monotonicity property given in Lemma 1 we deduce that

$$[\mathfrak{D}_{\pi^{k+1}}]^T v^k\leqslant\mathfrak{D}_{\pi^{k+1}}v^k=\mathfrak{D}v^k\leqslant v^k,\quad T=1,2,\ldots. \tag{40}$$

Relation (40) can be equivalently written as

$$\rho_1\big(c(x_1,u_1,x_2)+\rho_2(c(x_2,u_2,x_3)+\cdots+$$
$$\rho_T(c(x_T,u_T,x_{T+1})+v^k(x_{T+1}))\cdots)\big)\leqslant[\mathfrak{D}v^k](x_1)\leqslant v^k(x_1),$$

where $c(x_{t-1},u_{t-1},x_t)$, $t=2,3,\ldots,T+1$, is the cost sequence resulting from the policy $\Pi^{k+1}=\{\pi^{k+1},\pi^{k+1},\ldots,\pi^{k+1}\}$. Passing to the limit with $T\to\infty$, from Theorems 1 and 3 we conclude that the sequence $\{v^k\}$ is nonincreasing:

$$v^{k+1}(x)=J_\infty(\Pi^{k+1},x)\leqslant[\mathfrak{D}v^k](x)\leqslant v^k(x),\quad x\in\widetilde{\mathscr{X}},\ k=0,1,2,\ldots. \tag{41}$$

Since $v^k\geqslant J^*$, the sequence $\{v^k\}$ is monotonically convergent to some limit $v^\infty\geqslant J^*$. The function $\sigma(\cdot,x,\lambda\circ Q(x))$ is nondecreasing, and thus

$$\sigma(c_x+v^k,x,\lambda\circ Q(x))\downarrow\sigma(c_x+v^\infty,x,\lambda\circ Q(x)),\quad \text{as } k\to\infty,\ \forall\lambda\in\mathscr{P}(U(x)). \tag{42}$$

The left inequality in (41) also implies that

$$v^{k+1}(x)\leqslant\sigma(c_x+v^k,x,\lambda\circ Q(x)),\quad\forall\lambda\in\mathscr{P}(U(x)). \tag{43}$$

Passing to the limit with $k\to\infty$ on both sides of (43) and using (42), we conclude that

$$v^\infty(x)\leqslant\sigma(c_x+v^\infty,x,\lambda\circ Q(x)),\quad\forall\lambda\in\mathscr{P}(U(x)).$$

Because this is true for all $x\in\widetilde{\mathscr{X}}$ and all $\lambda\in\mathscr{P}(U(x))$, it follows that

$$v^\infty\leqslant\mathfrak{D}v^\infty.$$

By Lemma 2, $v^\infty\leqslant J^*$, and thus $v^\infty=J^*$. □

Observe that the convergence of the policy iteration method is not dependent on the cost function being nonnegative or nonpositive.

7.3. Specialized Nonsmooth Newton Method

In the evaluation step of the policy iteration method, we have to solve a system of nonlinear equations (35), which is nonsmooth for all risk mappings, except for the expected value mapping. To solve this system of equations, we adopt the specialized nonsmooth Newton method of Ruszczyński (2010), which uses the idea of the nonsmooth Newton method with linear auxiliary problems (for details, see Klatte and Kummer 2002, §10.1; Kummer 1988).

To find the unique solution of (35) with $v(x_A)=0$, we will solve iteratively an appropriate linear approximation of this system. Using the dual representation (7), the equation (35) can be equivalently written as follows:

$$v(x)=\max_{\mu\in\mathscr{A}(x,\pi^k(x)\circ Q(x))}\sum_{y\in\mathscr{X}}\sum_{u\in U(x)}[c(x,u,y)+v(y)]\mu(u,y),\quad x\in\widetilde{\mathscr{X}}. \tag{44}$$

Let v_l^k be an approximation of the solution of (44) at iteration l of the nonsmooth Newton method. In the description of the method, for simplicity of notation, we omit the index k, which remains fixed throughout the iterations. We find

$$M_l(\cdot\,|\,x)\in\operatorname*{argmax}_{\mu\in\mathscr{A}(x,\pi^k(x)\circ Q(x))}\sum_{y\in\mathscr{X}}\sum_{u\in U(x)}[c(x,u,y)+v_l(y)]\mu(u,y),\quad x\in\widetilde{\mathscr{X}}. \tag{45}$$

The maximum in Equation (45) is attained because the set $\mathscr{A}$ is bounded, convex, and closed, and the function being maximized is linear. Substituting M_l into (44), we obtain the following *linear* equation:

$$v(x)=\sum_{y\in\mathscr{X}}\sum_{u\in U(x)}[c(x,u,y)+v(y)]M_l(u,y\,|\,x),\quad x\in\widetilde{\mathscr{X}}. \tag{46}$$

The solution of this equation is our next approximation v_{l+1}, and the iteration continues.

We will show that the sequence $\{v_l\}$ obtained by this method converges to the unique solution of (35). At first, we need to provide some technical results.

Let us define the operator $\mathfrak{R}_l$ as follows:

$$[\mathfrak{R}_l v](x)=\sum_{y\in\mathscr{X}}\sum_{u\in U(x)}[c(x,u,y)+v(y)]M_l(u,y\,|\,x),\quad x\in\widetilde{\mathscr{X}}.$$

It is clear that the equation (46) can be equivalently written as $v=\mathfrak{R}_l v$.

LEMMA 3. *For any function* ψ^0 *on* $\mathscr{X}$, *with* $\psi^0(x_A)=0$, *the sequence*

$$\psi^{k+1}=\mathfrak{R}_l\psi^k,\quad k=0,1,2,\ldots, \tag{47}$$

is convergent to the unique solution of Equation (46).

