RISK-AVERSE CONTROL OF UNDISCOUNTED TRANSIENT MARKOV MODELS

by

OZLEM CAVUS

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Operations Research

Written under the direction of

Dr. Andrzej Ruszczyński

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2012

ABSTRACT OF THE DISSERTATION

Risk-Averse Control of Undiscounted Transient Markov Models

by Ozlem Cavus

Dissertation Director: Dr. Andrzej Ruszczyński

The classical optimal control problems for discrete-time, transient Markov processes are infinite horizon, undiscounted expected total cost or reward models. Some examples of these models are optimal stopping problems and stochastic shortest or longest path problems, which may have applications in health-care, finance, and maintenance. However, such expected value models implicitly assume the decision maker is risk-neutral, so they may not be appropriate for several real-life problems.

In this study, we use Markov risk measures to formulate a risk-averse version of the optimal control problem for transient Markov processes with general state and compact control spaces. We derive risk-averse dynamic programming equations and show that they have a unique solution which is also the optimal value of the Markov control problem. Furthermore, it is shown that a randomized policy may be strictly better than deterministic policies, when risk measures are employed.

We suggest two algorithms, value iteration and policy iteration methods, for solving the dynamic programming equations and show their convergence. In general, each policy evaluation step of the policy iteration algorithm requires solving a system of nonsmooth equations. We use a version of nonsmooth Newton method to solve these equations and show its global convergence.

We further consider a risk-averse finite horizon Markov control problem under randomized policies and derive a value iteration method for its solution.

Finally, we work on asset selling, organ transplant, and credit card examples to illustrate the theory for infinite horizon problem, and present numerical results.

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor Dr. Andrzej Ruszczyński for the continuous guidance, support, patience, and trust he showed me throughout this study. I am indebted to him for introducing me to this interesting research topic. I also owe earnest thankfulness to Dr. András Prékopa, Dr. Endre Boros and my committee members, Dr. Adi Ben-Israel, Dr. Darinka Dentcheva, Dr. Farid Alizadeh, Dr. Michael N. Katehakis for their assistance and valuable comments. I am truly thankful to Dr. İ. Kubal Altınel for encouraging me to pursue Ph.D. and for his constant guidance.

I would like to thank to Terry Hart, Clare Smietana, Lynn Agre, Katie D´Agosta, and my colleagues at RUTCOR-Rutgers Center for Operations Research for their help and the sincere, friendly working environment they have provided.

Many thanks to all my friends, especially Evrim, Hande, Nazlı, Gülin, Dilber, Gözde, Ferhat, Betül, Hülya, Berra, Mehmet Ali, and Sami. Life would not have been so enjoyable without them.

Special thanks to Cem İyigün who supported me in every way and has been my best friend throughout my graduate study.

Above all, I would like to thank to my mother Fatma, my father Mehmet, and my brother Özcan. I am really lucky for having such a great and supporting family.

Finally, I would like to acknowledge the funding sources that allowed me to complete my graduate study without any financial concern. I have been supported by RUTCOR as a teaching assistant during the first four years of my Ph.D. This research has also been supported by National Science Foundation under the grant number CMMI-0965689.

Dedication

To my parents...

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Preliminaries

1.1 Introduction

The optimal control problem for transient Markov processes is a classical problem in Operations Research (see Veinott [56], Pliska [44], Bertsekas and Tsitsiklis [6], Hernandez-Lerma and Lasserre [19], and the references therein). The research is focused on the expected total undiscounted cost model for stationary, infinite horizon Markov decision processes, with increased state and control space generality. Some specific examples of such models are stochastic shortest path problems (Bertsekas and Tsitsiklis [6]) and optimal stopping problems (*cf.* Çinlar [10], Dynkin and Yushkevich [13, 14], Puterman [45]).

In this study, we develop and solve a risk-averse model of this problem. In the literature, to our best knowledge, the studies for risk-averse models for transient Markov processes are based on the arrival probability criteria (see, e.g., Ohtsubo [37], [38]) and utility functions (see Howard and Matheson [21], Denardo and Rothblum [12], Patek [42]).

Howard and Matheson [21] use exponential utility function to incorporate risk to Markov decision processes with finite state and control spaces. They consider both finite and undiscounted infinite horizon problems where the decision maker can be either risk-averse or risk-seeking. Under the assumption that Markov chain is both irreducible and acyclic, they suggest a policy iteration algorithm to find the optimal solution of the infinite horizon problem and show its convergence. Later, Jaquette [22] works on a discounted, risk-averse version of the infinite horizon formulation proposed by Howard and Matheson [21], and shows that the optimal policy may not be stationary if discount factor is used. A risk-averse or risk-seeking version of the optimal stopping

problem is suggested by Denardo and Rothblum [12] using exponential utility function. They consider an undiscounted, transient Markov decision process under finite state and control spaces and assume that there exists a terminal state which is absorbing and cost-free. They derive a pair of dual linear programming formulations to find the optimal solution of the problem and provide the conditions of optimality. Patek [42] extends the work of Denardo and Rothblum [12] by relaxing their assumption of finite control space to a compact space but just considers a risk-averse problem with positive costs. Deriving the dynamic programming equations, value and policy iterations are suggested to solve them and the convergence of these algorithms are proved.

Another approach to include risk in Markov decision processes is to use the arrival probability criterion which minimizes (maximizes) the probability that total undiscounted cost (reward) is larger than a given threshold value. Yu et. al. [57] consider an absorbing Markov process with finite state and control spaces. The absorbing states are not cost-free and there exist one-time terminal rewards, therefore, this problem is not completely related to one that we work on. Ohtsubo [37] works on a similar problem where the absorbing state is cost-free and costs are nonnegative. The dynamic programming equations are derived and it is shown that the optimal value of this problem can be found by solving these equations by a value iteration method. Later, Ohtsubo [38] focuses on a similar problem with nonnegative rewards and proposes a policy iteration method in addition to value iteration.

Some other studies related to our work are by Levitt and Ben-Israel [31], Mannor and Tsitsiklis [33], [34]. These studies use the mean-variance risk measure to model risk in finite horizon Markov processes. Mannor and Tsikliklis [34] state that randomized policies may be better than deterministic policies but show this just using a constrained example and do not provide a general proof.

Different from the studies so far, we use the recent theory of dynamic risk measures (see Scandolo [52], Ruszczyński and Shapiro [49, 51], Cheridito, Delbaen and Kupper [8], Artzner et. al. [3], Klöppel and Schweizer [28], Pflug and Römisch [43], and the references therein) to develop and solve a new risk-averse formulation of the stochastic optimal control problem for transient Markov processes. Our results complement and

extend the results of Ruszczyński [48], where finite horizon undiscounted and infinite horizon *discounted* models with deterministic policies are considered. In our presentation, we closely follow the notation and development of [48].

Some applications of these problems concerned with expected performance criteria are given in the survey paper by White [58] 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 [9], 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.

A stochastic shortest path problem can be used to find the optimal replacement time of a system. Kurt and Kharoufe [30] 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 [55]).

So and Thomas [54] 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. account closure), 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.

1.1.1 Outline of the Dissertation

In the remaining sections of this chapter, we quickly review some basic concepts of controlled Markov models and dynamic risk measures. In chapter 2, we adapt and extend the theory of Markov risk measures suggested by Ruszczyński [48]. We then introduce and analyze the concept of a multikernel, which is essential for our theory. Chapter 3 is devoted to the analysis of a finite horizon model with randomized policies. The main model with infinite horizon and dynamic risk measures is analyzed and solved in chapter 4. In that chapter, we further compare randomized and deterministic polices, and give a condition where it is enough to consider just deterministic policies. In chapter 5, we suggest value and policy iteration methods for the solution of infinite horizon problem and show their convergence. Another solution approach, which we call as mathematical programming approach, is analyzed in chapter 6. Finally, chapter 7 illustrates our results on risk-averse versions of an optimal stopping problem of Karlin [26], of the organ transplant problem of Alagoz *et al.* [1], and of the credit card problem by So and Thomas [54].

Chapters 1, 2, 3, 4, and Sections 7.1 and 7.2 are based on the study [7] which is currently under review. Chapters 5 and 6, and Section 7.3 first appear here.

1.2 Controlled Markov Processes

In this section, we review the main concepts of controlled Markov models and we introduce relevant notation (for details, see [17, 18, 19]). Let $\mathcal{X}$ be a state space, and $\mathcal{U}$ a control space. We assume that $\mathcal{X}$ and $\mathcal{U}$ are Polish spaces, equipped with their Borel σ-algebras. A control set is a measurable multifunction $U : \mathcal{X} \rightrightarrows \mathcal{U}$; 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 measurable mapping from the graph of U to the set $\mathcal{P}(\mathcal{X})$ of probability measures on $\mathcal{X}$ (equipped with the topology of weak convergence).

The cost of transition from x to y, when control u is applied, is represented by $c(x, u, y)$, where $c : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to \mathbb{R}$. 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 *controlled Markov process* is represented by a state space $\mathcal{X}$, a control space $\mathcal{U}$, control sets U_t, controlled transition kernels Q_t, and cost functions c_t, $t = 1, 2, \ldots$. It is called a *stationary controlled Markov process* if there exist a control set U, transition kernel Q, and cost function c such that $U_t = U$, $Q_t = Q$, and $c_t = c$ for all $t = 1, 2, 3, \ldots$.

For $t = 1, 2, \ldots$ we define the space of state and control histories up to time t as $\mathcal{H}_t = \text{graph}(U)^t \times \mathcal{X}$. 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 $\pi_t : \mathcal{H}_t \to \mathcal{P}(\mathcal{U})$, $t = 1, 2, \ldots$, 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 $\pi_t : \mathcal{X} \to \mathcal{P}(\mathcal{U})$, $t = 1, 2, \ldots$, 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 $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{U})$ 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* $x_{\mathrm{A}} \in \mathcal{X}$, 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 $\widetilde{\mathcal{X}} = \mathcal{X} \setminus \{x_{\mathrm{A}}\}$, and the effective controlled substochastic kernel $\widetilde{Q}$ whose arguments are restricted to $\widetilde{\mathcal{X}}$ and whose values are nonnegative measures on $\widetilde{\mathcal{X}}$, so that $\widetilde{Q}\big(B\big|x, u\big) = Q\big(B\big|x, u\big)$, for all Borel sets $B \subset \widetilde{\mathcal{X}}$, all $x \in \widetilde{\mathcal{X}}$, and all $u \in U(x)$. Moreover, we assume that the following Pliska condition [44] is satisfied: a weight function $w : \mathcal{X} \to [1, \infty)$ and a constant K exist, such that for every Markov decision rule π we have

$$\sum_{j=1}^{\infty} \left\| \left(\widetilde{Q}^{\pi}\right)^{j} \right\|_{w} \leq K. \tag{1.1}$$

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

$$\|A\|_w = \sup_{x \in \widetilde{\mathcal{X}}} \frac{1}{w(x)} \int_{\widetilde{\mathcal{X}}} w(y)\ A(dy|x). \tag{1.2}$$

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

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

Hernadez-Lerma and Lasserre [19] 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 $\mathit{\Pi} = \{\pi_t\}_{t=1}^{\infty}$ so as to minimize the expected cost until absorption:

$$\min_{\mathit{\Pi}}\ \mathbb{E}\left[\sum_{t=1}^{\infty} c(x_t, u_t, x_{t+1})\right]. \tag{1.3}$$

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.

---

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

The optimal value can be found by solving the following dynamic programming equations (*cf.* Pliska [44] and Hernandez-Lerma and Lasserre [19]):

$$v(x) = \inf_{u \in U(x)} \int_{\mathcal{X}} \big(c(x,u,y) + v(y)\big)\, Q(dy|x,u), \quad x \in \widetilde{X},$$

$$v(x_A) = 0.$$

Here, the functions $v(x),\ x \in \mathcal{X}$ are called *value functions* and the minimizer $\hat{\pi}(x),\ x \in \mathcal{X}$:

$$\hat{\pi}(x) \in \operatorname*{arginf}_{u \in U(x)} \int_{\mathcal{X}} \big(c(x,u,y) + v(y)\big)\, Q(dy|x,u), \quad x \in \widetilde{X},$$

defines an optimal stationary Markov policy $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}, \ldots\}$.

Our aim is to introduce risk aversion to problem (1.3), and to replace the expected value operator by a dynamic risk measure. We shall show that the Pliska condition (1.1) 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 (1.3) 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.

1.3 Dynamic Risk Measures

Suppose T is a fixed time horizon. Each policy $\mathit{\Pi} = \{\pi_1, \pi_2, \ldots\}$ 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 $\mathcal{Z}_t$ of $\mathcal{F}_t$-measurable random variables on $\Omega,\ t = 1, \ldots, T$. In this study, we focus on the case when $\mathcal{Z}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P)$, for some $p \in [1, \infty]$. The reader is referred to Shapiro et. al. [53] and Ruszczyński [48] for details.

To evaluate risk of this cost sequence we use a dynamic time-consistent risk measure. Before giving the definition of that concept, we will provide some preliminaries.

It is convenient to introduce vector spaces $\mathcal{Z}_{t,\theta} = \mathcal{Z}_t \times \mathcal{Z}_{t+1} \times \cdots \times \mathcal{Z}_\theta$, where

$1 \leq t \leq \theta \leq T+1$ and the conditional risk measures $\rho_{t,\theta} : \mathcal{Z}_{t,\theta} \to \mathcal{Z}_t$ defined as follows:

$$\rho_{t,\theta}(Z_t, \ldots, Z_\theta) = Z_t + \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{\theta-1}(Z_\theta)\cdots\big)\Big). \tag{1.4}$$

Here, $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$, $t = 1, \ldots, T$, are one-step conditional risk measures assumed to satisfy the following conditions:

A1. *Convexity*: $\rho_t(\alpha Z+(1-\alpha)W) \leq \alpha\rho_t(Z)+(1-\alpha)\rho_t(W),\ \forall\, \alpha \in (0,1),\ Z, W \in \mathcal{Z}_{t+1}$;

A2. *Monotonicity*: If $Z \leq W$ then $\rho_t(Z) \leq \rho_t(W),\ \forall\, Z, W \in \mathcal{Z}_{t+1}$;

A3. *Predictable Translation Equivariance*: $\rho_t(Z+W) = Z + \rho_t(W),\ \forall\, Z \in \mathcal{Z}_t,\ W \in \mathcal{Z}_{t+1}$;

A4. *Positive Homogeneity*: $\rho_t(\beta Z) = \beta\rho_t(Z),\ \forall\, Z \in \mathcal{Z}_{t+1},\ \beta \geq 0$.

Some examples of one-step conditional risk measures satisfying above properties are *first-order mean-semideviation* (see, Ogryczak and Ruszczyński [39, 40], and Ruszczyński and Shapiro [50, Example 4.2], [51, Example 6.1]) and *Conditional Average Value at Risk* (see, *inter alia*, Ogryczak and Ruszczyński [41, Sec. 4], Pflug and Römisch [43, Sec. 2.2.3, 3.3.4], Rockafellar and Uryasev [46], Ruszczyński and Shapiro [50, Example 4.3], [51, Example 6.2]).

Example 1.3.1 The *first-order mean–semideviation* risk measure is defined by the function:

$$\rho_t(Z_{t+1}) = \mathbb{E}[Z_{t+1}|\mathcal{F}_t] + \kappa\mathbb{E}\big[\big(Z_{t+1} - \mathbb{E}[Z_{t+1}|\mathcal{F}_t]\big)_+|\mathcal{F}_t\big],$$

where $\kappa \in [0,1]$.

Example 1.3.2 The *Conditional Average Value at Risk* is calculated by the function:

$$\rho_t(Z_{t+1}) = \inf_{\eta \in \mathcal{Z}_t}\left\{\eta + \frac{1}{\alpha}\mathbb{E}\big[(Z_{t+1} - \eta)_+|\mathcal{F}_t\big]\right\},$$

with α being in the interval $[\alpha_{\min}, \alpha_{\max}] \subset (0,1)$.

Immediately from the definition of conditional risk measure (1.4) and using the properties of one-step conditional risk measures, we obtain the following properties of $\rho_{t,\theta} : \mathcal{Z}_{t,\theta} \to \mathcal{Z}_t$.

Lemma 1.3.3 *If the one-step conditional risk measures ρ_τ, $\tau = t, \dots, \theta - 1$, satisfy conditions (A1)–(A4), then*

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

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

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

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

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

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

A sequence $\{\rho_{t,T}\}_{t=1}^T$ of conditional risk measures $\rho_{t,T} : \mathcal{Z}_{t,T} \to \mathcal{Z}_t$ is called a *dynamic risk measure*. In this study, we assume that dynamic risk measures have time-consistency property. Here, we repeat the definition of time-consistency from [48].

Definition 1.3.4 *[48, Definition 3] Suppose a dynamic risk measure $\{\rho_{t,T}\}_{t=1}^T$ satisfies the following conditions*

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

for all $1 \le \tau < \theta \le T$ and all $Z, W \in \mathcal{Z}_{\tau,T}$. Then it is time-consistent *if*

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

In this study, we use the following form of a time-consistent dynamic measure of risk:

$$\begin{aligned} J_T(\Pi, x_1) = \rho_1\Big(c(x_1, u_1, x_2) + \rho_2\Big(c(x_2, u_2, x_3) + \cdots \\ + \rho_{T-1}\big(c(x_{T-1}, u_{T-1}, x_T) + \rho_T(c(x_T, u_T, x_{T+1}))\big)\cdots\Big)\Big). \end{aligned} \tag{1.5}$$

Ruszczyński [48, sec. 3] derives the nested formulation (1.5) 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 of [2] (see (B1)–(B4) in chapter 2).

Chapter 2

Markov Risk Measures

2.1 Markov Risk Measures for Randomized Policies

As indicated in [48], the fundamental difficulty of formulation (1.5) 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. In order to overcome this difficulty, in [48, sec. 4] a new construction of a one-step conditional measure of risk is introduced. Its arguments are functions on the state space $\mathcal{X}$, 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 $\mathcal{U} \times \mathcal{X}$ equipped with its product Borel σ-algebra $\mathcal{B}$.