Proof. Define $\delta^k = \psi^{k+1} - \psi^k$. It follows from (47) that

$$\delta^{k+1} = M_l \delta^k, \quad k = 0, 1, 2, \ldots.$$

Because each δ^k is a function of x only, we may consider the marginal measures

$$\tilde{M}_l(B \mid x) = M_l(\mathcal{U} \times B \mid x), \quad B \in \mathcal{B}(\widetilde{\mathcal{X}}).$$

Moreover, $\psi^k(x_A) = 0$, and we may restrict our considerations to functions on the effective state space $\widetilde{\mathcal{X}}$. We obtain

$$\delta^{k+1} = \tilde{M}_l \delta^k, \quad k = 0, 1, 2, \ldots.$$

Consequently,

$$\psi^{k+1} = \psi^0 + \sum_{j=0}^{k} \delta^j = \psi^0 + \sum_{j=0}^{k} (\tilde{M}_l)^j \delta^0. \tag{48}$$

By assumption, the model is risk transient, and $\tilde{M}_l$ is a measurable selector of the risk multikernel $\tilde{\mathfrak{M}}^{\pi^k}$. It follows from (13) that

$$\left\| \sum_{j=0}^{\infty} (\tilde{M}_l)^j \delta^0 \right\| \leqslant \sum_{j=0}^{\infty} \|(\tilde{M}_l)^j\| \, \|\delta^0\| < \infty.$$

Consequently, the series (48) is convergent to some limit ψ^∞. The affine operator $\mathfrak{R}_l$ is continuous, and thus passing to the limit in (47) we conclude that ψ^∞ satisfies Equation (46). If another solution φ to this equation existed, then their difference $\delta = \psi^\infty - \varphi$ would satisfy the equation

$$\delta = \tilde{M}_l \delta.$$

Iterating, we conclude that

$$\delta = (\tilde{M}_l)^k \delta, \quad k = 1, 2, \ldots.$$

By (13), the right-hand side converges to 0, as $k \to \infty$, and thus $\delta = 0$. □

We are now ready to prove convergence of the Newton method.

Theorem 7. *For any initial v_0, the sequence $\{v_l\}$ obtained by the Newton method is nondecreasing and convergent to the unique solution v^* of* (35).

Proof. By definition, for all v we have

$$\mathfrak{R}_l v \leqslant \mathfrak{D}_{\pi^k} v. \tag{49}$$

The operator $\mathfrak{R}_l$ is monotone owing to the fact that $M_l(\cdot \mid x)$, $x \in \mathcal{X}$, are probability measures. Therefore, if we apply the operator $\mathfrak{R}_l$ to inequality (49), and use (49) again, we obtain

$$[\mathfrak{R}_l]^2 v \leqslant \mathfrak{R}_l \mathfrak{D}_{\pi^k} v \leqslant [\mathfrak{D}_{\pi^k}]^2 v.$$

Iterating in this way, we get

$$[\mathfrak{R}_l]^T v \leqslant [\mathfrak{D}_{\pi^k}]^T v, \quad T = 1, 2, \ldots. \tag{50}$$

Passing to the limit with $T \to \infty$, from Lemma 3 we deduce that the left-hand side of (50) converges to v_{l+1}. Moreover, the right-hand side converges to the unique solution $\hat{v}$ of (44). Therefore, we get that $v_{l+1} \leqslant \hat{v}$, and thus the sequence $\{v_{l+1}\}$ is bounded from above. We will show that it is also nondecreasing.

For every $x \in \mathcal{X}$, we have

$$\begin{aligned} v_l(x) &= \sum_{y \in \mathcal{X}} \sum_{u \in U(x)} [c(x,u,y) + v_l(y)] M_{l-1}(u, y \mid x) \\ &\leqslant \max_{\mu \in \mathcal{A}(x, \pi^k(x) \circ Q(x))} \sum_{y \in \mathcal{X}} \sum_{u \in U(x)} [c(x,u,y) + v_l(y)] \mu(u,y) \\ &= \sum_{y \in \mathcal{X}} \sum_{u \in U(x)} [c(x,u,y) + v_l(y)] M_l(u, y \mid x) \\ &= [\mathfrak{D}_{\pi^k} v_l](x) = [\mathfrak{R}_l v_l](x). \end{aligned}$$

If we apply $\mathfrak{R}_l$ to above relation, owing to its monotonicity property, we obtain

$$v_l \leqslant \mathfrak{D}_{\pi^k} v_l \leqslant [\mathfrak{R}_l]^T v_l, \quad T = 1, 2, \ldots. \tag{51}$$

The right-hand side converges to v_{l+1}, as $T \to \infty$. Therefore,

$$v_l \leqslant \mathfrak{D}_{\pi^k} v_l \leqslant v_{l+1}, \tag{52}$$

and the sequence $\{v_l\}$ is nondecreasing. Since it is also bounded from above, it has some limit v^∞. Passing to the limit with $l \to \infty$ in (52), we obtain $v^\infty = \mathfrak{D}_{\pi^k} v^\infty$, and thus v^∞ is the unique solution of (35). □

7.4. Policy Evaluation by Convex Optimization

An alternative way to solve the policy evaluation equations (35) and (36) is to formulate and solve the following equivalent convex optimization problem:

$$\min \quad \sum_{x \in \mathcal{X}} v(x) \tag{53}$$

$$\text{s.t.} \quad v(x) \geqslant \sigma(c_x + v, x, \pi^k(x) \circ Q(x)), \quad x \in \widetilde{\mathcal{X}}, \tag{54}$$

$$v(x_A) = 0. \tag{55}$$

Since the risk transition mapping $\sigma(\cdot, x, \pi^k(x) \circ Q(x))$ is convex with respect to the first argument for all $x \in \widetilde{\mathcal{X}}$, the constraint (54) is convex.

Theorem 8. *Suppose the assumptions of Theorem* 3 *are satisfied. Then the solution of problem* (53)–(55) *is equal to $J_\infty(\Pi^k, \cdot)$.*

PROOF. By Theorem 3, the value function $J_\infty(\Pi^k,\cdot)$, which is the unique solution of the system (18)–(19), satisfies (54)–(55). Suppose the decision rule π^k is the only feasible decision rule in the problem. Then every feasible solution v of problem (53)–(55) satisfies (54), which can be written as $v \geqslant \mathfrak{D}v$. By virtue of Lemma 2(ii), $v(\cdot) \geqslant J_\infty(\Pi^k,\cdot)$. Therefore, $J_\infty(\Pi^k,\cdot)$ is an optimal solution of problem (53)–(55). Any other optimal solution $\bar{v}$ satisfies the inequality $\bar{v}(\cdot) \geqslant J_\infty(\Pi^k,\cdot)$ and the equation

$$\sum_{x\in\mathscr{X}} \bar{v}(x) = \sum_{x\in\mathscr{X}} J_\infty(\Pi^k,x).$$

It must, therefore, coincide with $J_\infty(\Pi^k,\cdot)$. □

The specialized Newton method discussed in §7.3 can be interpreted as a constraint linearization method for problem (53)–(55). We can also employ other methods of convex programming to this problem, in particular, exploiting the dual representation (7).