Suppose the current state is x and we use a randomized control λ. We define Q_x as the mapping $u \to Q(\cdot|x,u)$. The randomized control λ, together with the transition kernel Q defines a probability measure $\lambda \circ Q_x$ on the product space $\mathcal{U} \times \mathcal{X}$ as follows:

$$[\lambda \circ Q_x](B_u \times B_y) = \int_{B_u} Q(B_y|x,u)\ \lambda(du), \quad B_u \in \mathcal{B}(U), \quad B_y \in \mathcal{B}(\mathcal{X}). \tag{2.1}$$

The measure is extended to other sets in $\mathcal{B}$ 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 $\mathcal{U} \times \mathcal{X}$ defined as follows:

$$c_x(u,y) = c(x,u,y), \quad u \in \mathcal{U}, \quad y \in \mathcal{X}. \tag{2.2}$$

Let $\mathcal{V} = \mathcal{L}_p(\mathcal{U} \times \mathcal{X}, \mathcal{B}, P_0)$, where $p \in [1,\infty]$ and P_0 is some reference probability measure on $\mathcal{U} \times \mathcal{X}$. It is convenient to think of the dual space $\mathcal{V}'$ as the space of

signed measures m on $(\mathcal{U} \times \mathcal{X}, \mathcal{B})$, which are absolutely continuous with respect to P_0, with densities (Radon–Nikodym derivatives) lying in the space $\mathcal{L}_q(\mathcal{U} \times \mathcal{X}, \mathcal{B}, P_0)$, 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 $\mathcal{V}'$. The measure P_0 does not play any other role in our considerations. We consider the set of probability measures in $\mathcal{V}'$:

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

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

$$\langle \varphi, m \rangle = \int_{\mathcal{U} \times \mathcal{X}} \varphi(u, y)\, m(du \times dy), \quad \varphi \in \mathcal{V}, \quad m \in \mathcal{V}'. \tag{2.3}$$

The space $\mathcal{V}'$ (and thus $\mathcal{M}$) will be endowed with the weak* topology. For $p \in [1, \infty)$ we may endow $\mathcal{V}$ with the strong (norm) topology, or with the weak topology. For $p = \infty$, the space $\mathcal{V}$ will be endowed with is weak topology defined by the form (2.3), that is, the weak* topology on $\mathcal{L}_\infty(\mathcal{X}, \mathcal{B}, P_0)$.

Definition 2.1.1 *A measurable function $\sigma : \mathcal{V} \times \mathcal{X} \times \mathcal{M} \to \mathbb{R}$ 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 $\mathcal{V}$.*

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

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

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

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

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

Example 2.1.2 Consider the first-order mean–semideviation risk measure of Example 1.3.1, but with the state and the underlying probability measure as its arguments. We

define

$$\sigma(\varphi, x, m) = \langle \varphi, m \rangle + \kappa(x)\big\langle (\varphi - \langle \varphi, m \rangle)_+, m \big\rangle, \tag{2.4}$$

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

Example 2.1.3 Another important example is the Conditional Average Value at Risk (see, Example 1.3.2), 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 $\alpha : \mathcal{X} \to [\alpha_{\min}, \alpha_{\max}] \subset (0, 1)$ 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 $\varphi \in \mathcal{V}$ and a probability measure $\mu \in \mathcal{M}$ we can define the distribution function $F^{\mu}_{\varphi} : \mathbb{R} \to [0, 1]$ as follows

$$F^{\mu}_{\varphi}(\eta) = \mu\big\{(u, y) \in \mathcal{U} \times \mathcal{X} : \varphi(u, y) \leq \eta\big\}.$$

Definition 2.1.4 *A risk transition mapping $\sigma : \mathcal{V} \times \mathcal{X} \times \mathcal{M} \to \mathbb{R}$ is* law invariant*, if for all $\varphi, \psi \in \mathcal{V}$ and all $\mu, \nu \in \mathcal{M}$ such that $F^{\mu}_{\varphi} \equiv F^{\nu}_{\psi}$, we have $\sigma(\varphi, x, \mu) = \sigma(\psi, x, \nu)$ for all $x \in \mathcal{X}$.*

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 2.1.2 and 2.1.3 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 7.1.

Risk transition mappings allow for convenient formulation of risk-averse preferences for controlled Markov processes, where the cost is evaluated by formula (1.5). Consider a controlled Markov process $\{x_t\}$ with some Markov policy $\mathit{\Pi} = \{\pi_1, \pi_2, \dots\}$. For a fixed time t and a measurable function $g : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to \mathbb{R}$ the value of $Z_{t+1} = g(x_t, u_t, x_{t+1})$ is a random variable. We assume that g is *w-bounded*, that is,

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

for some constant $C > 0$ and for the weight function $w : \mathcal{X} \to [1, \infty)$, $w \in \mathcal{V}$. Then Z_{t+1} is an element of $\mathcal{Z}_{t+1}$. Let $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ 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 $\mathcal{Z}_t$, 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 $g_x : \mathcal{U} \times \mathcal{X} \to \mathbb{R}$ 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 2.1.5 *A one-step conditional risk measure $\rho_t : \mathcal{Z}_{t+1} \to \mathcal{Z}_t$ is a* Markov risk measure *with respect to the controlled Markov process $\{x_t\}$, if there exists a risk transition mapping $\sigma_t : \mathcal{V} \times \mathcal{X} \times \mathcal{M} \to \mathbb{R}$ such that for all w-bounded measurable functions $g : \mathcal{X} \times \mathcal{U} \times \mathcal{X} \to \mathbb{R}$ and for all feasible decision rules $\pi : \mathcal{X} \to \mathcal{P}(U)$ we have*

$$\rho_t\big(g(x_t, u_t, x_{t+1})\big) = \sigma_t\big(g_{x_t}, x_t, \pi_{x_t} \circ Q_{x_t}\big), \quad a.s. \tag{2.5}$$

Observe that the right hand side of formula (2.5) is parametrized by x_t, and thus it defines a special $\mathcal{F}_t$-measurable function of ω, whose dependence on the past is carried only via the state x_t.

Remark 2.1.6 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 $\mathcal{X}$ and define $\mathcal{V} = \mathcal{L}_p(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_0)$, $\mathcal{V}'$ - the set of measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ having densities with respect to P_0 in $\mathcal{L}_q(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_0)$, and $\mathcal{M} = \{m \in \mathcal{V}' : m(\mathcal{X}) = 1,\ m \geq 0\}$, exactly as in [48].

Remark 2.1.7 If, additionally, the stage-wise costs have the form $c(x_t, u_t, x_{t+1}, \xi_t)$, where ξ_t, $t = 1, 2, \ldots$, are some random variables distributed in 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:

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

All our considerations remain valid, just the notation complicates.

2.2 Stochastic Multikernels

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

Definition 2.2.1 *A* multikernel *is a measurable multifunction $\mathfrak{M}$ from $\mathcal{X}$ to the space* $\mathrm{rca}(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ *of regular measures on $(\mathcal{X}, \mathcal{B}(X))$. It is* stochastic*, if its values are sets of probability measures. It is* substochastic*, if $0 \leq M(B|x) \leq 1$ for all $M \in \mathfrak{M}(x)$, $B \in \mathcal{B}(\mathcal{X})$, and $x \in \mathcal{X}$. It is* convex *(*closed*), if for all $x \in \mathcal{X}$ 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 $x \in \mathcal{X}$. 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:

$$\big[M_1 \circ M_2\big]\big(B\big|x\big) = \int_{\mathcal{X}} M_2(B|y)\ M_1(dy|x), \quad \mathcal{B} \in \mathcal{B}(\mathcal{X}), \quad x \in \mathcal{X}. \tag{2.6}$$

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

Definition 2.2.2 *If $\mathfrak{M}_i : \mathcal{X} \rightrightarrows \mathrm{rca}(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, $i = 1, 2$ 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 2.2.2, 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 $\mathfrak{M} : \widetilde{\mathcal{X}} \rightrightarrows \mathrm{rca}(\widetilde{\mathcal{X}}, \mathcal{B}(\widetilde{\mathcal{X}}))$ is defined as follows:

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

where the norm $\|M\|_w$ is given by (1.2).

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 $x \in \mathcal{X}$ and $m \in \mathcal{M}$ (see Ruszczyński and Shapiro [50, Proposition 3.1]). Then it follows from [50, Theorem 2.2] that for every $x \in \mathcal{X}$ and $m \in \mathcal{M}$ a closed convex set $\mathcal{A}(x,m) \subset \mathcal{M}$ exists, such that for all $\varphi \in \mathcal{V}$ we have

$$\sigma(\varphi, x, m) = \max_{\mu \in \mathcal{A}(x,m)} \langle \varphi, \mu \rangle. \tag{2.7}$$

In fact, we also have

$$\mathcal{A}(x,m) = \partial_\varphi \sigma(0, x, m). \tag{2.8}$$

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

Example 2.2.3 For the mean-semideviation model of Example 2.1.2, following the derivations of Ruszczyński and Shapiro [50, Example 4.2], we have

$$\mathcal{A}(x,m) = \Big\{\mu \in \mathcal{M} : \exists \big(h \in \mathcal{L}_\infty(\mathcal{U}\times\mathcal{X}, \mathcal{B}, P_0)\big)\ \frac{d\mu}{dm} = 1 + h - \langle h, m\rangle,\ \|h\|_\infty \le \kappa(x),\ h \ge 0\Big\}. \tag{2.9}$$

Similar formulas can be derived for higher order measures.

Example 2.2.4 For the Conditional Average Value at Risk of Example 2.1.3, following the derivations of Ruszczyński and Shapiro [50, Example 4.3], we obtain

$$\mathcal{A}(x,m) = \left\{\mu \in \mathcal{M} : \frac{d\mu}{dm} \le \frac{1}{\alpha(x)}\right\}. \tag{2.10}$$

Consider the formula (2.5) and suppose that $g(x_t, u_t, x_{t+1}) = v(x_{t+1})$ for some measurable w-bounded function $v : \mathcal{X} \to \mathbb{R}$. Using the representation (2.7) we can write it as follows:

$$\rho_t\big(v(x_{t+1})\big) = \max_{\mu \in \mathcal{A}(x_t, \pi_{x_t} \circ Q_{x_t})} \langle v, \mu \rangle, \quad \text{a.s.} \tag{2.11}$$

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

$$\bar{\mu}(B) = \mu(\mathcal{U} \times B), \quad B \in \mathcal{B}(\mathcal{X}). \tag{2.12}$$

Denote by L the linear operator mapping each $\mu \in \mathcal{V}'$ to the corresponding marginal measure $\bar{\mu}$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, as defined in (2.12). For every x we can define the set of probability measures:

$$\mathfrak{M}^\pi_x = \big\{ L\mu : \mu \in \mathcal{A}(x, \pi_x \circ Q_x) \big\}, \quad x \in \mathcal{X}. \tag{2.13}$$

The multifunction $\mathfrak{M}^\pi : \mathcal{X} \rightrightarrows \mathcal{P}(\mathcal{X})$, assigning the set $\mathfrak{M}^\pi_x$ to each $x \in \mathcal{X}$, 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 $M^\pi \ll \mathfrak{M}^\pi$ are transition kernels.

It follows that formula (2.11) can be rewritten as follows:

$$\rho_t\big(v(x_{t+1})\big) = \max_{M \in \mathfrak{M}^\pi_{x_t}} \int_{\mathcal{X}} v(y)\, M(dy). \tag{2.14}$$

In the risk-neutral case we have

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

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

$$\mathfrak{M}^\pi_x = \big\{ Q^\pi_x \big\}, \quad x \in \mathcal{X}, \tag{2.15}$$

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 4.1, where we calculate some examples of transition multikernels.

Remark 2.2.5 *If $m \in \mathcal{A}(x,m)$ for all $x \in \mathcal{X}$ and $m \in \mathcal{M}$, then it follows from equation* (2.13) *that Q^π is a measurable selector of $\mathfrak{M}^\pi$. Moreover, it follows from* (2.7) *that for any function $\varphi \in \mathcal{V}$ we have*

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

It follows that the dynamic risk measure (1.5) *is bounded from below by the expected value of the total cost.*

The condition $m \in \mathcal{A}(x,m)$ is satisfied by the measures of risk in Examples 2.2.3 and 2.2.4.

Interestingly, uncertain transition matrices were used by Nilim and El Ghaoui in [36] 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 2.2.6 *Suppose $\varphi \in \mathcal{V}$ and $x \in \mathcal{X}$. If the controlled kernel $u \mapsto Q(\cdot|x,u)$ is continuous, and the multifunction $m \mapsto \mathcal{A}(\varphi, x, m)$ is lower semicontinuous, then the function $\lambda \mapsto \sigma(\varphi, x, \lambda \circ Q_x)$ is weakly* lower semicontinuous on $\mathcal{P}\big(U(x)\big)$.*

Proof For a continuous Q, the multifunction $\lambda \mapsto \mathcal{A}(x, \lambda \circ Q_x)$ inherits the continuity properties of $\mathcal{A}$. The function $\mu \mapsto \langle \varphi, \mu \rangle$ is continuous on $\mathcal{M}$ (in the weak* topology). The assertion of the theorem follows now from the dual representation (2.7) by [4, Theorem 1.4.16], whose proof remains valid in our setting as well. □

Some comments on the assumptions of Proposition 2.2.6 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., [18]). If the risk transition mapping $\sigma(\cdot,\cdot,\cdot)$ is continuous, then its subdifferential (2.8) is upper semicontinuous. However, in Proposition 2.2.6 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 2.2.3 and 2.2.4 are continuous with respect to m.

Chapter 3

Finite Horizon Problem

In this chapter, we consider a finite horizon model with randomized policies, which generalizes the model suggested by Ruszczyński [48] for deterministic policies. We derive the dynamic programming equations and prove that optimal solution of the problem can be found by iteratively solving these equations.

3.1 Dynamic Programming Equations for Finite Horizon Problems

We consider the Markov model at times $1, 2, \ldots, T+1$ under general policies $\Pi = \{\pi_1, \pi_2, \ldots, \pi_T\}$.

For the finite horizon problem, it is not necessary to assume that the Markov process is stationary. Therefore, we will use time-dependent control sets U_t, transition kernels Q_t, cost functions c_t, and multifunctions $\mathcal{A}_t$, $t = 1, \ldots, T$. Additionally, the assumption that the process is transient is not needed. We assume that the cost at the last stage is given by a function $v_{T+1}(x_{T+1})$.

Consider the problem

$$\min_{\Pi} \; J_T(\Pi, x_1), \tag{3.1}$$

where $J_T(\Pi, x_1)$ is defined by formula (1.5), with Markov conditional risk measures ρ_t, $t = 1, \ldots, T$ specified by risk transition mappings $\sigma_t(\cdot, \cdot, \cdot)$:

$$J_T(\Pi, x_1) = \rho_1\Big(c_1(x_1, u_1, x_2) + \rho_2\Big(c_2(x_2, u_2, x_3) + \cdots + \rho_T\big(c_T(x_T, u_T, x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big). \tag{3.2}$$

In the following theorem, we show that, similar to the risk-neutral case, the optimal solution of problem (3.1) can be found by solving appropriate dynamic programming

equations. A similar theorem is proved by Ruszczyński [48, Thm. 2] for deterministic policies.

Theorem 3.1.1 *If the following conditions are satisfied:*

(i) *The transition kernels $Q_t(x,\cdot)$ are continuous for every $x \in \mathcal{X}$ and $t = 1,\dots,T$;*

(ii) *The conditional risk measures ρ_t, $t = 1,\dots,T$ are Markovian and the multifunctions $\mathcal{A}_t(x,\cdot)$ are lower semicontinuous for every $x \in \mathcal{X}$;*

(iii) *The functions $c_t(\cdot,\cdot,\cdot)$, $t = 1,\dots,T$ are w-bounded, measurable, and lower semicontinuous with respect to the second argument;*

(iv) *The sets $U_t(x)$, $t = 1,\dots,T$ are compact for every $x \in \mathcal{X}$;*

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

then problem (3.1) *has an optimal solution and its optimal value $v_1(x)$ can be found by solving following dynamic programming equations:*

$$v_t(x) = \min_{\lambda_t \in \mathcal{P}(U_t(x))} \sigma_t\big(c_t(x,\cdot,\cdot) + v_{t+1}, x, \lambda_t \circ Q_t(x,\cdot)\big), \quad x \in \mathcal{X}, \quad t = T,\dots,1. \tag{3.3}$$

Furthermore, there exists an optimal Markov policy $\hat{\Pi} = \{\hat{\pi}_1,\dots,\hat{\pi}_T\}$ which satisfies the equations:

$$\hat{\pi}_t(x) \in \operatorname*{argmin}_{\lambda_t \in \mathcal{P}(U_t(x))} \sigma_t\big(c_t(x,\cdot,\cdot) + v_{t+1}, x, \lambda_t \circ Q_t(x,\cdot)\big), \quad x \in \mathcal{X}, \quad t = T,\dots,1. \tag{3.4}$$

Conversely, every solution of equations (3.3)–(3.4) *defines an optimal Markov policy $\hat{\Pi}$.*

Proof Our proof is similar to the proof of Ruszczyński [48, Thm. 2], but with adjustments due to the use of randomized strategies. Therefore, here, we just provide its short outline.

Since the conditional risk measures ρ_t, $t = 1,\dots,T$ are Markovian, we can apply the monotonicity condition (B2) and obtain the following forms which are equivalent to problem (3.1):

$$\min_{\pi_1,\dots,\pi_T} \Big\{\rho_1\Big(c_1(x_1,u_1,x_2) + \cdots + \rho_T\big(c_T(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big\} =$$

$$\min_{\pi_1,\dots,\pi_{T-1}} \Big\{\rho_1\Big(c_1(x_1,u_1,x_2) + \cdots + \min_{\pi_T} \rho_T\big(c_T(x_T,u_T,x_{T+1}) + v_{T+1}(x_{T+1})\big)\cdots\Big)\Big\}.$$

Since ρ_T is a Markov risk measure, the innermost optimization problem can be written as follows:

$$\min_{\lambda_T \in \mathcal{P}(U_T(x_T))} \sigma_T\big(c_T(x_T, \cdot, \cdot) + v_{T+1}, x_T, \lambda_T \circ Q_T(x_T, \cdot)\big). \tag{3.5}$$

which is equivalent to (3.3) for $t = T$. Furthermore, its solution is found by solving (3.4) for $t = T$.

The function $\lambda_T \mapsto \sigma_T\big(c_T(x_T, \cdot, \cdot) + v_{T+1}, x_T, \lambda_T \circ Q_T(x_T, \cdot)\big)$ is lower semicontinuous by Proposition 2.2.6. As the set of $\lambda_T \in \mathcal{P}(\mathcal{U})$ such that $\lambda_T(U_T(x_T)) = 1$ is weakly* compact, the optimal randomized decision rule $\pi_T(x)$, which is the minimizer in (3.5), exists.

After that, the horizon $T + 1$ is decreased to T, and the final cost becomes $v_T(x_T)$. The theorem is proved by continuing in this way for $T, T - 1, \ldots, 1$. □

Iteration of (3.3) gives that the value functions $v_t(\cdot)$ are the optimal values of the following subproblems formulated for a fixed $x_t = x$:

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

It follows from Theorem 3.1.1 that the optimal solution of the finite horizon problem (3.1) can be calculated by iteratively solving the equations (3.3)–(3.4) for $T, T-1, \ldots, 1$.

Chapter 4

Infinite Horizon Problem

4.1 Evaluation of Stationary Markov Policies in Infinite Horizon Problems

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

$$J_\infty(\mathit{\Pi}, x_1) = \lim_{T\to\infty} J_T(\mathit{\Pi}, x_1), \tag{4.1}$$

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

$$\begin{aligned} J_T(\mathit{\Pi}, x_1) &= \rho_1\Big(c(x_1, u_1, x_2) + \rho_2\Big(c(x_2, u_2, x_3) + \cdots + \rho_T\big(c(x_T, u_T, x_{T+1})\big)\cdots\Big)\Big) \\ &= \rho_{1,T+1}\big(0, c(x_1, u_1, x_2), c(x_2, u_2, x_3), \dots, c(x_T, u_T, x_{T+1})\big), \end{aligned} \tag{4.2}$$

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 3.1.1 with stationary transition kernel Q, cost function c, control sets U, and multifunction $\mathcal{A}$. We still have to index each conditional risk measure by time, because by definition it acts from the space $\mathcal{Z}_{t+1}$ to the space $\mathcal{Z}_t$.

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 (4.1) may be infinite, as the following example demonstrates.

Example 4.1.1 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 (4.1) 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 2.1.3, with $0 < \alpha \le \frac{1}{2}$, to define $\rho_1(\cdot)$. Using standard identities for the Average Value at Risk (see, e.g., [53, Thm. 6.2]), we obtain

$$\begin{aligned} J_\infty(1) &= \inf_{\eta\in\mathbb{R}} \Big\{\eta + \frac{1}{\alpha}\mathbb{E}\big[\big(1 + J_\infty(x_2) - \eta\big)_+\big]\Big\} \\ &= 1 + \inf_{\eta\in\mathbb{R}} \Big\{\eta + \frac{1}{\alpha}\mathbb{E}\big[\big(J_\infty(x_2) - \eta\big)_+\big]\Big\} = 1 + \frac{1}{\alpha}\int_{1-\alpha}^{1} F^{-1}(\beta)\, d\beta, \end{aligned} \tag{4.3}$$

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 (4.3) 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 2.1.2, we obtain

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

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 (1.1) 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 (2.13). Similar to the expected value case, it is convenient to consider the effective state space $\widetilde{\mathcal{X}} = \mathcal{X} \setminus \{x_{\mathrm{A}}\}$, and the *effective substochastic multikernel* $\widetilde{\mathfrak{M}}^\pi$ whose arguments are restricted to $\widetilde{\mathcal{X}}$ and whose values are convex sets of nonnegative measures on $\widetilde{\mathcal{X}}$, so that $\widetilde{M}\big(B\big|x,u\big) = M\big(B\big|x,u\big)$, for all Borel sets $B \subset \widetilde{\mathcal{X}}$, and all $M \in \widetilde{\mathfrak{M}}^\pi$.

Definition 4.1.2 *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 $w : \mathcal{X} \to [1,\infty)$, $w \in \mathcal{V}$, and a constant K exist such*

$$\sum_{j=1}^{\infty} \Big\| \big(\widetilde{\mathfrak{M}}^\pi\big)^j \Big\|_w \le K. \tag{4.4}$$

If the estimate (4.4) *is uniform for all Markov policies, the model is called* uniformly risk-transient.

In the special case of a risk-neutral model, Definition 4.1.2 reduces to the Pliska condition (1.1), owing to the equation (2.15).

Example 4.1.3 Consider the simple transient chain of Example 4.1.1 with the Average Value at Risk from Examples 2.1.3 and 2.2.4, where $0 < \alpha \le 1$. From (2.10) we obtain

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

As only one control is possible, formula (2.13) simplifies to

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

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

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

For $0 < \alpha \le \frac{1}{2}$ we can select $\widetilde{M} = 1 \in \widetilde{\mathfrak{M}}_1$ to show that $1 \in \big(\widetilde{\mathfrak{M}}_1\big)^j$ for all j, and thus condition (4.4) 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 (4.4) is satisfied.

Consider now the mean-semideviation model of Examples 2.1.2 and 2.2.3. From (2.9) we obtain

$$\mathcal{A}(x,m) = \Big\{(\mu_1,\mu_2) : \mu_j = m_j\left(1 + h_j - (h_1 m_1 + h_2 m_2)\right),\ 0 \le h_j \le \kappa,\ j = 1,2\Big\},$$
$$\mathfrak{M}_i = \Big\{(\mu_1,\mu_2) : \mu_j = Q_{ij}\left(1 + h_j - (h_1 Q_{i1} + h_2 Q_{i2})\right),\ 0 \le h_j \le \kappa,\ j = 1,2\Big\}, \quad i = 1,2.$$

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

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

For every $\kappa \in [0,1]$, Definition 4.1.2 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 $v : \mathcal{X} \to \mathbb{R}$ is a measurable function with $\|v\|_w < \infty$ for the weight function $w : \mathcal{X} \to [1,\infty)$, $w \in \mathcal{V}$, and with $v(x_{\mathrm{A}}) = 0$. In the lemma below, we consider $x_1 \in \mathcal{X}$ as a parameter of the problem, and thus $\rho_{1,T}\big(0,\dots,0,v(x_T)\big)$ is a function of x_1.

Lemma 4.1.4 *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 $\bar{v}_1 : \mathcal{X} \to \mathbb{R}$, $\|\bar{v}_1\|_w < \infty$, such that for all $x_1 \in \mathcal{X}$, and all $T \ge 1$*

$$\rho_{1,T}\big(0,\dots,0,v(x_T)\big) \le \bar{v}_1(x_1), \tag{4.5}$$

and

$$\big\|\bar{v}_1\big\|_w \le \Big\|\big(\widetilde{\mathfrak{M}}^{\pi}\big)^{T-1}\Big\|_w \cdot \big\|v\big\|_w, \tag{4.6}$$

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 (2.14), we obtain

$$\rho_{T-1}\big(v(x_T)\big) = \max_{m_{T-1}\in\mathfrak{M}^\pi_{x_{T-1}}} \int_{\mathcal{X}} v(y)\; m_{T-1}(dy). \tag{4.7}$$

It is a function of x_{T-1}, which we denote as $v_{T-1}(x_{T-1})$. Since $\|v\|_w < \infty$ and $w \in \mathcal{V}$, then $v \in \mathcal{V}$. As the sets $\mathfrak{M}^\pi_x$ are weakly* compact, the maximum in (4.7) 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

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

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., [47, 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

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

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

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

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 (2.6)). We obtain

$$\begin{aligned}v_{T-2}(x) &= \max_{M_{T-2}\ll\mathfrak{M}^\pi} \max_{M_{T-1}\ll\mathfrak{M}^\pi} \int_{\mathcal{X}} v(z)\; \big[M_{T-2}\circ M_{T-1}\big]\big(dz\big|x\big) \\ &\le \max_{M\ll(\mathfrak{M}^\pi)^2} \int_{\mathcal{X}} v(z)\; M\big(dz\big|x\big) = \bar{v}_{T-2}(x).\end{aligned}$$

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

$$\begin{aligned}\rho_1\Big(\rho_2\Big(\cdots\rho_{T-1}\big(v(x_T)\big)\cdots\Big)\Big) &\le \max_{M\ll(\mathfrak{M}^{\pi})^{T-1}}\int_{\mathcal{X}} v(z)\; M\big(dz\big|x_1\big)\\ &= \max_{\widetilde{M}\ll(\widetilde{\mathfrak{M}}^{\pi})^{T-1}}\int_{\widetilde{\mathcal{X}}} v(z)\;\widetilde{M}\big(dz\big|x_1\big).\end{aligned}$$

Denoting the right-hand side by $\bar{v}_1(x_1)$, we obtain the estimates (4.5)–(4.6). □

We can now provide sufficient conditions for the finiteness of the limit (4.1).

Theorem 4.1.5 *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* (4.1) *is finite and $\big\|J_{\infty}(\Pi,\cdot)\big\|_w < \infty$. If the model is uniformly risk-transient, then $\big\|J_{\infty}(\Pi,\cdot)\big\|_w$ is uniformly bounded.*

Proof By Lemma 1.3.3, 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 (4.2):

$$\begin{aligned}J_{T_2-1}(\Pi,x_1) &= \rho_{1,T_2}(0,Z_2,\dots,Z_{T_2})\\ &\le \rho_{1,T_2}(0,Z_2,\dots,Z_{T_1},0,\dots,0) + \sum_{j=T_1}^{T_2-1}\rho_{1,T_2}(0,\dots,0,Z_{j+1},0,\dots,0)\\ &= \rho_{1,T_1}(0,Z_2,\dots,Z_{T_1}) + \sum_{j=T_1}^{T_2-1}\rho_{1,j+1}(0,\dots,0,Z_{j+1}).\end{aligned} \tag{4.8}$$

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 $x \in \widetilde{\mathcal{X}}$, 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 $\mathcal{F}_j$-measurable, and in the last inequality – the subadditivity of the risk measures. Since $\|\bar{w}\|_w = 1$, Lemma 4.1.4 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

$$\|\bar{v}_j\|_w \le \Big\|\big(\widetilde{\mathfrak{M}}^\pi\big)^j\Big\|_w. \tag{4.9}$$

Substitution to (4.8) yields the estimate

$$\rho_{1,T_2}(0, Z_2, \dots, Z_{T_2}) \le \rho_{1,T_1}(0, Z_2, \dots, Z_{T_1}) + 2C\sum_{j=T_1-1}^{T_2} \bar{v}_j(x_1). \tag{4.10}$$

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 (4.10) applies to the new sequence. We obtain

$$\rho_{1,T_2}(0, Z_2, \dots, Z_{T_1}, -Z_{T_1+1}, \dots, -Z_{T_2}) \le \rho_{1,T_1}(0, Z_2, \dots, Z_{T_1}) + 2C\sum_{j=T_1-1}^{T_2} \bar{v}_j(x_1). \tag{4.11}$$

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

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

Substituting the estimate (4.11), 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 (4.10) yields

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

In view of (4.9), we conclude that

$$\big\|J_{T_2-1}(\Pi, \cdot) - J_{T_1-1}(\Pi, \cdot)\big\|_w \le 2C\sum_{j=T_1-1}^{T_2} \Big\|\big(\widetilde{\mathfrak{M}}^\pi\big)^j\Big\|_w. \tag{4.12}$$

By Definition 4.1.2, 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, $\big\|J_\infty(\Pi,\cdot)\big\|_w < \infty$. If the model is uniformly risk-transient, then the estimate (4.12) is same for all Markov policies Π, and thus $\big\|J_\infty(\Pi,\cdot)\big\|_w$ is uniformly bounded. □

Remark 4.1.6 *It is clear from the proof of Theorem 4.1.5, that*

$$J_\infty(\Pi, x_1) = \lim_{T\to\infty} \rho_{1,T}\big(0, Z_2, \dots, Z_T + f(x_T)\big), \tag{4.13}$$

for any measurable function $f : \mathcal{X} \to \mathbb{R}$, with $\big\|f\big\|_w < \infty$, 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 4.1.7 *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 $v : \mathcal{X} \to \mathbb{R}$, with $\|v\|_w < \infty$, satisfies the equations*

$$v(x) = \sigma\big(c_x + v, x, \pi(x) \circ Q_x\big), \quad x \in \widetilde{\mathcal{X}}, \tag{4.14}$$

$$v(x_{\mathrm{A}}) = 0, \tag{4.15}$$

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

Proof Denote $Z_t = c(x_{t-1}, u_{t-1}, x_t)$. Suppose a measurable function $v(\cdot)$ satisfies the dynamic programming equations (4.14)–(4.15). Since $\|v\|_w < \infty$ and $w \in \mathcal{V}$, then also $v \in \mathcal{V}$. By assumption, $c(\cdot,\cdot,\cdot)$ is w-bounded, and thus $c_x(\cdot,\cdot) \in \mathcal{V}$. Consequently, the right-hand side of (4.14) is well-defined. By iteration of (4.14), we obtain for all $x_1 \in \mathcal{X}$ the following equation:

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

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

$$\rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1} + v(x_{T+1})\big) \le \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1}\big) + \rho_{1,T+1}\big(0, 0, \dots, v(x_{T+1})\big). \tag{4.16}$$

By Lemma 4.1.4,

$$v(x_1) = \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1} + v(x_{T+1})\big) \le \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1})\big) + d_T(x_1), \quad (4.17)$$

with

$$\big\|d_T\big\|_w \le \Big\|\big(\widetilde{\mathfrak{M}}^{\pi}\big)^{T-1}\Big\|_w \cdot \big\|v\big\|_w. \quad (4.18)$$

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

$$\begin{aligned} &2\rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1}\big) \\ &\le \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1} + v(x_{T+1})\big) + \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1} - v(x_{T+1})\big) \\ &= v(x_1) + \rho_{1,T+1}\big(0, Z_2, \dots, Z_{T+1} - v(x_{T+1})\big). \end{aligned} \quad (4.19)$$

Similar to (4.16)–(4.17),

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

Substituting into (4.19) we obtain

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

Combining this estimate with (4.17) and using (4.18) we conclude that

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

Thus $v(\cdot) \equiv J_\infty(\mathit{\Pi}, \cdot)$, 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 (4.2) as follows:

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

The function $\rho_1(\cdot)$, as a finite-valued coherent measure of risk on a Banach lattice, is continuous (see [50, Prop. 3.1]). Since $\big\|J_T(\mathit{\Pi}, \cdot) - J_\infty(\mathit{\Pi}, \cdot)\big\|_w \to 0$, as $T \to \infty$, then the sequence $\big\{J_T(\mathit{\Pi}, \cdot)\big\}$ is also convergent in the space $\mathcal{V}$. Therefore,

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

This is identical with equation (4.14) with $v(\cdot) \equiv J_\infty(\mathit{\Pi}, \cdot)$. Equation (4.15) is obvious.

□

4.2 Dynamic Programming Equations for Infinite Horizon Problems

We shall now focus on the optimal value function

$$J^*(x) = \inf_{\Pi \in \Pi^{\mathrm{RM}}} J_\infty(\Pi, x), \quad x \in \mathcal{X}, \tag{4.20}$$

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

Theorem 4.2.1 *Assume that the following conditions are satisfied:*

(i) *For every $x \in \mathcal{X}$ the transition kernel $Q(x, \cdot)$ is continuous;*

(ii) *The conditional risk measures ρ_t, $t = 1, \dots, T$, are Markov and such that for every $x \in \mathcal{X}$ the multifunction $\mathcal{A}(x, \cdot)$ 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 $x \in \mathcal{X}$ the set $U(x)$ is compact;*

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

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

$$v(x) = \inf_{\lambda \in \mathcal{P}(U(x))} \sigma\big(c_x + v, x, \lambda \circ Q_x\big), \quad x \in \widetilde{X}, \tag{4.21}$$

$$v(x_{\mathrm{A}}) = 0, \tag{4.22}$$

if and only if $v(x) = J^*(x)$ for all $x \in \mathcal{X}$. Moreover, the minimizer $\pi^*(x)$, $x \in \mathcal{X}$, on the right hand side of (4.21) *exists and defines an optimal randomized Markov policy $\Pi^* = \{\pi^*, \pi^*, \dots\}$.*

Proof Suppose $J^*(\cdot)$ is given by (4.20). The set of policies of the form $\{\lambda, \pi, \pi, \dots\}$ is larger than Π^{RM}, and thus

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

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

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

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

$$J^*(x) \geq \inf_{\lambda \in \mathcal{P}(U(x))} \sigma\big(c_x + J^*, x, \lambda \circ Q_x\big), \quad x \in \mathcal{X}. \tag{4.23}$$

The mapping $\lambda \mapsto \sigma\big(c_x + J^*, x, \lambda \circ Q_x\big)$ is continuous for all x, and the set of $\lambda \in \mathcal{P}(\mathcal{U})$ such that $\lambda(U(x)) = 1$ is weakly* compact. Therefore, there exists a minimizer $\pi^*(x)$ on the right hand side of (4.23). Hence,

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

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

$$J^*(x_1) \geq \rho_{1,T}\big(0, Z_2, \dots, Z_T + J^*(x_T)\big), \tag{4.24}$$

with the sequence of controls and states resulting from the stationary Markov policy $\mathit{\Pi}^* = \{\pi^*, \pi^*, \dots\}$. Owing to Remark 4.1.6, we can pass to the limit on the right-hand side and obtain the inequality:

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

It follows that $\mathit{\Pi}^*$ is the optimal stationary Markov policy, and thus $J^*(\cdot) = J_\infty(\mathit{\Pi}^*, \cdot)$. By Theorem 4.1.7, relation (4.23) is an equation, which proves (4.21)–(4.22).

To prove the converse implication, suppose $v(\cdot)$ satisfies (4.21)–(4.22) 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 $\lambda \in \mathcal{P}(\mathcal{U})$ such that $\lambda(U(x)) = 1$, there exists a randomized control $\hat{\pi}(\cdot)$, which is a minimizer on the right hand side of (4.21). We obtain the equation

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

By Theorem 4.1.7,

$$v(x) = J_\infty(\hat{\mathit{\Pi}}, x) \geq J^*(x), \quad x \in \mathcal{X}, \tag{4.25}$$

where $\hat{\mathit{\Pi}} = \{\hat{\pi}, \hat{\pi}, \dots\}$. On the other hand, it follows from (4.21) that for the optimal policy $\mathit{\Pi}^* = \{\pi^*, \pi^*, \dots\}$ we have

$$v(x) \leq \sigma\big(c_x + v, x, \pi^*(x) \circ Q_x\big), \quad x \in \mathcal{X}. \tag{4.26}$$

The risk transition mapping σ is nondecreasing with respect to the first argument. Therefore, iterating inequality (4.26) we obtain an inequality corresponding to (4.24):

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

Passing to the limit with $T \to \infty$ and applying Remark 4.1.6, we conclude that

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

The last estimate together with (4.25) 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 4.2.2 *Assume that the conditions of Theorem 4.2.1 are satisfied, together with the following assumption: there exists a constant C such that $J_\infty(\Lambda, x) \ge -Cw(x)$ for all $x \in \mathcal{X}$ and for all policies Λ. Then a function $v : \mathcal{X} \to \mathbb{R}$, with $\|v\|_w < \infty$, satisfies the equations* (4.21)–(4.22) *if and only if $v(x) = \hat{J}(x)$ for all $x \in \mathcal{X}$. Moreover, the minimizer $\pi^*(x)$, $x \in \mathcal{X}$, on the right hand side of* (4.21) *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

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

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

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

Thus $\hat{J}(\cdot)$ satisfies an inequality analogous to (4.23):

$$\hat{J}(x) \geq \inf_{\lambda\in\mathcal{P}(U(x))} \sigma\big(c_x + \hat{J}, x, \lambda\circ Q_x\big), \quad x\in\mathcal{X}. \tag{4.27}$$

We can now repeat the argument from the proof of Theorem 4.2.1. Denoting by $\hat{\lambda}$ the minimizer above, iterating inequality (4.27), and passing to the limit we conclude that

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

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 4.2.1. □

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 $\mu\in\mathcal{A}(x,\mu)$, for all $x\in\mathcal{X}$ and $\mu\in\mathcal{M}$. Indeed, by virtue of Remark 2.2.5, the dynamic risk measure is bounded from below by the expected value of the cost, which is finite in this case.