8. Numerical Illustration

8.1. Credit Card Problem

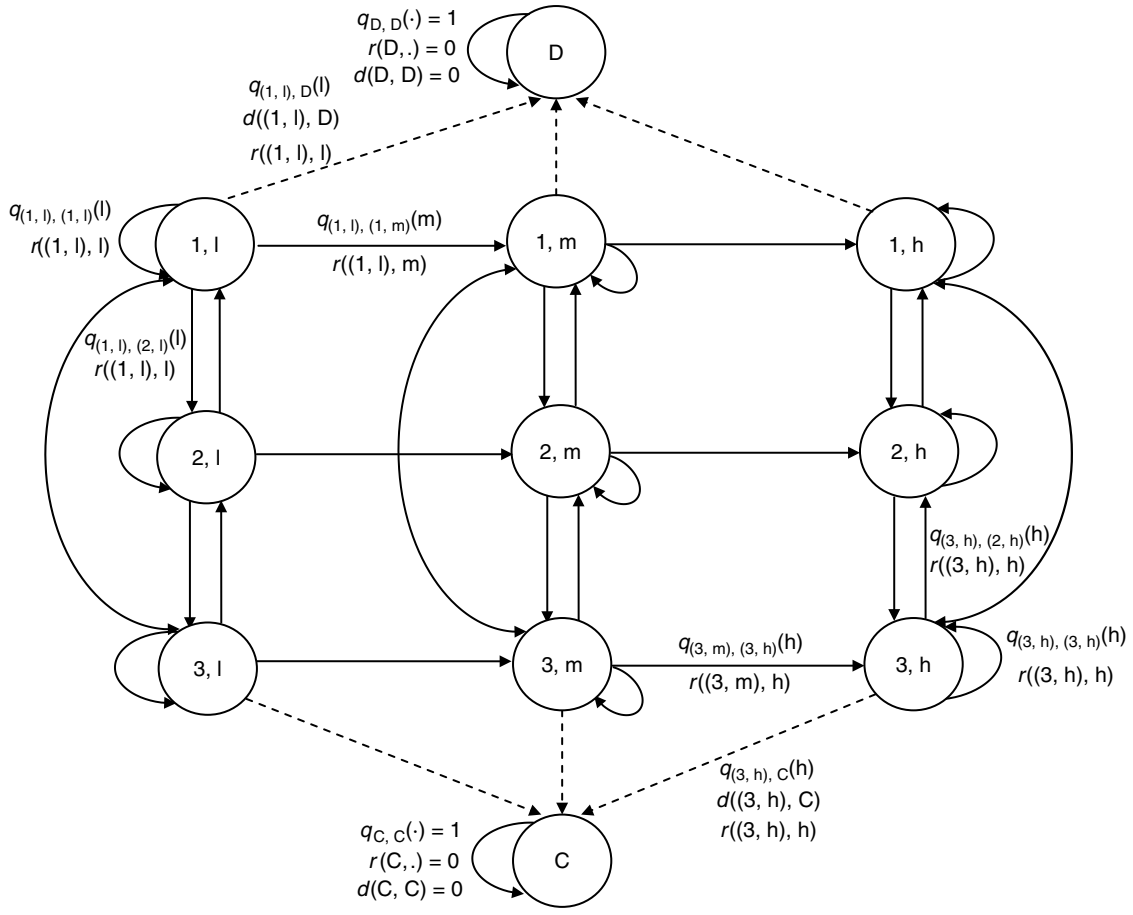
In this section, we illustrate our results on a simplified and modified version of the credit card example discussed by So and Thomas (2011). We use a discrete-time, absorbing Markov decision chain illustrated in Figure 1.

The states of the system are denoted by (i,j), $i=1,2,3$, $j=$"l","m","h", where i represents the type of the customer, and j is the credit limit given. We consider three customer types with $i=1$ representing a customer who does not pay the debt in a timely manner, type $i=3$ representing a responsible customer, and type $i=2$ an intermediate level customer. There are three credit limits: "low" (denoted by "l"), "medium" (denoted by "m"), and "high" (denoted by "h"). The state space includes two additional states "account closure" (denoted by "C'') and "default" (denoted by "D''), both of which are absorbing states.

Following So and Thomas (2011), we do not consider decreasing the credit limit at any of the states. Two controls are possible for states (i,l), $i=1,2,3$, either to keep the credit limit unchanged (represented by "l") or increase it to the medium limit (represented by "m"). Similarly, for states (i,m), $i=1,2,3$, the admissible controls are "m" and "h." The states (i,h), $i=1,2,3$ have one possible control: keep the credit limit at the high level (represented by "h"). There is only one formal control "Continue" at the absorbing states C and D.

The decision to keep the credit limit unchanged results in a transition to the same state, or to a state with a different

Figure 1. The credit card model.



customer type but the same credit limit, or to one of the absorbing states C and D. For example, under the control "m," the possible transitions from the state $(2,\mathrm{m})$ are to the states $(1,\mathrm{m})$, $(2,\mathrm{m})$, $(3,\mathrm{m})$, C, and D. If it is decided to increase the credit limit, then with probability one a transition is made to a new state with the same customer type as the current state, but with the higher credit limit. For example, if the credit limit is increased to "h" at state $(2,\mathrm{m})$, then a transition to state $(2,\mathrm{h})$ will occur with probability one.

The rewards are the profits obtained at each time step. We consider two different profit values: the first one, denoted by $r(x,u)$, $x\in\mathscr{X}$, $u\in U(x)$, is the profit obtained at state x under the control u, and the second one, $d(x,y)$, $x\in\mathscr{X}$, $y\in\mathscr{X}$, is the profit collected from the transition from state x to state y. We assume that $r(x,u)=0$, $x\in\{\mathrm{C},\mathrm{D}\}$, $u\in U(x)$, and $d(\mathrm{C},\mathrm{C})=0$, $d(\mathrm{D},\mathrm{D})=0$.