4.3 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 (4.21), is nonlinear, in general, as opposed to the expected

value model, where

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

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

$$\inf_{\lambda \in \mathcal{P}(U(x))} \int\limits_{U(x)} \int\limits_{\mathcal{X}} \big(c(x,u,y) + v(y)\big)\; Q(dy|x,u)\; \lambda(du|x)$$

$$= \inf_{u \in U(x)} \int\limits_{\mathcal{X}} \big(c(x,u,y) + v(y)\big)\; Q(dy|x,u).$$

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 (4.21):

$$v(x) = \min_{\lambda \in \mathcal{P}^{\delta}(U(x))} \sigma\big(c_x + v, x, \lambda \circ Q_x\big), \quad x \in \mathcal{X}, \tag{4.28}$$

where $\mathcal{P}^{\delta}(U(x))$ denotes the set of Dirac measures supported at $U(x)$. For a fixed $x \in \mathcal{X}$ 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 $y \in \mathcal{X}$, and the measure $\lambda \circ Q_x$ is the measure $Q(\cdot|x,u)$ on the state space $\mathcal{X}$. We can, therefore, rewrite (4.28) in a simpler form

$$v(x) = \min_{u \in U(x)} \Big\{ c(x,u) + \sigma\big(v, x, Q(\cdot|x,u)\big) \Big\}, \quad x \in \mathcal{X}, \tag{4.29}$$

where (with a slight abuse of notation) $\sigma : \mathcal{L}_p(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_x) \times \mathcal{X} \times \mathcal{L}_q(\mathcal{X}, \mathcal{B}(\mathcal{X}), P_x) \to \mathbb{R}$, and $\sigma(\cdot,\cdot,\cdot)$ is a coherent measure of risk with respect to its first argument. In equation (4.29) we also used the translation property of coherent measures of risk. This is almost exactly the form of the dynamic programming equation which is derived in [48] 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 Conditional Average Value at Risk of Example 2.1.3:

$$\sigma(\varphi, x, \mu) = \inf_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{\alpha(x)} \int\limits_{U(x) \times \mathcal{X}} (\varphi(u,y) - \eta)_+ \; \mu(du \times dy) \right\}. \tag{4.30}$$

Lemma 4.3.1 *If the risk transition mapping has the form* (4.30) *then the dynamic programming equations* (4.21) *have a solution in deterministic decision rules.*

Proof Interchanging the integration and the infimum in the definition of Conditional Average Value at Risk, we obtain a lower bound

$$\begin{aligned}\sigma(\varphi, x, \lambda \circ Q_x) &= \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{\alpha(x)} \int_{U(x)} \int_{\mathcal{X}} (\varphi(u,y) - \eta)_+ \, Q(dy|x,u) \, \lambda(du|x) \right\} \\ &= \inf_{\eta\in\mathbb{R}} \int_{U(x)} \int_{\mathcal{X}} \left(\eta + \frac{1}{\alpha(x)} (\varphi(u,y) - \eta)_+ \right) Q(dy|x,u) \, \lambda(du|x) \\ &\geq \int_{U(x)} \inf_{\eta\in\mathbb{R}} \int_{\mathcal{X}} \left(\eta + \frac{1}{\alpha(x)} (\varphi(u,y) - \eta)_+ \right) Q(dy|x,u) \, \lambda(du|x).\end{aligned}$$

The above inequality becomes an equation for every Dirac measure λ. On the right-hand side of (4.21) we have

$$\begin{aligned}&\inf_{\lambda\in\mathcal{P}(U(x))} \sigma(c_x + v, x, \lambda \circ Q_x) \\ &\quad \geq \inf_{\lambda\in\mathcal{P}(U(x))} \int_{U(x)} \inf_{\eta\in\mathbb{R}} \int_{\mathcal{X}} \left(\eta + \frac{1}{\alpha(x)} \big(c(x,u,y) + v(y) - \eta\big)_+ \right) Q(dy|x,u) \, \lambda(du|x).\end{aligned}$$

As the right hand side achieves its minimum over $\lambda \in \mathcal{P}(U(x))$ 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 Conditional Average Value at Risk, deterministic Markov policies are optimal. □

Chapter 5

Value and Policy Iteration Methods for Infinite Horizon Problem

In this chapter, we suggest two iterative methods to solve the dynamic programming equations (4.21) and (4.22): *risk-averse value iteration* and *risk-averse policy iteration*. The solution of (4.22) is obvious, therefore we mainly aim to solve (4.21) using these methods. Throughout this chapter, we closely follow the notation and construction of [48].

The suggested methods are similar to the classical value iteration [5] and policy iteration methods [20] for the expected value case. However, the risk-averse policy iteration method is more complicated than the corresponding risk-neutral one since it requires solving a system of nonsmooth equations (reduces to a system of linear equations in the risk-neutral case) in order to evaluate the current policy. To solve these equations, we adopt and modify the *specialized nonsmooth Newton method* of Ruszczyński [48] proposed for risk-averse infinite horizon discounted models.

In the following two sections, we explain the methods and prove their convergence. We further show the global convergence of the Newton method.

Throughout this chapter, we assume that the infimum in equation (4.21) exists, therefore, it can be replaced with minimum.

5.1 Risk-Averse Value Iteration Method

In order to find the unique solution v^* of the dynamic programming equations (4.21) and (4.22), we adopt and extend the classical value iteration method of Bellman [5]. A similar method is also suggested in [48] for the dynamic programming equations corresponding to risk-averse infinite horizon discounted models with deterministic policies.

Let v^k be a certain approximation of v^*, then from equations (4.21) and (4.22), we obtain the following iterative method:

$$v^{k+1}(x) = \min_{\lambda \in \mathcal{P}(U(x))} \sigma\big(c_x + v^k, x, \lambda \circ Q_x\big), \quad x \in \widetilde{\mathcal{X}}, \quad k = 0, 1, 2, \ldots,$$

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

We provide the steps of this method in Algorithm 1.

Algorithm 1 Risk-Averse Value Iteration

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

The algorithm stops when the value functions do not change. However, in practice, approximate satisfaction of this stopping condition is required.

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

$$[\mathfrak{D}v](x) = \min_{\lambda \in \mathcal{P}(U(x))} \sigma(c_x + v, x, \lambda \circ Q_x), \quad x \in \widetilde{\mathcal{X}}, \tag{5.1}$$

$$[\mathfrak{D}_\pi v](x) = \sigma(c_x + v, x, \pi_x \circ Q_x), \quad x \in \widetilde{\mathcal{X}}, \tag{5.2}$$

where $\pi_x \in \mathcal{P}(U(x))$. To prove the convergence, we first provide the following two lemmas similar to Lemma 1 and Lemma 3 in [48].

Lemma 5.1.1 *For any φ and ψ in $\mathcal{V}$ such that $\varphi \geq \psi$, we have the relations*

$$\mathfrak{D}_\pi \varphi \geq \mathfrak{D}_\pi \psi \text{ and } \mathfrak{D}\varphi \geq \mathfrak{D}\psi.$$

Proof The proof is similar to the proof of Lemma 1 in [48], which we will provide here for completeness. From the dual representation (2.7), we have

$$[\mathfrak{D}_\pi v](x) = \max_{\mu \in \mathcal{A}(x, \pi_x \circ Q_x)} \langle c_x + v, \mu \rangle. \tag{5.3}$$

Since the elements of sets $\mathcal{A}(x, \pi_x \circ Q_x)$ are just probability measures, $\mathfrak{D}_\pi \varphi \geq \mathfrak{D}_\pi \psi$ for $\varphi \geq \psi$.

Suppose that the minimum in (5.1) is attained at π_φ and π_ψ for φ and ψ, respectively. Then, we get

$$\mathfrak{D}\varphi = \mathfrak{D}_{\pi_\varphi}\varphi \geq \mathfrak{D}_{\pi_\varphi}\psi \geq \mathfrak{D}_{\pi_\psi}\psi = \mathfrak{D}\psi.$$

□

Lemma 5.1.2 *Suppose that the controlled Markov model is uniformly risk-transient with $c(\cdot,\cdot,\cdot)$ and $\varphi(\cdot)$ being w-bounded*

(i) *if $\varphi \leq \mathfrak{D}\varphi$, then $\varphi \leq J^*$;*

(ii) *if $\varphi \geq \mathfrak{D}\varphi$, then $\varphi \geq J^*$.*

Proof We will follow the proof of Lemma 3 in [48] with some adjustments.

(i) If $\varphi \leq \mathfrak{D}\varphi$, then for any π such that $\pi(x) \in \mathcal{P}(U(x)),\ x \in \mathcal{X}$, we have

$$\varphi \leq \mathfrak{D}\varphi \leq \mathfrak{D}_\pi \varphi. \tag{5.4}$$

If we apply the operator $\mathfrak{D}_\pi$ to relation (5.4), then from the monotonicity property stated in Lemma 5.1.1, we obtain the following relation

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

Proceeding in this way, we get

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

Let the Markov policy $\Pi = \{\pi, \pi, \ldots\}$ result in the cost sequence $Z_t = c(x_{t-1}, u_{t-1}, x_t)$, $t = 2, 3, \ldots$. From the equality (5.2), it is clear that the right hand side of the last displayed inequality is equivalent to (3.2) with $v_{T+1} = \varphi$ and $\Pi = \{\pi, \pi, \ldots, \pi\}$. Then, we get

$$\varphi(x_1) \leq \big[[\mathfrak{D}_\pi]^T\varphi\big](x_1) = \rho_{1,T+1}\big(0, Z_2, \ldots, Z_{T+1} + \varphi(x_{T+1})\big).$$

Passing to the limit with $T \to \infty$ and using Remark 4.1.6, we conclude that

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

Since above inequality is correct for any stationary Markov policy $\Pi = \{\pi, \pi, \ldots\}$, then $\varphi \leq J^*$.

(ii) If $\varphi \geq \mathfrak{D}\varphi$, then there exists a π such that $\pi(x) \in \mathcal{P}(U(x)),\ x \in \mathcal{X}$ giving that

$$\varphi \geq \mathfrak{D}_\pi \varphi = \mathfrak{D}\varphi. \tag{5.5}$$

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

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

Similar to the proof of part (i)

$$\varphi(x_1) \geq \big[[\mathfrak{D}_\pi]^T \varphi\big](x_1) = \rho_{1,T+1}\big(0, Z_2, \ldots, Z_{T+1} + \varphi(x_{T+1})\big). \tag{5.6}$$

If we pass to the limit with $T \to \infty$ in (5.6), again from Remark 4.1.6, we obtain

$$\varphi(x) \geq J_\infty(\Pi, x) \geq J^*(x), \quad x \in \mathcal{X}.$$

□

Theorem 5.1.3 *Suppose the conditions of Theorem 4.2.2 are satisfied:*

(i) *if $c(x, u, y) \geq 0$ for all $x, y \in \mathcal{X}$ and $u \in U(x)$, then for $v^0 = 0$, the sequence $\{v_k\}$ obtained by the value iteration method is nondecreasing and convergent to the unique solution v^* of* (4.21) *and* (4.22).

(ii) *if $c(x, u, y) \leq 0$ for all $x, y \in \mathcal{X}$ and $u \in U(x)$, then for $v^0 = 0$, the sequence $\{v_k\}$ is nonincreasing and converges to v^*.*

Proof (i) Owing to the monotonicity property (B2) and the fact that $c(x, u, y) \geq 0$, $v^0 \leq \mathfrak{D}v^0$ for $v^0 = 0$. From Lemma 5.1.1 and Lemma 5.1.2, we obtain

$$v^* \geq v^{k+1} \geq v^k, \quad k = 0, 1, 2, \ldots.$$

We have a nondecreasing and bounded sequence which is thus convergent to some limit v^∞. Since the operator $\mathfrak{D}$ is continuous, letting $k \to \infty$ in the equation $v^{k+1} = \mathfrak{D}v^k$, we obtain $v^\infty = \mathfrak{D}v^\infty$. This implies that $v^\infty = v^*$, the sequence $\{v^k\}$ converges to v^*.

(ii) Similarly, if $c(x,u,y) \leq 0$, then $v^0 \geq Dv^0$ for $v^0 = 0$ and the sequence $\{v_k\}$ is nonincreasing:

$$v^k \geq v^{k+1} \geq v^*, \quad k = 0,1,2,\ldots.$$

Using the same argument in the proof of (i), we deduce that $v^\infty = v^*$. □

5.2 Risk-Averse Policy Iteration Method

As an alternative way for solving the dynamic programming equations (4.21) and (4.22), we suggest the risk-averse policy iteration method which is analogous to the classical policy iteration method of Howard [20]. A similar approach is proposed in [48] for discounted infinite horizon problems with the feasible set being restricted to deterministic policies.

The steps of our method are explained in Algorithm 2.

Algorithm 2 Risk-Averse Policy Iteration

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

For a stationary policy $\Pi^k = \{\pi^k, \pi^k, \ldots\}$, the *policy evaluation step* solves the

following system of equations to find $J_\infty(\Pi^k, x) = v^k(x),\ x \in \mathcal{X}$:

$$v(x) = \sigma(c_x + v, x, \pi_x^k \circ Q_x), \quad x \in \widetilde{\mathcal{X}}, \tag{5.7}$$

$$v(x_{\mathrm{A}}) = 0. \tag{5.8}$$

Then the *policy improvement step* finds a new decision rule $\pi^{k+1}(.)$ if it provides an improvement in the value function:

$$\pi_x^{k+1} \leftarrow \underset{\lambda \in \mathcal{P}(U(x))}{\operatorname{argmin}}\ \sigma\big(c_x + v^k, x, \lambda \circ Q_x\big), \quad x \in \widetilde{\mathcal{X}}, \tag{5.9}$$

and these steps are repeated until the value function does not change. In practice, an approximate satisfaction of the stopping condition is required.

Let the operators $\mathfrak{D}$ and $\mathfrak{D}_\pi$ be defined as (5.1) and (5.2), respectively. Then (5.7) can be equivalently written as

$$v^k = \mathfrak{D}_{\pi^k} v^k. \tag{5.10}$$

Similarly, (5.9) is equivalent to

$$\mathfrak{D}_{\pi^{k+1}} v^k = \mathfrak{D} v^k. \tag{5.11}$$

Theorem 5.2.1 *Suppose the conditions of Theorem 4.2.2 are satisfied. Then for any* π^0 *such that* $\pi^0(x) \in \mathcal{P}(U(x)),\ x \in \mathcal{X}$, *the sequence* $\{v^k\}$ *obtained by the policy iteration method is nonincreasing and convergent to the unique solution* v^* *of* (4.21) *and* (4.22).

Proof We will follow the proof of Theorem 6 in [48] with some adjustments due to allowing randomized policies and not using a discount factor. Using the equations (5.10) and (5.11), we obtain

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

Applying the operator $\mathfrak{D}_{\pi^{k+1}}$ to above relation, from the monotonicity property given in Lemma 5.1.1, we deduce that

$$[\mathfrak{D}_{\pi^{k+1}}]^T v^k \le \mathfrak{D}_{\pi^{k+1}} v^k = \mathfrak{D} v^k \le v^k, \quad T = 1, 2, \ldots. \tag{5.12}$$

Relation (5.12) can be equivalently written as

$$\rho_{1,T+1}\big(0, Z_2, \ldots, Z_{T+1} + v^k(x_{T+1})\big) \le [\mathfrak{D} v^k](x_1) \le v^k(x_1),$$

where $Z_t = c(x_{t-1}, u_{t-1}, x_t),\ t = 2, 3, \ldots, T+1$ 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 Remark 4.1.6 and Theorem 4.1.7, we conclude that the sequence $\{v^k\}$ is nonincreasing:

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

Since $v^*(x) \leq J_\infty(\Pi^{k+1}, x)$, it is trivial that $v^k \geq v^*$. Let $\varphi^k = \mathfrak{D}_{\pi^{k+1}} v^k$, then the relation (5.12) states that

$$\varphi^k = \mathfrak{D}v^k \leq v^k. \tag{5.13}$$

As the operator $\mathfrak{D}$ is continuous, passing to the limit with $k \to \infty$ in the relation (5.13), we obtain

$$\varphi^\infty = \mathfrak{D}v^\infty \leq v^\infty. \tag{5.14}$$

Following the argument stated above, we deduce that

$$v^{k+1} = \lim_{T\to\infty} [\mathfrak{D}_{\pi^{k+1}}]^T v^k = \lim_{T\to\infty} [\mathfrak{D}_{\pi^{k+1}}]^{T-1} \varphi^k \leq \varphi^k,$$

where the right hand side inequality comes from relation (5.12). Passing to the limit with $k \to \infty$, we get $v^\infty \leq \varphi^\infty$. Combining this with relation (5.14), we state that $\varphi^\infty = v^\infty$, the algorithm converges to a unique limit function. Furthermore, replacing φ^∞ with v^∞ in (5.14), we get $v^\infty = v^*$. □

5.2.1 Specialized Nonsmooth Newton Method

In the evaluation step of the policy iteration method, we have to solve a system of nonlinear equations (5.7), which is nonsmooth for all risk mappings, except for the expected value mapping. In order to solve this system of equations, we adopt the *specialized nonsmooth Newton method* of Ruszczyński [48] which uses the idea of the nonsmooth Newton method with linear auxiliary problems (see [27, Section 10.1] and [29] for details).

To find the unique solution of (5.7) with $v(x_\mathrm{A}) = 0$, we will solve iteratively an appropriate linear approximation of it. Using the dual representation (2.7), (5.7) can be equivalently written as

$$v(x) = \max_{\mu(x) \in \mathcal{A}(x, \pi_x^k \circ Q_x)} \langle c_x + v, \mu(x) \rangle, \quad x \in \widetilde{\mathcal{X}}. \tag{5.15}$$

Let v_l be an approximate solution of (5.15) at iteration l of the nonsmooth Newton method. Then, we find a kernel μ_l by solving

$$\mu_l(x) \in \underset{\mu(x)\in\mathcal{A}(x,\pi_x^k\circ Q_x)}{\operatorname{argmax}} \langle c_x + v_l, \mu(x)\rangle, \quad x \in \mathcal{X}. \tag{5.16}$$

The maximum in equation (5.16) is attained because the set $\mathcal{A}$ is bounded, convex, and closed, and the function being maximized is linear. If plug in μ_l in (5.15), then we obtain the following system of linear equations

$$v(x) = \langle c_x + v, \mu_l(x)\rangle, \quad x \in \widetilde{\mathcal{X}}, \tag{5.17}$$

which suggests a computational recipe for solving (5.7) (see Algorithm 3).

Algorithm 3 Specialized Nonsmooth Newton Method

procedure NONSMOOTHNEWTONMETHOD(v_0)
  $l \leftarrow 0$
  repeat
    $\mu_{l+1}(x) \leftarrow \underset{\mu(x)\in\mathcal{A}(x,\pi_x^k\circ Q_x)}{\operatorname{argmax}} \langle c_x + v_l, \mu(x)\rangle, \quad x \in \mathcal{X}$
    $v(x_{\mathrm{A}}) \leftarrow 0$
    Solve $v(x) = \langle c_x + v, \mu_l(x)\rangle, \quad x \in \widetilde{\mathcal{X}}$ for v
    $v_{l+1} \leftarrow v$
    $l \leftarrow l + 1$
  until $v_l = v_{l-1}$
  return v_l
end procedure

We will show that the sequence $\{v_l\}$, $l = 1, 2, \ldots$ obtained by this method converges to the unique solution of (5.7). But, first of all, we need to provide some technical results.

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

$$[\mathfrak{R}_l v](x) = \langle c_x + v, \mu_l(x)\rangle.$$

It is clear that the equation (5.17) can be equivalently written as $v(x) = [\mathfrak{R}_l v](x)$. We will also use $\mathbb{E}_l$ to denote the expected value with respect to the substochastic kernel μ_l.

Lemma 5.2.2 *Equation* (5.17) *has a unique solution* $v_{l+1}(x_1) = \mathbb{E}_l[\sum_{t=1}^{\infty} c_{x_t}]$. *And, for any* v_0, $[\mathfrak{R}_l]^T v_0$ *converges to* v_{l+1} *as* $T \to \infty$.

Proof Following the proofs of Lemma 9.4.8 and Proposition 9.5.11 in [19], we will show that (5.17) is satisfied if and only if $v(x_1) = \mathbb{E}_l[\sum_{t=1}^{\infty} c_{x_t}]$. Iterating (5.17), we obtain

$$v(x_1) = [\mathfrak{R}_l^T v](x_1) = \mathbb{E}_l\Big[\sum_{t=1}^{T} c_{x_t}\Big] + \mathbb{E}_l[v(x_{T+1})]. \tag{5.18}$$

From (2.13), we deduce that

$$\mathbb{E}_l[v(x_{T+1})] = \int_{\widetilde{X}} v(y) M^T(dy|x_1), \tag{5.19}$$

where $M(x) = L\mu_l(x)$. Definition 4.1.2 together with the norm definition of the stochastic multikernel give that

$$\sum_{j=1}^{\infty} \|(M)^j\|_w \leq K, \tag{5.20}$$

which, trivially, implies that $(M)^j \to 0$ as $j \to \infty$.

If we pass to the limit with $T \to \infty$ in (5.18), then it follows from (5.19) and (5.20) that $\mathbb{E}_l[v(x_{T+1})] \to 0$. Therefore, if (5.17) is satisfied, then $v(x_1) = \mathbb{E}_l[\sum_{t=1}^{\infty} c_{x_t}]$. To show the other side, we can easily check that $v(x_1) = \mathbb{E}_l[\sum_{t=1}^{\infty} c_{x_t}]$ satisfies the equation (5.17), which can also be written as:

$$v(x_1) = \mathbb{E}_l[c_{x_1}] + \mathbb{E}_l[v(x_2)].$$

Above equation is equivalent to

$$\begin{aligned} v(x_1) &= \mathbb{E}_l[c_{x_1}] + \mathbb{E}_l\Big[\mathbb{E}_l\Big[\sum_{t=2}^{\infty} c_{x_t}|x_1, u_1, x_2\Big]\Big] \\ &= \mathbb{E}_l[c_{x_1}] + \mathbb{E}_l\Big[\sum_{t=2}^{\infty} c_{x_t}\Big] = \mathbb{E}_l\Big[\sum_{t=1}^{\infty} c_{x_t}\Big]. \end{aligned} \tag{5.21}$$

Additionally, it follows from (5.18) that $[\mathfrak{R}_l]^T v_0 \to v_{l+1}$ as $T \to \infty$. □

Theorem 5.2.3 *For any initial v_0, the sequence $\{v_l\}$ obtained by the Newton method is nondecreasing and convergent to the unique solution $\hat{v}$ of* (5.7).

Proof The proof follows in a way similar to the proof of Theorem 7 in [48]. By definition, we have

$$\mathfrak{R}_l v \leq \mathfrak{D}_{\pi^k} v. \tag{5.22}$$

The operator $\mathfrak{R}_l$ is monotone owing to the fact that $\mu_l(x)$, $x \in \mathcal{X}$ are probability measures. Therefore, if we apply the operator $\mathfrak{R}_l$ to inequality (5.22), we obtain

$$[\mathfrak{R}_l]^T v \leq [\mathfrak{D}_{\pi^k}]^T v, \quad T = 1, 2, \ldots. \tag{5.23}$$

Passing to the limit with $T \to \infty$, from Lemma (5.2.2), we deduce that the left hand side of (5.23) converges to v_{l+1}. Moreover, the right hand side converges to the unique solution $\bar{v}$ of (5.15). Therefore, we get that $v_{l+1} \leq \bar{v}$, the sequence $\{v_{l+1}\}$ is bounded from above. We will show that it is also nondecreasing.

For every x, we have

$$v_l(x) = \langle c_x + v_l, \mu_{l-1}(x) \rangle \leq \max_{\mu(x) \in \mathcal{A}(x, \pi^k_x \circ Q_x)} \langle c_x + v_l, \mu(x) \rangle = [\mathfrak{D}_{\pi^k} v_l](x) = [\mathfrak{R}_l v_l](x).$$

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

$$v_l \leq \mathfrak{D}_{\pi^k} v_l \leq [\mathfrak{R}_l]^T v_l, \quad T = 1, 2, \ldots. \tag{5.24}$$

Passing to the limit with $T \to \infty$, the right hand side converges to v_{l+1}. Therefore, we get

$$v_l \leq \mathfrak{D}_{\pi^k} v_l \leq v_{l+1}. \tag{5.25}$$

The sequence $\{v_l\}$ is nondecreasing. Since it is also bounded from above, it has a limit. Let $\hat{v} = \lim_{l \to \infty} v_l$. If we pass to the limit with $l \to \infty$ in (5.25), we obtain $\hat{v} = \mathcal{D}_{\pi^k} \hat{v}$, $\hat{v}$ is the unique solution of (5.7). $\square$

Chapter 6

Mathematical Programming Approach for Infinite Horizon Problem

If the state and control spaces are finite, then, in addition to the iterative methods described in Chapter 5, mathematical programming approach can also be used to solve infinite horizon Markov decision problems. These models will be linear for the expected value case (for details, see [11], [24], [25], [32], [45], and the references therein). However, since the risk transition mapping $\big(v, \pi_x\big) \to \sigma(c_x + v, x, \pi_x \circ Q_x)$ is nonlinear in general, we expect to have nonlinear models for the risk-averse problems.

Throughout this chapter, we will assume that both the state and control spaces are finite. In section 6.1, we will derive the mathematical formulation for the problem with randomized policies, whereas in section 6.2, we will assume that the feasible set is restricted to deterministic policies.

6.1 Randomized Policies

From the dynamic programming equation (4.21), we obtain that

$$v(x) \leq \sigma\big(c_x + v, x, \pi_x \circ Q_x\big), \quad x \in \widetilde{X}, \tag{6.1}$$

for any randomized decision rule π. Then, the unique solution $v(\cdot) = J^*(\cdot)$ of the equations (4.21) and (4.22) can be found by solving the following optimization problem:

$$\begin{aligned}
\textbf{(P1)} \quad \max \quad & \langle \mathbf{1}, v \rangle && (6.2)\\
\text{s.t.} \quad & v(x) \leq \sigma\big(c_x + v, x, \pi_x \circ Q_x\big), \quad x \in \widetilde{X}, && (6.3)\\
& v(x_{\mathrm{A}}) = 0, && (6.4)\\
& \langle \mathbf{1}, \pi_x \rangle = 1, \quad x \in \mathcal{X}, && (6.5)\\
& \pi_x \geq 0, \quad x \in \mathcal{X}. && (6.6)
\end{aligned}$$

Here, $\mathbf{1}$ is a column vector with all entries being equal to 1, v and $\pi_x,\ x \in \mathcal{X}$ are the decision vectors. Throughout this chapter, we will assume that $\langle \cdot, \cdot \rangle$ denotes the usual scalar product. It follows from the dual representation (2.7) that the constraint (6.3) can be replaced by the following inequality:

$$v(x) \leq \max_{\mu(x) \in \mathcal{A}(x, \pi_x \circ Q_x)} \langle c_x + v, \mu(x) \rangle, \quad x \in \widetilde{X}, \tag{6.7}$$

with $\mathcal{A}(x, \pi_x \circ Q_x)$ being a convex set for every $x \in \mathcal{X}$ and $\pi(x) \in \mathcal{P}(U(x))$. Note that, if $\mu^*(x)$ is the maximizer of the right hand side of (6.7), then we have

$$v(x) \leq \langle c_x + v, \mu^*(x) \rangle, \quad x \in \widetilde{X}. \tag{6.8}$$

Using this fact, Problem (P1) can be equivalently formulated as below:

$$\begin{aligned}
\textbf{(P2)} \quad \max \quad & \langle \mathbf{1}, v \rangle && (6.9)\\
\text{s.t.} \quad & v(x) \leq \langle c_x + v, \mu(x) \rangle, \quad x \in \widetilde{X}, && (6.10)\\
& v(x_{\mathrm{A}}) = 0, && (6.11)\\
& \langle \mathbf{1}, \pi_x \rangle = 1, \quad x \in \mathcal{X}, && (6.12)\\
& \pi_x \geq 0, \quad x \in \mathcal{X}, && (6.13)\\
& \mu(x) \in \mathcal{A}(x, \pi_x \circ Q_x), \quad x \in \widetilde{X}. && (6.14)
\end{aligned}$$

Constraint (6.10) is nonconvex. Furthermore, if π_x is not fixed, then the set $\mathcal{A}(x, \pi_x \circ Q_x)$ will be nonconvex (see Example 6.1.1) for any $x \in \mathcal{X}$, in general. Therefore, Problem (P2) is nonconvex.

Example 6.1.1 Consider the first-order mean-semideviation risk measure of Examples 2.1.2 and 2.2.3. Using (2.9), for finite state and control spaces, we get

$$\begin{aligned}
\mathcal{A}(x, \pi_x \circ Q_x) = \Big\{ & g \in \mathcal{M} : g(x,u,y) \\
& = \pi_x(u) Q(x,u,y) \big(1 + h(x,u,y) - \sum_z h(x,u,z) \pi_x(u) Q(x,u,z)\big), \\
& 0 \leq h(x,u,y) \leq \kappa(x), y \in \mathcal{X}, u \in U(x) \Big\}. \qquad (6.15)
\end{aligned}$$

Let π_x^1 and $h^1(x,.,.)$ define $\mu^1(x)$ such that $\mu^1(x) \in \mathcal{A}(x, \pi_x^1 \circ Q_x)$. Similarly, π_x^2 and $h^2(x,.,.)$ give $\mu^2(x) \in \mathcal{A}(x, \pi_x^2 \circ Q_x)$. Taking a convex combination of $(\mu^1(x), \pi_x^1, h^1(x,.,.))$

and $(\mu^2(x), \pi_x^2, h^2(x,.,.))$ with $\theta > 0$, we obtain

$$(\overline{\mu}(x), \overline{\pi}_x, \overline{h}(x,.,.)) = \theta(\mu^1(x), \pi_x^1, h^1(x,.,.)) + (1-\theta)(\mu^2(x), \pi_x^2, h^2(x,.,.)).$$

It can be easily checked that

$$\overline{\mu}(x,u,y) \neq \overline{\pi}_x(u)Q(x,u,y)\big(1 + \overline{h}(x,u,y) - \sum_z \overline{h}(x,u,z)\overline{\pi}_x(u)Q(x,u,z)\big).$$

This gives that $\overline{\mu}(x) \notin \mathcal{A}(x, \overline{\pi}_x \circ Q_x)$, therefore, the set $\mathcal{A}(x, \pi_x \circ Q_x)$ is not convex if π_x is not fixed.

6.2 Deterministic Policies

For deterministic policies, the dynamic programming equation (4.21) gets the form

$$v(x) = \min_{u \in U(x)} \sigma\big(c_x + v, x, Q(x,u)\big), \quad x \in \widetilde{X}. \tag{6.16}$$

Following the arguments of previous section, we obtain the mathematical formulation (P3).

$$\begin{aligned}
\textbf{(P3)} \quad \max \quad & \langle \mathbf{1}, v \rangle && (6.17)\\
\text{s.t.} \quad & v(x) \leq \langle c_x + v, \mu(x,u) \rangle, \quad x \in \widetilde{X}, u \in U(x), && (6.18)\\
& v(x_{\mathrm{A}}) = 0, && (6.19)\\
& \mu(x,u) \in \mathcal{A}(x, Q(x,u)), \quad x \in \widetilde{X}, u \in U(x), && (6.20)
\end{aligned}$$

with the set $\mathcal{A}(x, Q(x,u))$ being convex for every $x \in \widetilde{X}$ and $u \in U(x)$. We will show that for the first-order mean-semideviation (Example 2.1.2 and 2.2.3) and conditional average value at risk (Example 2.1.3 and 2.2.4) measures, constraint (6.20) will be linear.

Example 6.2.1 For the first-order mean-semideviation risk measure, it follows from (2.9) that constraint (6.20) has the following linear form:

$$\begin{aligned}
&\mu(x,u,y) = Q(x,u,y)(1 + h(x,u,y) - \langle h(x,u), Q(x,u) \rangle), \quad x \in \widetilde{X}, u \in U(x), y \in \mathcal{X},\\
&0 \leq h(x,u,y) \leq \kappa(x), \quad x \in \widetilde{X}, u \in U(x), y \in \mathcal{X},\\
&\langle \mathbf{1}, \mu(x,u) \rangle = 1, \quad x \in \widetilde{X}, u \in U(x).
\end{aligned}$$

Example 6.2.2 For the conditional average value at risk measure, using (2.10), we obtain the following linear form for (6.20):

$$0 \leq \mu(x,u) \leq \frac{Q(x,u)}{\alpha(x)}, \quad x \in \widetilde{X}, u \in U(x),$$

$$\langle \mathbf{1}, \mu(x,u) \rangle = 1, \quad x \in \widetilde{X}, u \in U(x).$$

Chapter 7

Illustrative Examples

In the following three sections, we illustrate our models and results on three simple examples. We first consider the classical asset selling problem originating from Karlin [26]. In this example, we restrict the feasible policies to be deterministic and derive the structure of the optimal policy. The second example is a simple organ transplant problem inspired from Alagoz et al. [1]. We show that, for a risk-averse patient, a randomized policy may be better than a deterministic policy. Finally, we solve a credit card example with pure policies, which is a simplified version of the problem studied by So and Thomas [54].

7.1 Asset Selling Problem

In this section, we consider an asset selling problem where random offers Y_t arrive in time periods $t = 1, 2, \dots$. We assume that Y_t are independent and identically distributed, integrable random variables coming from a discrete distribution. In each time period, we select one of the decisions "sell" corresponding to accepting the highest offer received so far and quitting the process or "wait" corresponding to waiting in hope of a better offer. A positive observation cost c is incurred in each period. Therefore, if the process stops at a random time τ, then the total cost incurred will be $Z = c\tau - \max_{0\leq j\leq\tau} Y_j$. This problem is a simple and classical example of an *optimal stopping problem* (see, e.g., Çinlar [10], Dynkin and Yushkevich [13, 14], and Puterman [45]).

Sticking to our notation, we introduce the state space $\mathcal{X} = \{x_{\mathrm{A}}\}\cup\{0, 1, 2, \dots\}$, where x_{A} is the absorbing state reached after the decision "sell," and the other states represent the highest offer received so far. The control space is $\mathcal{U} = \{0, 1\}$, with 0 representing

"wait" and 1 representing "sell." The states evolve according to the equation

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

The above formula also defines the transition kernel Q:

$$Q(x_{t+1} \leq a | x_t, u_t = 0) = \begin{cases} 0 & \text{if } a < x_t, \\ P\{Y_{t+1} \leq a\} & \text{if } a \geq x_t. \end{cases}$$

Here P denotes the probability and a is some constant. If $u_t = 1$, then the transition will be to the absorbing state x_{A} with probability one. And, the cost is

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

It is known that, the optimal solution of the expected value version of this problem is obtained by solving the following equation:

$$c = \mathbb{E}\big[(Y - x)_+\big]. \tag{7.1}$$

Let $\hat{x}$ be the solution of (7.1), then the optimal policy is to accept the first offer that is greater than or equal to $\hat{x}$.

Using our theory, we will solve the risk-averse version of the problem. We will restrict our consideration to deterministic policies and derive the structure of the optimal policy.

For this example, equation (4.29) takes the form:

$$v(x) = \min\Big\{ -x, c + \sigma\big(v, x, Q_x\big)\Big\}, \quad x \in \widetilde{\mathcal{X}}, \tag{7.2}$$

where σ is a stationary risk transition mapping.