The objective is to maximize the one-time profit one would be willing to collect at time zero instead of a random sequence of future profits. To apply our theory, we will work with the negatives of profit values and their present time equivalents represented by measures of risk. The corresponding minimization problem of a dynamic measure of risk will be solved. We assume that feasible policies are limited to deterministic ones, and we use the first-order mean–semideviation (see Equation (8)) as the risk measure. Then, the dynamic programming Equation (21) takes on the following form:

$$v(x)=\min_{u\in U(x)}\Bigg\{\underbrace{\sum_{y\in\mathscr{X}}(v(y)-r(x,u)-d(x,y))q_{x,y}(u)}_{\text{expected value }\psi}+\kappa\underbrace{\sum_{z\in\mathscr{X}}(v(z)-r(x,u)-d(x,z)-\psi)_+q_{x,z}(u)}_{\text{semideviation}}\Bigg\},\quad x\in\widetilde{\mathscr{X}},\qquad(56)$$

where $q_{x,y}(u)$ is the probability of making a transition to state $y\in\mathscr{X}$ from $x\in\mathscr{X}$ under the control $u\in U(x)$. Using the fact that $\sum_{y\in\mathscr{X}}r(x,u)q_{x,y}(u)=r(x,u)$, we can rewrite (56) as follows:

$$v(x)=\min_{u\in U(x)}\Bigg\{-r(x,u)+\underbrace{\sum_{y\in\mathscr{X}}(v(y)-d(x,y))q_{x,y}(u)}_{\bar{\psi}}+\kappa\sum_{z\in\mathscr{X}}(v(z)-d(x,z)-\bar{\psi})_+q_{x,z}(u)\Bigg\},\quad x\in\widetilde{\mathscr{X}}.\qquad(57)$$

We use both value and policy iteration methods to solve the dynamic programming Equation (57) with $v(\mathrm{C})=0$ and $v(\mathrm{D})=0$. As explained in §6, value iteration is just the iteration of Equation (57).

To find the unique solution of the nonsmooth equation system appearing in the policy evaluation step of the policy iteration algorithm (see Algorithm 2), we apply Newton's method of §7.3 and the convex optimization method of §7.4.

To calculate M_{l+1} at iteration $l+1$ of Newton's method, we solve the following optimization problem for all $x\in\mathscr{X}$:

$$\begin{aligned}\max_{\mu,h}\ &\sum_{y\in\mathscr{X}}(v_l(y)-r(x,\pi^k(x))-d(x,y))\mu(y)\\ \text{s.t. }&\mu(y)=q_{x,y}(\pi^k(x))\Bigg(1+h(y)-\sum_{z\in\mathscr{X}}h(z)q_{x,z}(\pi^k(x))\Bigg),\quad y\in\mathscr{X},\\ &\sum_{y\in\mathscr{X}}\mu(y)=1,\\ &h(y)\leqslant\kappa,\quad y\in\mathscr{X},\\ &\mu(y),h(y)\geqslant 0,\quad y\in\mathscr{X},\end{aligned}$$

where $\pi^k(x)\in U(x), x\in\mathscr{X}$ is the decision rule at iteration k of the policy iteration algorithm. Then, v_{l+1} is calculated by solving the following system of linear equations:

$$v(x)=\sum_{y\in\mathscr{X}}(v(y)-r(x,\pi^k(x))-d(x,y))\mu(y),\quad x\in\widetilde{\mathscr{X}},$$
$$v(\mathrm{D})=0,\quad v(\mathrm{C})=0.$$

The convex optimization problem (53)–(55) with first-order mean–semideviation risk measure has the following form:

$$\begin{aligned}\min_{v,\psi,\varphi}\ &\sum_{x\in\mathscr{X}}v(x)\\ \text{s.t. }&\psi(x)=\sum_{y\in\mathscr{X}}(v(y)-r(x,\pi^k(x))-d(x,y))q_{x,y}(\pi^k(x)),\quad x\in\widetilde{\mathscr{X}},\\ &v(x)\geqslant\psi(x)+\kappa\sum_{y\in\mathscr{X}}\varphi(x,y)q_{x,y}(\pi^k(x)),\quad x\in\widetilde{\mathscr{X}},\\ &\varphi(x,y)\geqslant v(y)-r(x,\pi^k(x))-d(x,y)-\psi(x),\quad x\in\widetilde{\mathscr{X}},\ y\in\mathscr{X},\\ &\varphi(x,y)\geqslant 0,\quad x\in\widetilde{\mathscr{X}},\ y\in\mathscr{X},\\ &v(x_A)=0.\end{aligned}$$

In this problem, $\psi(x)$ represents the expected value of one-step risk accumulation at state x, and $\varphi(x,y)$ is the upper semideviation in the case where transition is made to state y. Because we are using the first-order mean–semideviation, the problem is in fact linear.

8.2. Numerical Results

For numerical illustration, we used the transition probabilities given in Table 1 with "—" signs indicating transition probabilities equal to zero.

Table 1. Transition probabilities.

Limit	State	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)	C	D
l	(1, l)	0.84	—	—	0.120	—	—	0.01	—	—	0.001	0.029
	(1, m)	—	—	—	—	—	—	—	—	—	—	—
	(1, h)	—	—	—	—	—	—	—	—	—	—	—
	(2, l)	0.040	—	—	0.739	—	—	0.200	—	—	0.011	0.010
	(2, m)	—	—	—	—	—	—	—	—	—	—	—
	(2, h)	—	—	—	—	—	—	—	—	—	—	—
	(3, l)	0.004	—	—	0.010	—	—	0.963	—	—	0.020	0.003
	(3, m)	—	—	—	—	—	—	—	—	—	—	—
	(3, h)	—	—	—	—	—	—	—	—	—	—	—
m	(1, l)	—	1	—	—	—	—	—	—	—	—	—
	(1, m)	—	0.835	—	—	0.100	—	—	0.005	—	0.005	0.055
	(1, h)	—	—	—	—	—	—	—	—	—	—	—
	(2, l)	—	—	—	—	1	—	—	—	—	—	—
	(2, m)	—	0.049	—	—	0.860	—	—	0.073	—	0.002	0.016
	(2, h)	—	—	—	—	—	—	—	—	—	—	—
	(3, l)	—	—	—	—	—	—	—	1	—	—	—
	(3, m)	—	0.006	—	—	0.070	—	—	0.914	—	0.004	0.006
	(3, h)	—	—	—	—	—	—	—	—	—	—	—
h	(1, l)	—	—	—	—	—	—	—	—	—	—	—
	(1, m)	—	—	1	—	—	—	—	—	—	—	—
	(1, h)	—	—	0.829	—	—	0.060	—	—	0.001	0.010	0.100
	(2, l)	—	—	—	—	—	—	—	—	—	—	—
	(2, m)	—	—	—	—	—	1	—	—	—	—	—
	(2, h)	—	—	0.055	—	—	0.858	—	—	0.060	0.001	0.026
	(3, l)	—	—	—	—	—	—	—	—	—	—	—
	(3, m)	—	—	—	—	—	—	—	—	1	—	—
	(3, h)	—	—	0.009	—	—	0.079	—	—	0.900	0.002	0.010