We assume that σ is law invariant (this concept is defined in Definition 2.1.4). Notice that, 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. Then from Definition 2.1.4, we obtain

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

Suppose that the risk transition mapping σ does not depend on its second argument. This means that our attitude to risk does not depend on the current state x. Then, using the dual representation (2.7), we may rewrite the last equation as follows:

$$\sigma\big(v, x, Q_x\big) = \max_{\mu\in\mathcal{A}} \mathbb{E}_\mu\big[v\big(\max(x, Y)\big)\big]. \tag{7.3}$$

Since σ is not dependent on x, the closed convex set of probability measures $\mathcal{A}$ is fixed for any $x \in \mathcal{X}$. From (7.3), equation (7.2) takes on the form

$$v(x) = \min\Big\{ -x, c + \max_{\mu\in\mathcal{A}} \mathbb{E}_\mu\big[v\big(\max(x, Y)\big)\big]\Big\}, \quad x \in \widetilde{\mathcal{X}}. \tag{7.4}$$

Observe that $v(x) \le -x$ and thus $v\big(\max(x, Y)\big) \le -\max(x, Y)$. The last displayed inequality combined with (7.4) implies that

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

If at state x, the optimal policy is to accept the offer x, then $v(x) = -x$. This gives the following relation:

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

It is clear that $\max(x, Y) = x + (Y - x)_+$. After elementary simplifications, we obtain

$$\min_{\mu\in\mathcal{A}} \mathbb{E}_\mu\big[(Y - x)_+\big] \le c. \tag{7.5}$$

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

$$\min_{\mu\in\mathcal{A}} \mathbb{E}_\mu\big[(Y - x^*)_+\big] = c; \tag{7.6}$$

if $x < x^*$, then wait.

The corresponding value function will be:

$$v^*(x) = -\max(x, x^*). \tag{7.7}$$

Equation (7.4) can be verified by direct substitution. For $x \le x^*$, if we substitute (7.7) in (7.4), we obtain

$$\begin{aligned}
&\min\Big\{ -x, c + \max_{\mu\in\mathcal{A}} \mathbb{E}_\mu\Big[-\max\big(\max(x, Y), x^*\big)\Big]\Big\} \\
&= \min\Big\{ -x, c - \min_{\mu\in\mathcal{A}} \mathbb{E}_\mu\big[(Y - x^*)_+\big] - x^*\Big\} = -\max(x, x^*) = -x^*.
\end{aligned}$$

The second equality follows from equation (7.6). For $x > x^*$, in a similar way, using the inequality (7.5), we get

$$\begin{aligned}&\min\Big\{-x, c+\max_{\mu\in\mathcal{A}}\mathbb{E}_\mu\Big[-\max\big(\max(x,Y),x^*\big)\Big]\Big\}\\&=\min\Big\{-x, c-\min_{\mu\in\mathcal{A}}\mathbb{E}_\mu\big[(Y-x)_+\big]-x\Big\}=-x=-\max(x,x^*).\end{aligned}$$

Observe that the solution (7.6) of the risk-averse problem is closely related to the solution (7.1) of the expected value problem. The only difference is that we have to account for the least favorable distribution of the offers. Therefore, if $P_Y \in \mathcal{A}$, then the critical level $x^* \leq \hat{x}$.

7.2 Organ Transplant Problem

In this section, we illustrate our theory on a risk-averse and simplified version of the organ transplant problem discussed in Alagoz et. al. [1]. Similar to Alagoz et. al. [1], we assume that the organ is transferred from a living-donor. However, we do not consider different stages of sickness and combine them in one state denoted by S.

We consider the discrete-time, absorbing Markov chain depicted in Figure 7.1. State S is the initial state and represents a patient in need of an organ transplant. State L represents life after a successful transplant. State D is an absorbing state representing death.

In state S, two controls (decisions) are possible for the patient: "Wait" (denoted by W) or "Transplant" (denoted by T). Under the control W, the transitions from state S are either to state D or back to state S. T 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. In other two states only one (formal) control 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 zero.

Generally, in a cost minimization problem, the value of a dynamic measure of risk

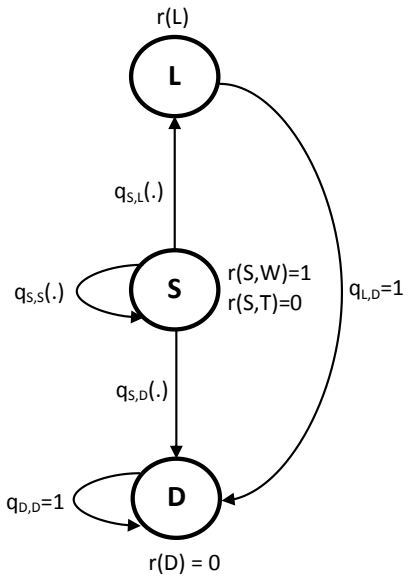

Figure 7.1: The organ transplant model.

(1.5) 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 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 $\lambda = (\lambda_{\mathrm{W}}, \lambda_{\mathrm{T}})$ be the randomized policy in the state S and let $\Lambda = \big\{\lambda \in \mathbb{R}^2 : \lambda_{\mathrm{W}} + \lambda_{\mathrm{T}} = 1,\ \lambda \geq 0\big\}$. We use the first order mean–semideviation risk mapping of Example 2.1.2. For $\kappa = 1$, the dynamic programming equation (4.21) at S takes on the form

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

In the semideviation parts, we wrote ψ 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.

7.2.1 The Survival Model

We will describe 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 successful transplant, as depicted in Figure 7.2.

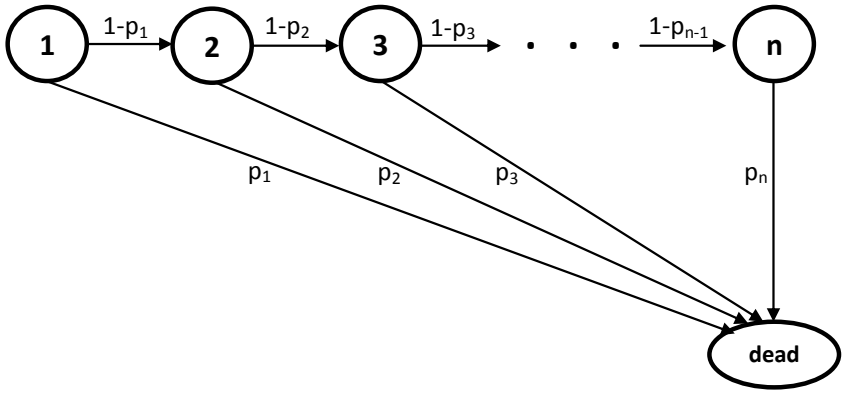

Figure 7.2: The survival model.

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 our notation, we define the cost $c(\cdot) = -r(\cdot)$. For illustration, we apply the mean–semideviation model of Example 2.1.2 with $\kappa = 1$.

The risk transition mapping has the form:

$$\sigma(\varphi, i, \nu) = \underbrace{\mathbb{E}_\nu[\varphi]}_{\text{expected value}} + \kappa \underbrace{\mathbb{E}_\nu\big[\big(\varphi - \mathbb{E}_\nu[\varphi]\big)_+\big]}_{\text{semideviation}}. \tag{7.8}$$

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

In (7.8), 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 (3.3). As there is no control to choose, the minimization operation 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, $v_i(i) \le 0$, for $i = n-1, n-2, \ldots, 1$.

Let us calculate the mean and semideviation components of (7.8) at states $i = 1, \ldots, n-1$:

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

In the last equation we used the fact that $v_{i+1}(i+1) \le 0$. For $i = 1, \ldots, n-1$, the dynamic programming equations (3.3) 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 deterministic equivalent length of life $r(\mathrm{L})$. For $\kappa = 0$ the above formulas give the negative of the expected length of life with new organ.

7.2.2 Numerical Illustration

In our calculations we used the transition data provided in Table 7.1. 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 7.1: Transition probabilities from state S.

For the survival model, we used the distribution function, $F(x)$, of lifetime of the American population from Jasiulewicz [23]. 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 [23], are given in Table 7.2.

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 7.2: 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 $r(\mathrm{L})$.

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 $r(\mathrm{L}) = 610.46$.

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. Such a result is not possible in an expected value model, no matter what the data are. Because an optimal policy in an expected value model is always a deterministic policy: either transplant immediately or never.

7.3 Credit Card Problem

In this section, we work on a simplified and modified version of the credit card example discussed by So and Thomas [54]. We use a discrete-time, absorbing Markov decision chain illustrated in Figure 7.3.

The states of the system are denoted by $(i, j),\ i = 1, 2, 3;\ j = \mathrm{l}, \mathrm{m}, \mathrm{h}$, where i represents the type of the customer and j is the credit limit given. We consider three customer types with "1" representing a customer who does not pay his/her debt in a timely manner, type "3" representing a responsible customer, and type "2" a moderate 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.

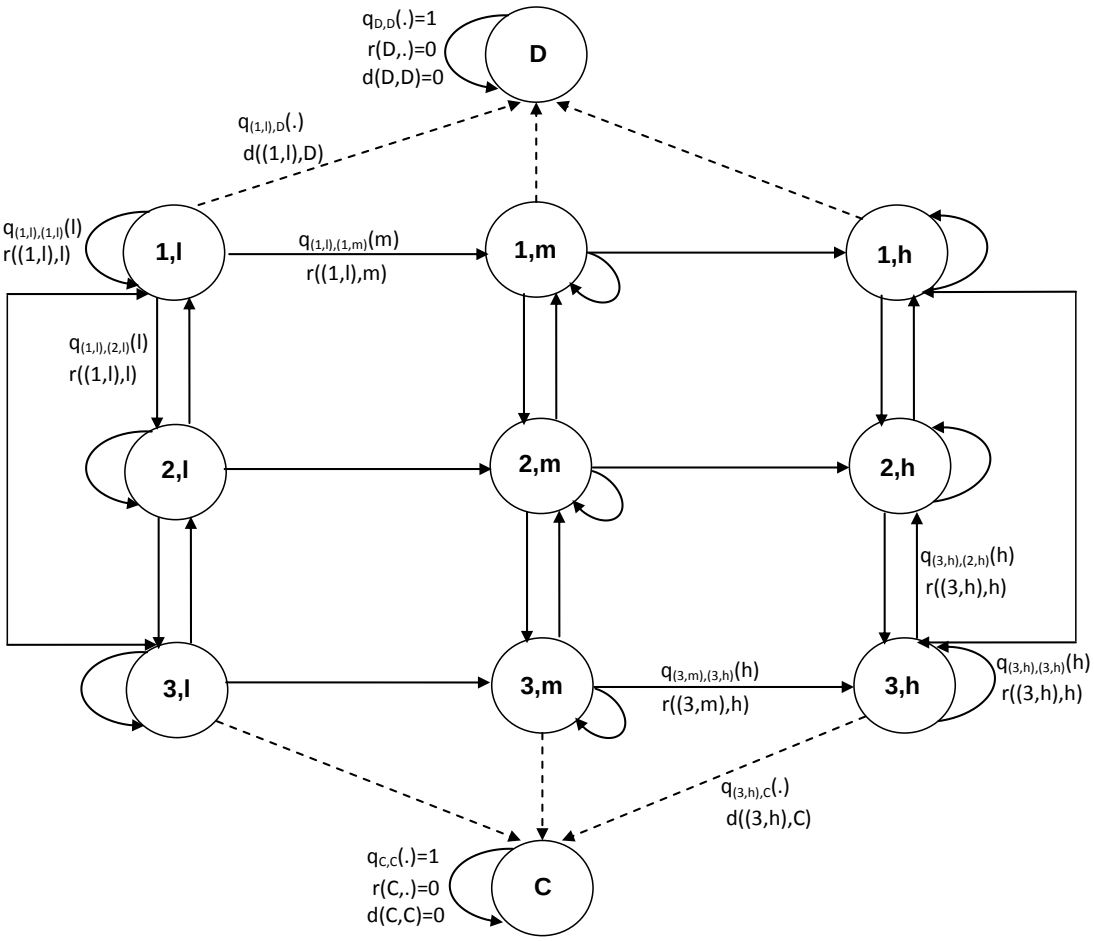

Figure 7.3: The credit card model.

Following So and Thomas [54], we do not consider decreasing the credit limit for 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 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, to keep the credit limit at high level (represented by h). There is only one formal control "Continue" for 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 customer type but same credit limit, or one of the absorbing states C and D. For example, under the control m, the set of all possible transitions from state $(2, \mathrm{m})$ is $\{(1, \mathrm{m}), (2, \mathrm{m}), (3, \mathrm{m}), \mathrm{C}, \mathrm{D}\}$. 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 new credit limit. In other words, if the credit limit is increased to h at state $(2, \mathrm{m})$, then the transition will be to state $(2, \mathrm{h})$ with probability one.

The rewards are the profits obtained at each time step. We consider two different profit values, the first one denoted by $r(x,u), x \in \mathcal{X}, u \in U(x)$ is the profit obtained at state x under the control u, and the second one, $d(x,y), x \in \mathcal{X}, y \in \mathcal{X}$ is the profit collected under 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. However, in order to stick to our notation, we will work with the negative of profit values represented by measures of risk, therefore, a minimization problem will be solved. We assume that feasible policies are limited to deterministic ones and we use the first-order mean–semideviation (see equation (2.4)) as the risk measure. Then, the dynamic programming equation (4.21) takes on the form

$$v(x) = \min_{u \in U(x)} \Bigg\{ \underbrace{\sum_{y \in \mathcal{X}} \Big(v(y) - r(x,u) - d(x,y) \Big) q_{x,y}(u)}_{\text{expected value } \psi} + \kappa \underbrace{\sum_{z \in \mathcal{X}} \Big(v(z) - r(x,u) - d(x,z) - \psi \Big)_+ q_{x,z}(u)}_{\text{semideviation}} \Bigg\}, \quad x \in \widetilde{X}, \quad (7.9)$$

which can also equivalently be written as follows using the fact that $\sum_{y \in \mathcal{X}} r(x,u) q_{x,y}(u) = r(x,u)$:

$$v(x) = \min_{u \in U(x)} \Bigg\{ -r(x,u) + \underbrace{\sum_{y \in \mathcal{X}} \Big(v(y) - d(x,y) \Big) q_{x,y}(u)}_{\overline{\psi}} + \kappa \sum_{z \in \mathcal{X}} \Big(v(z) - d(x,z) - \overline{\psi} \Big)_+ q_{x,z}(u) \Bigg\}, \quad x \in \widetilde{X}. \quad (7.10)$$

We use both value and policy iteration methods to solve the dynamic programming equation (7.10) with $v(\mathrm{C}) = 0$ and $v(\mathrm{D}) = 0$. As explained in section 5.1, value iteration is just the iteration of equation (7.10).

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 the Newton's method of section 5.2.1. In order to calculate μ_{l+1} at iteration $l+1$ of the Newton's method (see step 4 of Algorithm 3), we solve the following optimization problem for all $x \in \mathcal{X}$:

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

where $\pi^k(x) \in U(x), x \in \mathcal{X}$ 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

$$\begin{aligned}
& v(x) = \sum_{y\in\mathcal{X}} \Big(v(y) - r\big(x,\pi^k(x)\big) - d(x,y)\Big)\mu(x,y), \quad x \in \widetilde{X}, \\
& v(\mathrm{D}) = 0, \\
& v(\mathrm{C}) = 0.
\end{aligned}$$

7.3.1 Numerical Illustration

For numerical illustration, we used the transition probabilities given in Table 7.3. State and control dependent profit values $r(x,u), x \in \mathcal{X}, u \in U(x)$ are provided in Table 7.4 and the transition profits $d(x,y), x \in \mathcal{X}, u \in U(x)$ are given in Table 7.5. 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 [54].

We solved two different problems for this example. In the first problem, we assumed that the decision makers, namely creditors, are risk-neutral. Whereas, we considered risk-averse decision makers for the second problem. Since, in general, the operator $\mathfrak{D} : \mathcal{V} \to \mathcal{V}$ (see (5.1)) will be nonlinear, we did not allow randomized policies for the risk-averse case of this example and limited feasible policies to deterministic ones.

Limit	State	(1,l)	(1,m)	(1,h)	(2,l)	(2,m)	(2,h)	(3,l)	(3,m)	(3,h)	C	D
	(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
l	(2,m)	-	-	-	-	-	-	-	-	-	-	-
	(2,h)	-	-	-	-	-	-	-	-	-	-	-
	(3,l)	0.004	-	-	0.010	-	-	0.963	-	-	0.020	0.003
	(3,m)	-	-	-	-	-	-	-	-	-	-	-
	(3,h)	-	-	-	-	-	-	-	-	-	-	-
	(1,l)	-	1	-	-	-	-	-	-	-	-	-
	(1,m)	-	0.835	-	-	0.100	-	-	0.005	-	0.005	0.055
	(1,h)	-	-	-	-	-	-	-	-	-	-	-
	(2,l)	-	-	-	-	1	-	-	-	-	-	-
m	(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)	-	-	-	-	-	-	-	-	-	-	-
	(1,l)	-	-	-	-	-	-	-	-	-	-	-
	(1,m)	-	-	1	-	-	-	-	-	-	-	-
	(1,h)	-	-	0.829	-	-	0.060	-	-	0.001	0.010	0.100
	(2,l)	-	-	-	-	-	-	-	-	-	-	-
h	(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

Table 7.3: Transition probabilities.

The optimal policies and values of the expected value (risk-neutral) problem are given in Table 7.6. 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 different κ values. Optimal policies and values have been calculated using the iterative methods described in Chapter 5. The algorithms were coded in MATLAB R2011b and MOSEK optimization toolbox for MATLAB [35] was used to solve the optimization problem in Newton's method.

The convergence of the value iteration method is proved in Theorem 5.1.3 for problems with all nonpositive or nonnegative cost values. In this example, the profit values are not restricted to be all nonnegative or nonpositive, therefore, Theorem 5.1.3 does

Limit \ State	(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	-	2240	1920	-	650	560	-	90	80

Table 7.4: Profit values for state and control pairs.

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	-3700
(1,h)	-	-	-	-	-	-	-	-	-	1000	-15000
(2,l)	-	-	-	-	-	-	-	-	-	18	-400
(2,m)	-	-	-	-	-	-	-	-	-	30	-2500
(2,h)	-	-	-	-	-	-	-	-	-	500	-10000
(3,l)	-	-	-	-	-	-	-	-	-	5	-250
(3,m)	-	-	-	-	-	-	-	-	-	15	-1250
(3,h)	-	-	-	-	-	-	-	-	-	300	-4500

Table 7.5: Transition profits.

not apply here. However, using Lemma 5.1.2, we can state that if at any iteration k of the value iteration method, the value function v^k satisfies the relation $v^k \leq \mathfrak{D}v^k = v^{k+1}$, then (using an argument similar to the proof of Theorem 5.1.3) the remaining sequence obtained by the value iteration method will be nondecreasing and convergent to the optimal value v^*. Similarly, if $v^k \geq \mathfrak{D}v^k = v^{k+1}$, a nonincreasing remaining sequence converging to v^* is generated. For this example, the initial value function was set to zero, $v^0 = 0$, for the value iteration method. We observed that even when the sequence was not monotone at initial iterations of the value iteration algorithm, it became monotone very soon guaranteeing the convergence. The initial value function was also zero for the 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 7.7 and 7.8.

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 policy.

State	(1,l)	(1,m)	(1,h)	(2,l)	(2,m)	(2,h)	(3,l)	(3,m)	(3,h)
Values $v(\cdot)$	-7407.60	-7063.60	-4823.60	-7179.09	-7132.09	-6482.09	-6262.99	-6257.99	-5910.98
Policies	m	h	h	m	h	h	m	m	h

Table 7.6: Optimal values and policies for the expected value problem.

κ \ State	(1,l)	(1,m)	(1,h)	(2,l)	(2,m)	(2,h)	(3,l)	(3,m)	(3,h)
0.025	-7006.47	-6662.47	-4422.47	-6779.78	-6732.78	-6082.78	-5890.73	-5885.73	-5529.64
0.1	-6022.33	-5557.60	-3317.60	-5680.78	-5633.78	-4983.78	-4871.23	-4866.23	-4484.51
0.2	-4879.94	-4271.36	-2031.36	-4404.95	-4357.95	-3707.95	-3694.24	-3689.24	-3280.65
0.3	-3890.29	-3150.33	-910.33	-3298.83	-3251.83	-2601.83	-2684.25	-2679.25	-2246.70
0.4	-3025.84	-2166.80	73.20	-2331.68	-2284.68	-1634.68	-1814.65	-1809.65	-1351.35
0.5	-2263.92	-1296.49	943.51	-1477.88	-1430.88	-780.88	-1065.10	-1060.10	-568.84
0.6	-1583.41	-519.29	1720.71	-712.82	-665.82	-15.82	-419.64	-414.64	129.33
0.7	-973.84	178.30	2418.30	-25.64	21.36	671.36	137.76	142.76	753.34
0.8	-500.31	600.94	3047.74	493.20	641.34	1291.34	633.92	638.92	1311.99
0.9	-139.64	879.55	3618.58	878.60	1053.13	1853.64	1004.58	1009.58	1814.67
1	-2.70	989.73	4140.69	994.50	1145.21	2375.02	1095.70	1100.70	2299.66

Table 7.7: Optimal values, $v(\cdot)$, of the risk-averse problem for different κ's.

When we work with the negative of profit values, the parameter κ of the first-order mean–semideviation can be interpreted as a penalty parameter which penalizes the upper deviations from 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 7.8, it can be seen that for very small κ values, the optimal policy is same both for risk-averse and risk-neutral problems, which is a trivial result of previous assertion. Similarly, when κ gets smaller, optimal values get closer to the optimal values of expected value problem (see Table 7.7).

The number of iterations needed by both value and policy iteration methods for different κ values can be found in Table 7.9. For $\kappa = 1$, value iteration method required 1231 iterations, whereas, policy iteration method found the optimal solution in just three iterations. The first iteration of the policy iteration method required six Newton steps, second and third iterations required two and three Newton steps, respectively. It can be seen that policy iteration found the optimal solution at most four steps and

κ \ State	(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

Table 7.8: Optimal policies of the risk-averse problem for different κ's.

each step required at most six Newton iterations. However, the value iteration method required much more steps, changing between 525 and 1354.

κ	# of Value Iterations	# of Policy Iterations	# of Newton Iterations
0.025	869	3	4,3,3
0.1	797	4	3,3,2,3
0.2	746	4	3,3,2,2
0.3	689	4	4,2,2,2
0.4	658	4	4,2,2,2
0.5	661	4	4,2,2,2
0.6	761	3	4,3,3
0.7	893	3	4,2,3
0.8	525	3	4,3,2
0.9	1354	3	5,2,3
1	1231	3	6,2,3

Table 7.9: Number of iterations for the risk-averse problem.

Expected Total Profits for 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

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

for a given stationary policy $\Pi = \{\pi, \pi, \ldots\}$. The expected total profit function $\varphi(x)$, $x \in \mathcal{X}$ can be found by solving the following equation with $\varphi(\mathrm{C}) = 0$ and $\varphi(\mathrm{D}) = 0$ (*cf.*

Hernández-Lerma and Lasserre [19, Lemma 9.4.8]):

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

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 7.10. 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.

κ \ State	(1,l)	(1,m)	(1,h)	(2,l)	(2,m)	(2,h)	(3,l)	(3,m)	(3,h)
0.025	7407.60	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.1	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.2	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.3	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.4	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.5	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.6	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.7	7363.82	7063.60	4823.60	7179.09	7132.09	6482.09	6262.99	6257.99	5910.98
0.8	6250.72	5095.83	4823.60	5706.40	7132.09	6482.09	6125.71	6120.71	5910.98
0.9	2096.97	845.98	4823.60	648.85	408.31	6482.09	356.36	351.36	5910.98
1	2096.97	845.98	4823.60	648.85	408.31	6482.09	356.36	351.36	5910.98

Table 7.10: Expected total profits for the risk-averse problem for different κ's.

In order to estimate the distribution of the total profit, we simulated the Markov process under the optimal policies of the expected model and risk-averse model with $\kappa = 1$. We used Microsoft Excel based simulation tool YASAI (Version 2.3 [16]) (see [15]). The sample size was 32760 and the random number seed used was 10000. The graphs of resulting cumulative distribution functions are provided in Figures 7.4 - 7.12.

The first order mean–semideviation of Example 2.1.2 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, Dentcheva and Ruszczynski [53, sec. 6.3.3]). However, consistency with the second order stochastic dominance is guaranteed without any additional conditions (see Ogryczak and Ruszczyński [39, 40, 41], and Shapiro, Dentcheva and

Ruszczynski [53, sec. 6.3.3]).

Due to consistency with stochastic orders, the first order mean–semideviation should never prefer stochastically dominated outcomes, which can be observed from Figures 7.4 - 7.12. Total profit under the optimal policy of the risk-averse model with $\kappa = 1$ is 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, there is only one possible control for these states, which is to keep the credit limit unchanged, and possible transitions are to states with high credit limit, or to C and D.

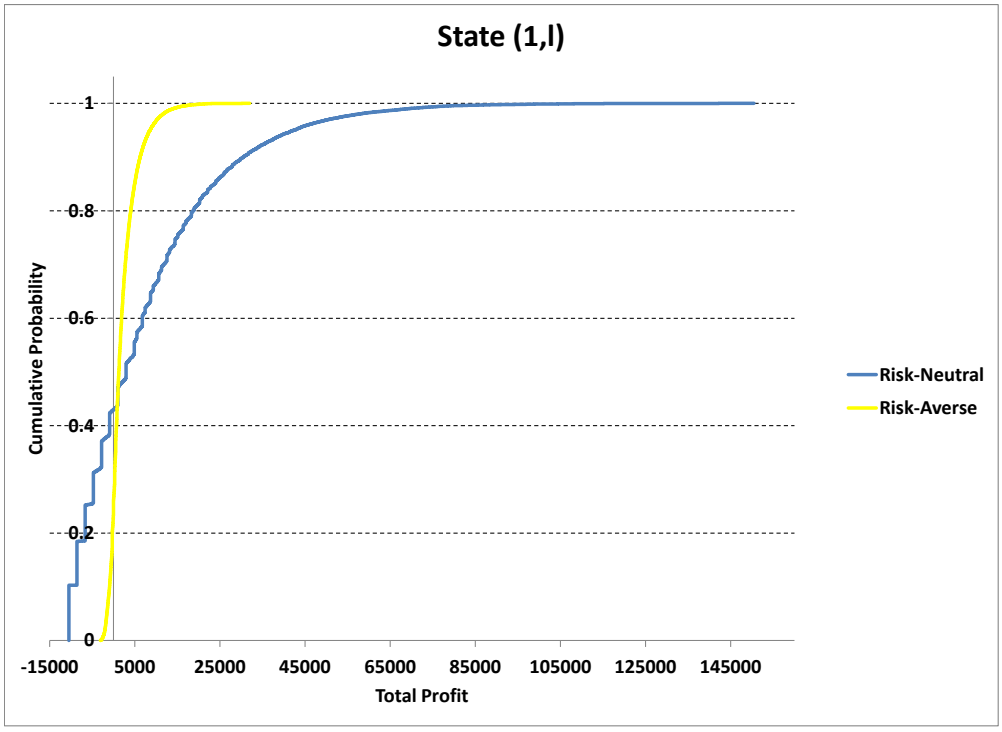

Figure 7.4: Cumulative probability distribution functions of total profit at state $(1, \mathrm{l})$.

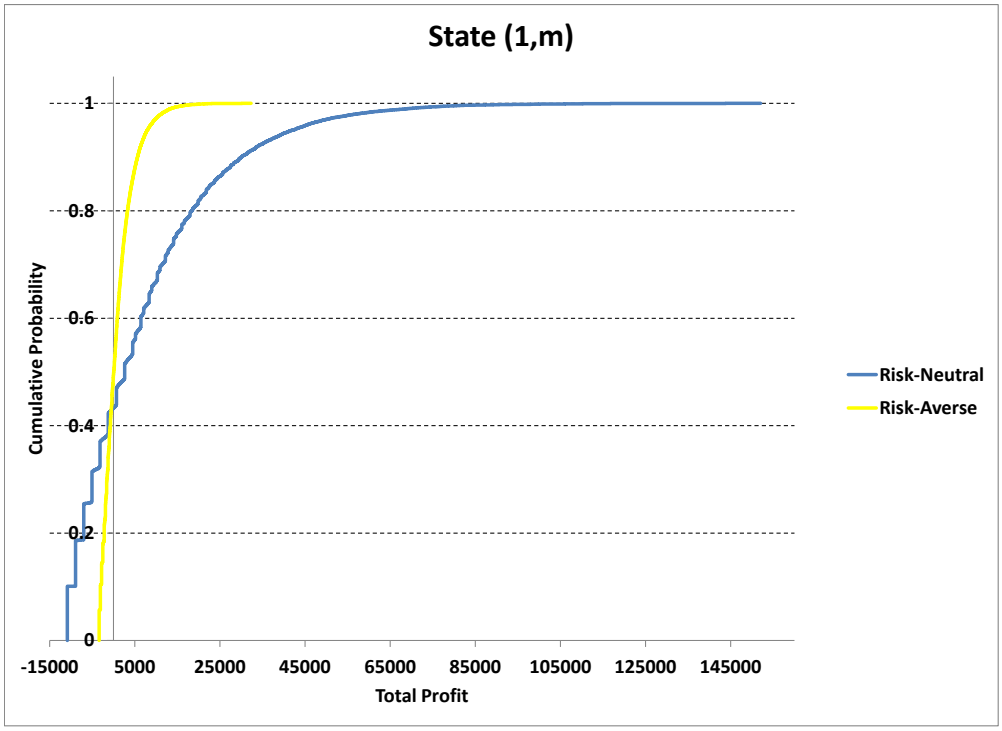

Figure 7.5: Cumulative probability distribution functions of total profit at state $(1, \mathrm{m})$.

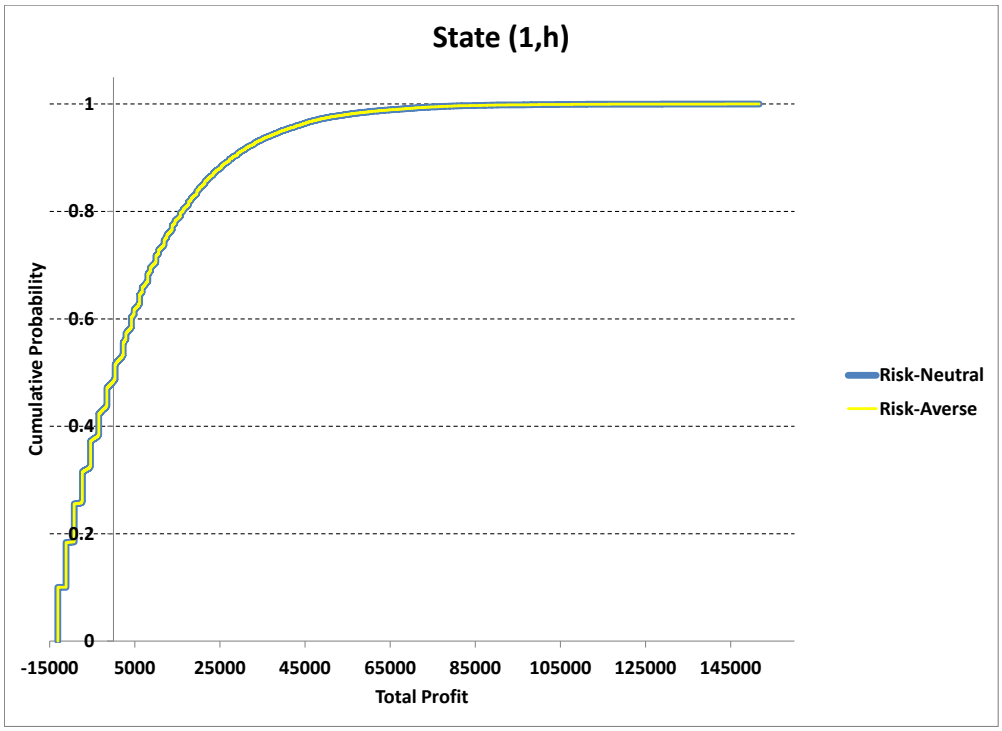

Figure 7.6: Cumulative probability distribution functions of total profit at state $(1, \mathrm{h})$.

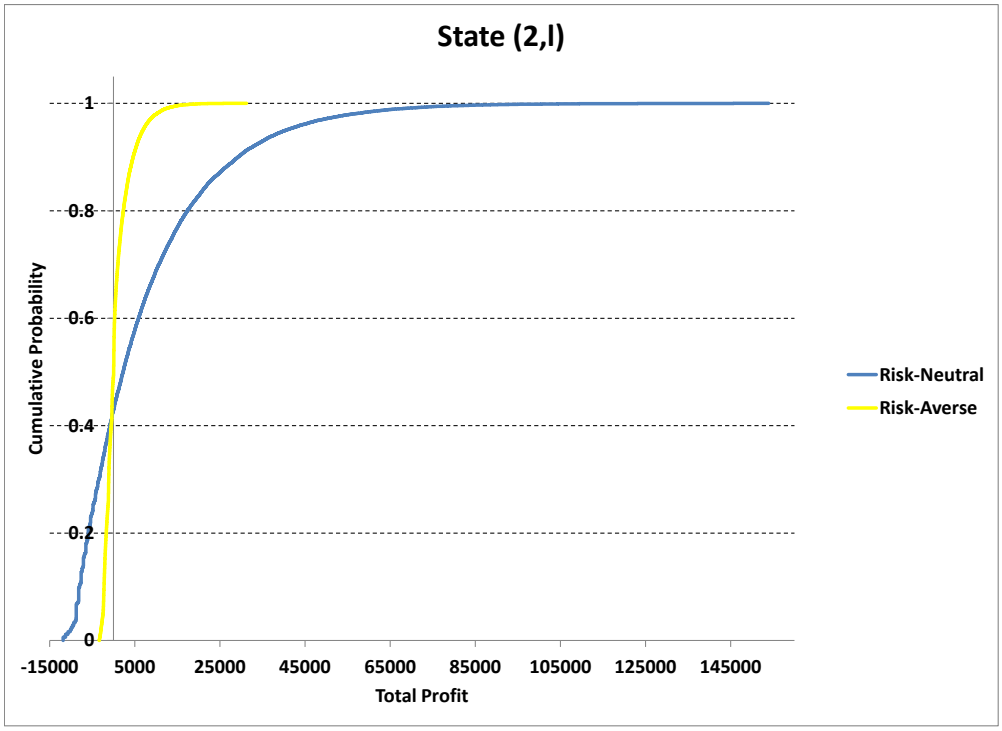

Figure 7.7: Cumulative probability distribution functions of total profit at state $(2, \mathrm{l})$.

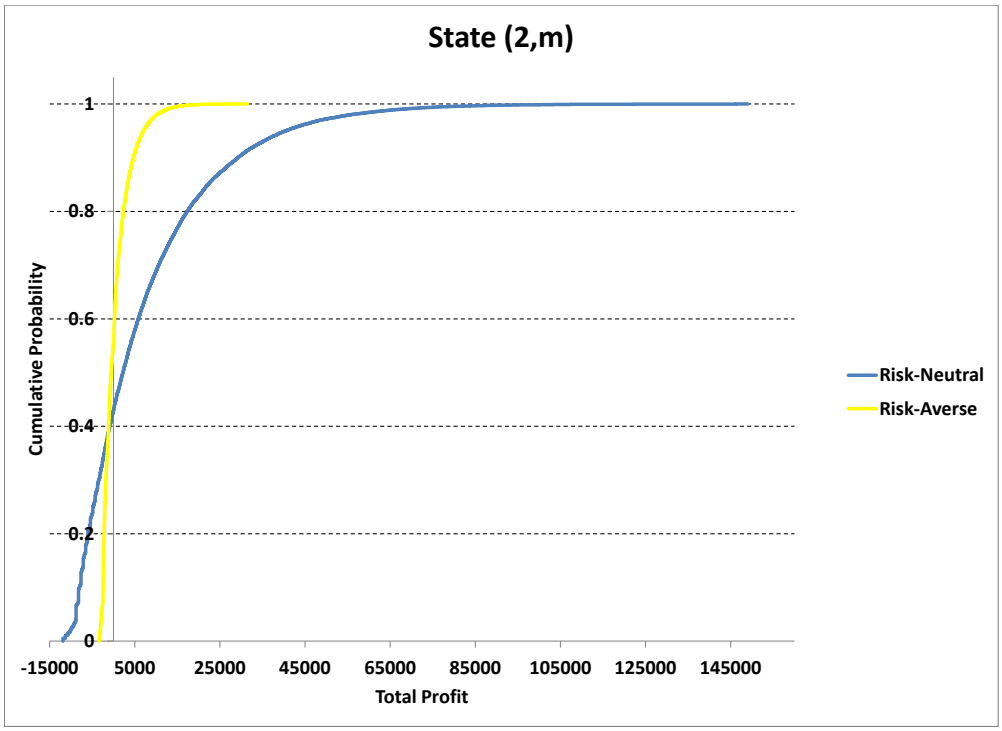

Figure 7.8: Cumulative probability distribution functions of total profit at state $(2, \mathrm{m})$.