State and control dependent profit values $r(x,u)$, $x \in \mathscr{X}$, $u \in U(x)$, are provided in Table 2, and the transition profits $d(x,y)$, $x \in \mathscr{X}$, $y \in \mathscr{X}$, are given in Table 3. The empty cells in Table 2 mean that the corresponding state–control pairs are inadmissible. The "—" signs in Table 3 mean that corresponding transition profits are zero. All data used in this example are not real and do not correspond to a real case, but they are determined on the basis of partial information provided by So and Thomas (2011).

Table 2. Profit values for state and control pairs.

	State								
Limit	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)
l	270			18			−10		
m	344	300		47	30		5	4	
h		2,240	1,920		650	560		90	80

We solved two different problems for this example. In the first problem, we assumed that the decision makers, namely, creditors, are risk neutral. In the second problem, we considered risk-averse decision makers. Since, in general, the operator $\mathfrak{D}\colon \mathscr{V} \to \mathscr{V}$ (see (24)) will be nonlinear, we did not allow randomized policies for the risk-averse case of this example, and we limited feasible policies to deterministic ones.

The optimal policies and values of the expected value (risk-neutral) problem are given in Table 4. Here, the optimal value function is the negative of the expected total profit function earned under the optimal policy.

We modeled the risk-averse problem using the first-order mean–semideviation as the risk measure and solved it with

Table 3. Transition profits.

State	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)	C	D
(1, l)	—	—	—	—	—	—	—	—	—	40	−550
(1, m)	—	—	—	—	—	—	—	—	—	100	−3,700
(1, h)	—	—	—	—	—	—	—	—	—	1,000	−15,000
(2, l)	—	—	—	—	—	—	—	—	—	18	−400
(2, m)	—	—	—	—	—	—	—	—	—	30	−2,500
(2, h)	—	—	—	—	—	—	—	—	—	500	−10,000
(3, l)	—	—	—	—	—	—	—	—	—	5	−250
(3, m)	—	—	—	—	—	—	—	—	—	15	−1,250
(3, h)	—	—	—	—	—	—	—	—	—	300	−4,500

Table 4. Optimal values and policy for the expected value problem.

State	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)
Values $v(\cdot)$	−7,407.60	−7,063.60	−4,823.60	−7,179.09	−7,132.09	−6,482.09	−6,262.99	−6,257.99	−5,910.98
Policy	m	h	h	m	h	h	m	m	h

different values of the parameter κ. Optimal policies and values have been calculated using the two iterative methods presented in this paper. The algorithms have been coded in MATLAB R2011b and the MOSEK optimization toolbox for MATLAB (see MOSEK 2012) has been integrated. All numerical experiments have been carried out on a PC with an Intel Core i7-2620M 2.70 GHz processor and 6 GB of RAM.

The convergence of the value iteration method is proved in Theorem 5 for problems with all nonpositive or nonnegative cost values. In this example, the profit values are not restricted to being all nonnegative or nonpositive; therefore, Theorem 5 does not apply here. However, using Lemma 2, we can state that if at any iteration k of the value iteration method the value function v^k satisfies the relation $v^k \leqslant \mathfrak{D}v^k = v^{k+1}$, then (using an argument similar to the proof of Theorem 5) the remaining sequence obtained by the value iteration method will be nondecreasing and convergent to the optimal value function J^*. Similarly, if $v^k \geqslant \mathfrak{D}v^k = v^{k+1}$, a nonincreasing remaining sequence converging to J^* is generated. For this example, the initial value function was set to zero, $v^0 \equiv 0$, for the value iteration method. We observed that even when the sequence was not monotonic at initial iterations of the value iteration algorithm, it became monotonic very soon, which guaranteed convergence. The initial value function was also set to zero for Newton method, and the initial policy used for the policy iteration method was to keep the credit limit unchanged.

The optimal values and policies for the risk-averse problem are summarized in Tables 5 and 6.

Since the optimal solutions of both problems for the absorbing states C and D are trivial, they are not provided in the tables. The optimal value is always zero for the absorbing states, and the formal control "Continue" is the optimal control.

When we work with the negatives of profits, the parameter κ of the first-order mean–semideviation can be interpreted as a penalty parameter that penalizes the upper deviations from the mean. This means that the decision maker is less (more) risk averse if κ values are lower (higher). The risk-averse model is equivalent to the expected value model for $\kappa = 0$.

From Table 6, it can be seen that for very small values of κ, the optimal policy is the same for both risk-averse and risk-neutral problems, which is a trivial result of the previous assertion. Similarly, when κ gets smaller, optimal values get closer to the optimal values of expected value problem (see Table 5).

The numbers of iterations needed by both value and policy iteration methods for different values of κ can be found in Table 7. For $\kappa = 1$, the value iteration method required 1,231 iterations, whereas the policy iteration method found the optimal solution in just 3 iterations. When Newton's method was used, the first iteration of the policy iteration method required 6 Newton iterations, the second and third iterations required 2 and 3 Newton iterations, respectively. It can be seen that the policy iteration found the optimal solution in at most 4 iterations, and each iteration required at most 6 Newton iterations when Newton's method was used. However, the value iteration method required much more steps, changing between 525 and 1,354. Policy evaluation by convex optimization method was compared to policy evaluation by Newton's method by comparing the execution times of the entire run of the policy iteration method; the results can be seen in Table 7.

Table 5. Optimal values, $J^*(\cdot)$, of the risk-averse problem for different κ's.

	State								
κ	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)
0.025	−7,006.47	−6,662.47	−4,422.47	−6,779.78	−6,732.78	−6,082.78	−5,890.73	−5,885.73	−5,529.64
0.1	−6,022.33	−5,557.60	−3,317.60	−5,680.78	−5,633.78	−4,983.78	−4,871.23	−4,866.23	−4,484.51
0.2	−4,879.94	−4,271.36	−2,031.36	−4,404.95	−4,357.95	−3,707.95	−3,694.24	−3,689.24	−3,280.65
0.3	−3,890.29	−3,150.33	−910.33	−3,298.83	−3,251.83	−2,601.83	−2,684.25	−2,679.25	−2,246.70
0.4	−3,025.84	−2,166.80	73.20	−2,331.68	−2,284.68	−1,634.68	−1,814.65	−1,809.65	−1,351.35
0.5	−2,263.92	−1,296.49	943.51	−1,477.88	−1,430.88	−780.88	−1,065.10	−1,060.10	−568.84
0.6	−1,583.41	−519.29	1,720.71	−712.82	−665.82	−15.82	−419.64	−414.64	129.33
0.7	−973.84	178.30	2,418.30	−25.64	21.36	671.36	137.76	142.76	753.34
0.8	−500.31	600.94	3,047.74	493.20	641.34	1,291.34	633.92	638.92	1,311.99
0.9	−139.64	879.55	3,618.58	878.60	1,053.13	1,853.64	1,004.58	1,009.58	1,814.67
1	−2.70	989.73	4,140.69	994.50	1,145.21	2,375.02	1,095.70	1,100.70	2,299.66

Table 6. Optimal policy of the risk-averse problem for different κ's.

	State								
κ	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)
0.025	m	h	h	m	h	h	m	m	h
0.1	l	h	h	m	h	h	m	m	h
0.2	l	h	h	m	h	h	m	m	h
0.3	l	h	h	m	h	h	m	m	h
0.4	l	h	h	m	h	h	m	m	h
0.5	l	h	h	m	h	h	m	m	h
0.6	l	h	h	m	h	h	m	m	h
0.7	l	h	h	m	h	h	m	m	h
0.8	l	m	h	l	h	h	m	m	h
0.9	l	m	h	l	m	h	m	m	h
1	l	m	h	l	m	h	m	m	h

8.3. Total Profit Distribution for the Risk-Averse Model

We calculated the expected total profits of each state under the optimal policies of the risk-averse problem with different κ's. This is equivalent to calculating

$$\varphi(x_1)=\mathbb{E}\left[\sum_{t=1}^{\infty} c(x_t,\pi(x_t),x_{t+1})\right], \quad x_1\in\widetilde{\mathscr{X}},$$

for a given stationary policy $\Pi=\{\pi,\pi,\ldots\}$. The expected total profit function $\varphi(x)$, $x\in\mathscr{X}$ can be found by solving the following equation with $\varphi(\mathrm{C})=0$ and $\varphi(\mathrm{D})=0$ (cf. Hernández-Lerma and Lasserre 1999, Lemma 9.4.8):

$$\varphi(x)=r(x,\pi^*(x))+\sum_{y\in\mathscr{X}}(d(x,y)+\varphi(y))\cdot q_{x,y}(\pi^*(x)), \quad x\in\widetilde{\mathscr{X}},$$

where $\Pi=\{\pi^*,\pi^*,\ldots\}$ is the optimal policy of the risk-averse problem. The expected total profits calculated using the above equation can be found in Table 8. For $\kappa=0.025$, the optimal policy of the risk-averse problem is same as the optimal policy of the expected value model; therefore both models give the same expected total profits. When κ gets larger, the decision maker becomes more risk averse and forgoes some profit for more secure policies.

To estimate the distribution of the total profit, we simulated the Markov process under the optimal policies of the expected value model and the risk-averse model with two values of κ: 0.8 and 1. We used the Microsoft Excel-based simulation tool YASAI 2.3 of Eckstein and Riedmueller (2011) of Eckstein and Riedmueller (2002). The sample

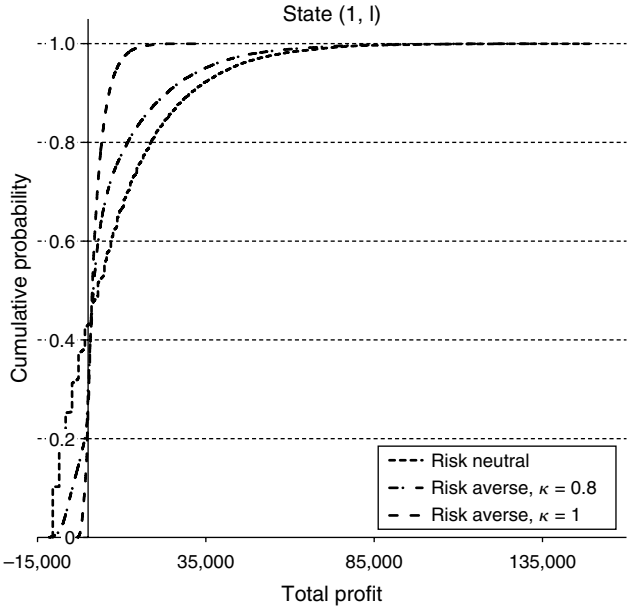
Table 7. Number of iterations for the risk-averse problem.

	Value iteration	Policy iteration with Newton's method			Policy iteration with convex optimization method	
κ	# of value iterations	# of policy iterations	# of Newton iterations	Time (seconds)	# of policy iterations	Time (seconds)
0.025	869	3	4, 3, 3	0.470592	3	0.085575
0.1	797	4	3, 3, 2, 3	0.443240	4	0.108498
0.2	746	4	3, 3, 2, 2	0.384024	4	0.108682
0.3	689	4	4, 2, 2, 2	0.465086	4	0.126204
0.4	658	4	4, 2, 2, 2	0.388726	4	0.096055
0.5	661	4	4, 2, 2, 2	0.422561	4	0.119027
0.6	761	3	4, 3, 3	0.421394	3	0.111233
0.7	893	3	4, 2, 3	0.347835	3	0.108685
0.8	525	3	4, 3, 2	0.353331	3	0.090320
0.9	1,354	3	5, 2, 3	0.398920	3	0.087521
1	1,231	3	6, 2, 3	0.413536	3	0.092212

Table 8. Expected total profits for the risk-averse problem for different κ's.