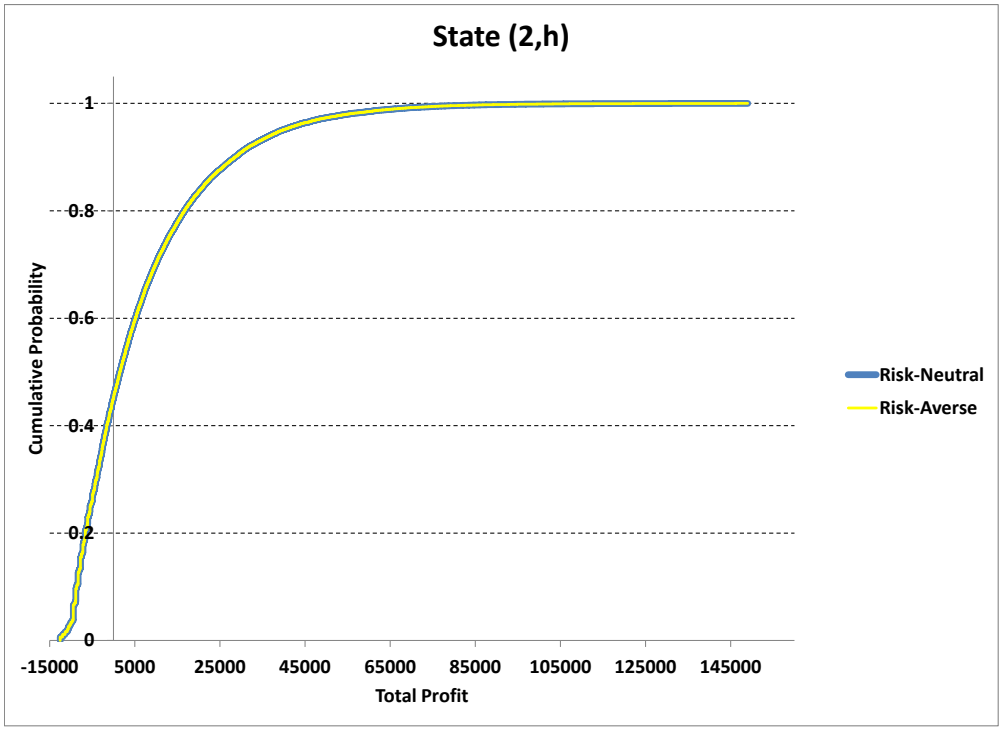

Figure 7.9: Cumulative probability distribution functions of total profit at state $(2, \mathrm{h})$.

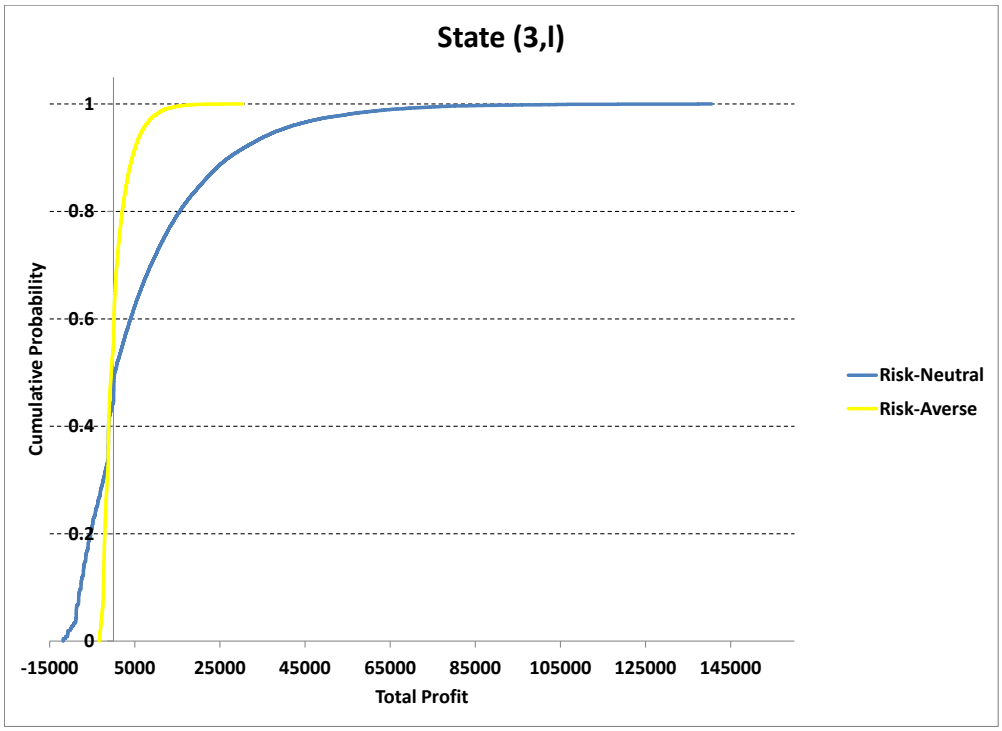

Figure 7.10: Cumulative probability distribution functions of total profit at state $(3, \mathrm{l})$.

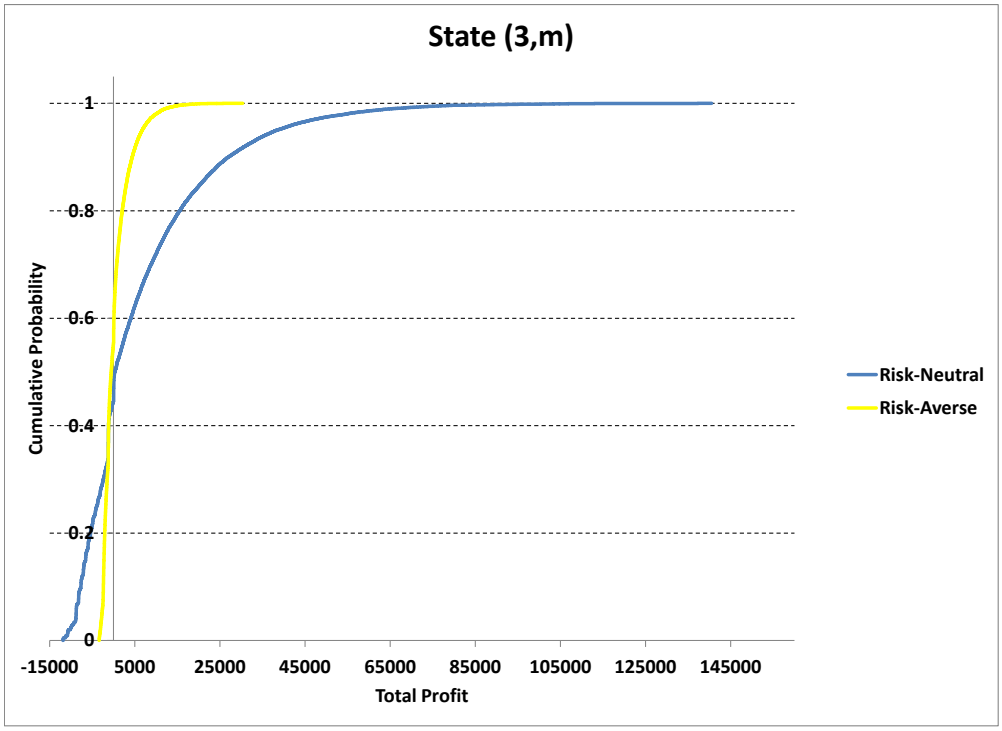

Figure 7.11: Cumulative probability distribution functions of total profit at state $(3, \mathrm{m})$.

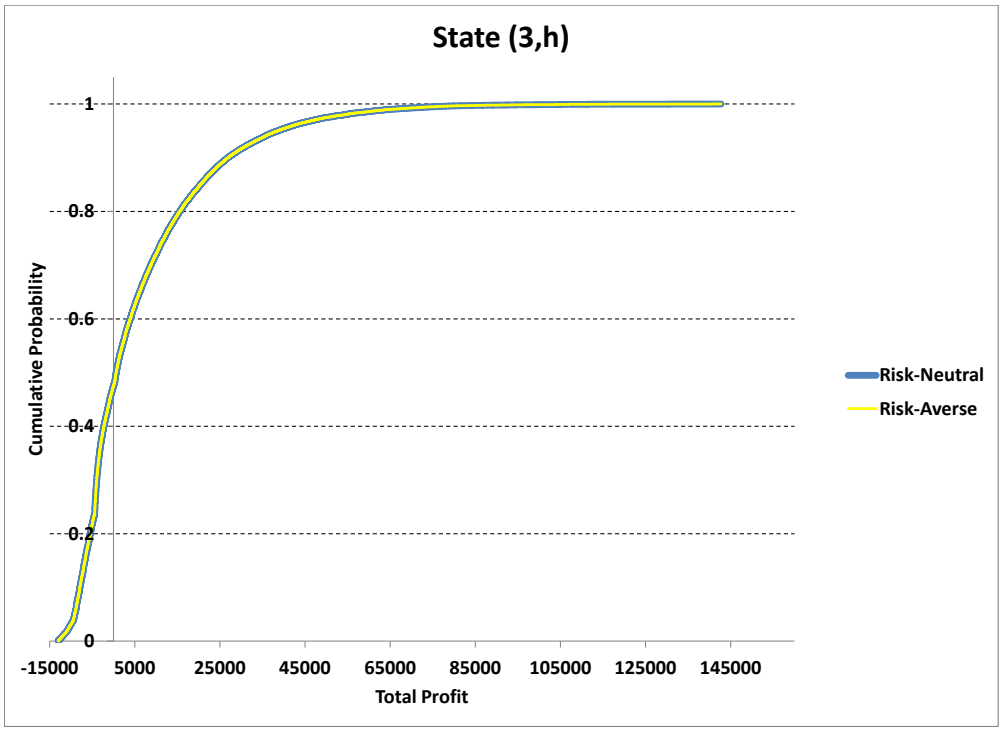

Figure 7.12: Cumulative probability distribution functions of total profit at state $(3, \mathrm{h})$.

Chapter 8

Conclusion and Future Study

We adopted and extended the concept of Markov risk measures suggested by Ruszczyński [48] to randomized policies and used it to develop a new risk-averse formulation for the discrete-time, infinite horizon, undiscounted, transient Markov process. We showed that, when risk measures are employed, there may not exist a finite optimal value function of the problem even the process is transient. Therefore, we defined a new concept called risk-transient Markov model, which generalizes the Pliska condition (1.1).

We derived the risk-averse dynamic programming equations for risk-transient and stationary Markov models with general state and compact control spaces. We showed that the dynamic programming equations have a unique solution which is equivalent to the optimal value function of the risk-averse Markov control problem and the optimal policy can be found using these equations.

For expected value models, the optimal policy will always be deterministic, therefore, it is enough to limit the consideration to deterministic policies. We showed that this fact may not be valid when risk measures are employed and a randomized policy may be optimal, in general. However, we proved that for Conditional Average Value at Risk, similar to the expected value models, it is sufficient to consider just deterministic policies.

We suggested risk-averse value and policy iteration methods to solve the dynamic programming equations and proved the convergence of the methods to the unique solution of these equations. For the expected value models, the classical policy iteration method requires solving a system of linear equations, which turns into a system of nonsmooth equations, in general, for the risk-averse models. We adopted the specialized nonsmooth Newton method of Ruszczyński [48] to solve this nonsmooth equation

system and proved its global convergence for risk-transient Markov control models.

Assuming that the state and control spaces are finite, we proposed another method, that we call mathematical programming approach, for solving the risk-averse control problems for transient models. However, the formulation that we suggested is nonconvex, therefore, this approach is not promising compared to the iterative methods.

In order to explain the results of the study, we focused on three different examples (asset selling, organ transplant, and credit card problems) which were adopted from some problems existing in the literature. We derived the structure of the optimal policy for the asset selling problem and showed that an optimal control-limit policy exists. The results of the organ transplant example support our theory that randomized policies may be optimal when risk measures are used. We compared the iterative methods on the credit card example, where policy iteration method required much less iterations than the value iteration.

As a future study, the rate of convergence of the proposed iterative methods can be studied. When randomized policies are considered, policy and value iteration methods, in general, require solving a nonlinear optimization problem for improving the current policy and value, respectively. The corresponding problem can be solved by enumeration if just deterministic policies are considered. However, this convenience is not possible for randomized policies, therefore, a commercial solver or a solution algorithm is required. In the future, efficient solution methods may be suggested for this nonlinear optimization problem.

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(10), 1420–1430, 2004.

[2] Artzner, P., F. Delbaen, J.-M. Eber, and D. Heath, Coherent measures of risk, *Mathematical Finance*, 9(3), 203–228, 1999.

[3] Artzner, P., F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, Coherent multiperiod risk adjusted values and Bellman's principle, *Annals of Operations Research*, 152(1), 5–22, 2007.

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

[5] Bellman, R., *Dynamic Programming*, Princeton University Press, Princeton, NJ, 1957.

[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] Çavuş, Ö. and A. Ruszczyński, Risk-Averse Control of Undiscounted Transient Markov Models, *submitted*, 2012.

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

[9] 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.

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

[11] Denardo, E. V., On linear programming in a Markov decision problem, *Management Science*, 16(5), 281–288, 1970.

[12] Denardo, E. V. and U. G. Rothblum, Optimal stopping, exponential utility, and linear programming, *Mathematical Programming*, 16(1), 228–244, 1979.

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

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

[15] Eckstein, J. and S. T. Riedmueller, YASAI: Yet Another Add-in for Teaching Elementary Monte Carlo Simulation in Excel, *INFORMS Trans. Ed.*, 2(2), 12–26, 2002.

[16] YASAI (Version 2.3) [Software], 2011. Available from http://www.yasai.rutgers.edu/.

[17] Feinberg, E. A. and A. Shwartz (Eds.), *Handbook of Markov Decision Processes: Methods and Applications*, Kluwer, Boston, 2002.

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

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

[20] Howard, R. A., *Dynamic Programming and Markov Processes*, Wiley, New York, 1960.

[21] Howard, R. A. and J. E. Matheson, Risk-sensitive Markov decision processes, *Management Science*, 18(7), 356–369, 1972.

[22] Jacuette, C. J., A utility criterion for Markov decision processes, *Management Science*, 23(1), 43–49, 1976.

[23] Jasiulewicz, H., Application of mixture models to approximation of age-at-death distribution, *Insurance: Mathematics and Economics*, 19(3), 237–241, 1997.

[24] Kallenberg, L. C. M., *Linear Programming and Finite Markovian Control Problems*, Mathematical Centre Tracts 148, Amsterdam, 1983.

[25] Kallenberg, L. C. M., Survey of linear programming for standard and nonstandard Markovian control problems. Part I: Theory, *Mathematical Methods of Operations Research*, 40, 1–42, 1994.

[26] 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.

[27] Klatte, D. and B. Kummer, *Nonsmooth Equations in Optimization*, Kluwer, Dordrecht, 2002.

[28] Klöppel, S. and M. Schweizer, Dynamic indifference valuation via convex risk measures, *Mathematical Finance*, 17(4), 599–627, 2007.

[29] Kummer, B., Newton's method for non-differentiable functions, in: *Advances in Mathematical Optimization* , J. Guddat et al. (Eds.), Academie-Verlag, Berlin, 1988, pp. 114–125.

[30] Kurt, M. and J. P. Kharoufeh, Monotone optimal replacement policies for a Markovian deteriorating system in a controllable environment, *Operations Research Letters*, 38(4), 273–279, 2010.

[31] Levitt, S. and A. Ben-Israel, On modeling risk in Markov decision processes, in: *Optimization and Related Topics*, Rubinov, A., and B. Glover (Eds.), Kluwer Academic Publishers, Dordrecht, 2001, pp. 27–40.

[32] Manne, A. S., Linear programming and sequential decisions, *Management Science*, 6(3), 259–267, 1960.

[33] Mannor, S. and J. N. Tsitsiklis, Mean-variance optimization in Markov decision processes, in: *Proceedings of the* 28^{th} *International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

[34] Mannor, S. and J. Tsitsiklis, Mean-variance optimization in Markov decision processes, *CoRR*, abs/1104.5601, 2011. URL http://arxiv.org/abs/1104.5601.

[35] MOSEK Optimization Toolbox for MATLAB. Available from http://mosek.com/.

[36] Nilim, A. and L. El Ghaoui, Robust control of Markov decision processes with uncertain transition matrices, *Operations Research*, 53(5), 780-798, 2005.

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

[38] Ohtsubo, Y., Optimal threshold probability in undiscounted Markov decision processes with a target set, *Applied Mathematics and Computation*, 149(2), 519–532, 2004.

[39] Ogryczak, W. and A. Ruszczyński, From stochastic dominance to mean-risk models: Semideviations as risk measures, *European Journal of Operational Research*, 116(1), 33–50, 1999.

[40] Ogryczak, W. and A. Ruszczyński, On consistency of stochastic dominance and mean-semideviation models, *Mathematical Programming*, 89(2), 217–232, 2001.

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

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

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

[44] Pliska, S. R., On the transient case for Markov decision chains with general state spaces, in: *Dynamic Programming and Its Applications*, M. L. Puterman (Eds.), Academic Press, New York, 1978, pp. 335–349.

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

[46] R. T. Rockafellar and S. Uryasev, Conditional value-at-risk for general loss distributions, *Journal of Banking and Finance*, 26(7), 1443–1471, 2002.

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

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

[49] 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-Verlag, London, 2006, pp. 119–157.

[50] Ruszczyński, A. and A. Shapiro, Optimization of convex risk functions, *Mathematics of Operations Research*, 31(3), 433–452, 2006.

[51] Ruszczyński, A. and A. Shapiro, Conditional risk mappings, *Mathematics of Operations Research*, 31(3), 544–561, 2006.

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

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

[54] So, M. M. C. and L. C. Thomas, Modelling the profitability of credit cards by Markov decision processes, *European Journal of Operational Research*, 212(1), 123–130, 2011.

[55] 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(5), 457–466, 1979.

[56] Veinott, A. F., Discrete dynamic programming with sensitive discount optimality criteria, *The Annals of Mathematical Statistics*, 40(5), 1635–1660, 1969.

[57] Yu, S. X., Y. Lin, and P. Yan, Optimization models for the first arrival target distribution function in discrete time, *Journal of Mathematical Analysis and Applications*, 225(1), 193–223, 1998.

[58] White, D. J., A survey of applications of Markov decision processes, *The Journal of the Operational Research Society*, 44(11), 1073–1096, 1993.

Vita

Özlem Çavuş

2007-2012 *Ph.D. in Operations Research*
RUTCOR, Rutgers University, NJ, USA

2004-2007 *M.S. in Industrial Engineering*
Bogazici University, Istanbul, Turkey

1999-2004 *B.S. in Industrial Engineering*
Bogazici University, Istanbul, Turkey

2011-2012 Graduate Assistant, RUTCOR, Rutgers University

2007-2011 Teaching Assistant, RUTCOR, Rutgers University

2004-2007 Teaching Assistant, Department of Industrial Engineering, Bogazici University