	State								
κ	(1, l)	(1, m)	(1, h)	(2, l)	(2, m)	(2, h)	(3, l)	(3, m)	(3, h)
0.025	7,407.60	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.1	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.2	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.3	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.4	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.5	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.6	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.7	7,363.82	7,063.60	4,823.60	7,179.09	7,132.09	6,482.09	6,262.99	6,257.99	5,910.98
0.8	6,250.72	5,095.83	4,823.60	5,706.40	7,132.09	6,482.09	6,125.71	6,120.71	5,910.98
0.9	2,096.97	845.98	4,823.60	648.85	408.31	6,482.09	356.36	351.36	5,910.98
1	2,096.97	845.98	4,823.60	648.85	408.31	6,482.09	356.36	351.36	5,910.98

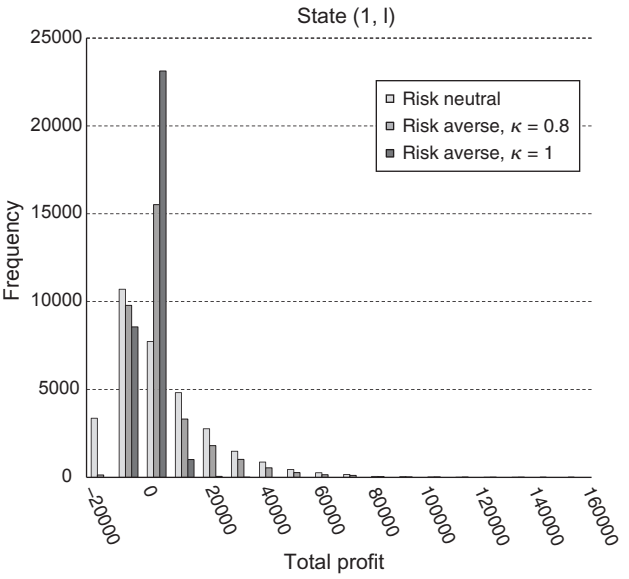
Figure 2. Empirical cumulative probability distribution functions of the total profit at state (1,l).



size was 32,760, and the random number seed used was 10,000. The graphs of the resulting empirical cumulative distribution functions of the total profit, when the initial state is (1,l), are provided in Figure 2. The corresponding histograms are shown in Figure 3.

The first-order mean–semideviation of Example 1 is consistent with stochastic orders. For coherent measures of risk, consistency with the first-order stochastic dominance follows from axiom (A2), under the condition that the probability space Ω is nonatomic (see Shapiro et al. 2009, §6.3.3). However, for the first-order mean–semideviation, consistency with the second-order stochastic dominance is guaranteed without any additional conditions (see Ogryczak and Ruszczyński 1999, 2001, 2002; Shapiro et al. 2009, §6.3.3).

Figure 3. Histograms of the total profit at state (1,l).



Because of consistency with stochastic orders, the first-order mean–semideviation should never prefer stochastically dominated outcomes, which can be observed from Figure 2. Total profits under the optimal policies of the risk-averse model with $\kappa = 0.8$ and $\kappa = 1$ are not stochastically dominated by the total profit of the expected value (risk-neutral) model.

For states with high credit limit, $(\cdot,\mathrm{h})$, the cumulative probability distributions of the total profit are the same for both risk-averse and risk-neutral models. This is because only one control is possible for these states, which is to keep the credit limit unchanged, and the possible transitions are to states with high credit limit, or to C and D. At all other states, the distributions are similar to those for state (1,l).

Acknowledgments

The authors thank two anonymous referees and the associate editor for their insightful comments, which helped improve the presentation of the results. This research was supported by the National Science Foundation [Award CMMI-0965689]. The first author was partially funded by TUBITAK [Grant 213M442].

References

Artzner P, Delbaen F, Eber JM, Heath D (1999) Coherent measures of risk. *Math. Finance* 9:203–228.

Artzner P, Delbaen F, Eber J-M, Heath D, Ku H (2007) Coherent multiperiod risk adjusted values and Bellman's principle. *Ann. Oper. Res.* 152:5–22.

Aubin JP, Frankowska H (1990) *Set-Valued Analysis* (Birkhäuser, Boston).

Bellman R (1957) *Dynamic Programming* (Princeton University Press, Princeton, NJ).

Bertsekas DP, Tsitsiklis JN (1991) An analysis of stochastic shortest-path problems. *Math. Oper. Res.* 16(3):580–595.

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

Boda K, Filar JA (2006) Time consistent dynamic risk measures. *Math. Methods Oper. Res.* 63:169–186.

Cavazos-Cadena R, Fernández-Gaucherand E (1999) Controlled Markov chains with risk-sensitive criteria: average cost, optimality equations and optimal solutions. *Math. Methods Oper. Res.* 49:299–324.

Çavuş Ö, Ruszczyński A (2012) Risk-averse control of undiscounted transient Markov models. http://www.optimization-online.org/.

Cheridito P, Delbaen F, Kupper M (2006) Dynamic monetary risk measures for bounded discrete-time processes. *Electronic J. Probab.* 11:57–106.

Çinlar E (1975) *Introduction to Stochastic Processes* (Prentice-Hall, Englewood Ciffs, NJ).

Coraluppi SP, Marcus SI (1999) Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica* 35:301–309.

Coraluppi SP, Marcus SI (2000) Mixed risk-neutral/minimax control of discrete-time, finite-state Markov decision processes. *IEEE Trans. Automatic Control* 45:528–532.

Denardo EV, Rothblum UG (1979) Optimal stopping, exponential utility, and linear programming. *Math. Programming* 16:228–244.

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

Dynkin EB, Yushkevich AA (1969) *Markov Processes: Theory and Problems* (Plenum, New York).

Dynkin EB, Yushkevich AA (1979) *Controlled Markov Processes* (Springer-Verlag, New York).

Eckstein J, Riedmueller ST (2002) YASAI: Yet another add-in for teaching elementary Monte Carlo simulation in Excel. *INFORMS Trans. Ed.* 2:12–26.

Eckstein J, Riedmueller ST (2011) YASAI (Version 2.3). Accessed July 2012, http://www.yasai.rutgers.edu/download.html.

Feinberg EA, Shwartz A (2002) *Handbook of Markov Decision Processes: Methods and Applications* (Kluwer, Dordrecht, The Netherlands).

Fernàndez-Gaucherand E, Marcus SI (1997) Risk-sensitive optimal control of hidden Markov models: Structural results. *IEEE Trans. Automatic Control* 42:1418–1422.

Filar JA, Lee HM (1985) Gain–variability tradeoffs in undiscounted Markov decision processes. *Proc. 24th IEEE Conf. Decision and Control* (IEEE, Piscataway, NJ), 1106–1112.

Filar JA, Kallenberg LCM, Lee HM (1989) Variance-penalized Markov decision processes. *Math. Oper. Res.* 14(1):147–161.

Fleming WH, Hernández-Hernández D (1997) Risk sensitive control of finite state machines on an infinite horizon. *SIAM J. Control Optim.* 35:1790–1810.

Hernández-Hernández D, Marcus SI (1996) Risk-sensitive control of Markov processes in countable state space. *Systems Control Lett.* 29:147–155.

Hernández-Hernández D, Marcus SI (1999) Existence of risk sensitive optimal stationary policies for controlled Markov processes. *Appl. Math. Optim.* 40:273–285.

Hernández-Lerma O, Lasserre JB (1996) *Discrete-Time Markov Control Processes. Basic Optimality Criteria* (Springer, New York).

Hernández-Lerma O, Lasserre JB (1999) *Further Topics on Discrete-Time Markov Control Processes* (Springer, New York).

Howard RA (1960) *Dynamic Programming and Markov Processes* (John Wiley & Sons, New York).

Howard RA, Matheson JE (1972) Risk-sensitive Markov decision processes. *Management Sci.* 18(1):356–369.

Jaquette SC (1973) Markov decision processes with a new optimality criterion: Discrete time. *Ann. Statist.* 1:496–505.

Jaquette SC (1976) A utility criterion for Markov decision processes. *Management Sci.* 23(1):43–49.

Klatte D, Kummer B (2002) *Nonsmooth Equations in Optimization* (Kluwer, Dordrecht, The Netherlands).

Kummer B (1988) Newton's method for non-differentiable functions. Guddat J, Bank B, Hollatz H, Kall P, Klatte D, Kummer B, Lommatzsch K, Tammer K, Vlach M, Zimmermann K, eds. *Advances in Mathematical Optimization* (Academie-Verlag, Berlin), 114–125.

Levitt S, Ben-Israel A (2001) On modeling risk in Markov decision processes. Rubinov A, Glover B, eds. *Optimization and Related Topics* (Kluwer Academic Publishers, Dordrecht, The Netherlands), 27–40.

Mannor S, Tsitsiklis JN (2011) Mean-variance optimization in Markov decision processes. *Proc. 28th Internat. Conf. Machine Learn., Bellevue, WA* (Omnipress, Madison, WI).

MOSEK (2012) Optimization toolbox for MATLAB. http://mosek.com/.

Nie Y, Wu X (2009) Shortest path problem considering on-time arrival probability. *Transportation Res. B* 43:597–613.

Ogryczak W, Ruszczyński A (1999) From stochastic dominance to mean-risk models: Semideviations as risk measures. *Eur. J. Oper. Res.* 116:33–50.

Ogryczak W, Ruszczyński A (2001) On consistency of stochastic dominance and mean–semideviation models. *Math. Programming* 89: 217–232.

Ogryczak W, Ruszczyński A (2002) Dual stochastic dominance and related mean-risk models. *SIAM J. Optim.* 13:60–78.

Ohtsubo Y (2003) Minimizing risk models in stochastic shortest path problems. *Math. Methods Oper. Res.* 57(1):79–88.

Ohtsubo Y (2004) Optimal threshold probability in undiscounted Markov decision processes with a target set. *Appl. Math. Comput.* 149: 519–532.

Patek SD (2001) On terminating Markov decision processes with a risk averse objective function. *Automatica* 37(9):1379–1386.

Pflug GC, Römisch W (2007) *Modeling, Measuring and Managing Risk* (World Scientific, Singapore).

Pliska SR (1979) On the transient case for Markov decision chains with general state spaces. Puterman ML, eds. *Dynamic Programming and Its Applications* (Academic Press, New York), 335–349.

Puterman ML (1994) *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York).

Riedel F (2004) Dynamic coherent risk measures. *Stochastic Processes Their Appl.* 112:185–200.

Rockafellar RT, Uryasev SP (2002) Conditional value-at-risk for general loss distributions. *J. Banking Finance* 26:1443–1471.

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

Ruszczyński A, Shapiro A (2005) Optimization of risk measures. Calafiore G, Dabbene F, eds. *Probabilistic and Randomized Methods for Design Under Uncertainty* (Springer, London).

Ruszczyński A, Shapiro A (2006a) Optimization of convex risk functions. *Math. Oper. Res.* 31(3):433–452.

Ruszczyński A, Shapiro A (2006b) Conditional risk mappings. *Math. Oper. Res.* 31(3):544–561.

Scandolo G (2003) Risk measures in a dynamic setting. Ph.D. thesis, Università degli Studi di Milano, Milan.

Shapiro A, Dentcheva D, Ruszczyński A (2009) *Lectures on Stochastic Programming* (SIAM Publications, Philadelphia).

So MMC, Thomas LC (2011) Modelling the profitability of credit cards by Markov decision processes. *Eur. J. Oper. Res.* 212:123–130.

Veinott AF (1969) Discrete dynamic programming with sensitive discount optimality criteria. *Ann. Math. Statist.* 40:1635–1660.

White DJ (1988) Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *J. Optim. Theory Appl.* 56:1–29.

Wu CB, Lin YL (1999) Minimizing risk models in Markov decision processes with policies depending on target values. *J. Math. Anal. Appl.* 231:47–67.

Özlem Çavuş is an assistant professor of industrial engineering at Bilkent University, Turkey.

Andrzej Ruszczyński is a distinguished professor at Rutgers University. His research interests are in the area of stochastic programming, in particular risk-averse optimization and risk-averse control.