METHODS OF TEMPORAL DIFFERENCES FOR RISK-AVERSE DYNAMIC PROGRAMMING AND LEARNING

by

ÜMİT EMRE KÖSE

A dissertation submitted to the

Graduate School-Newark

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Management

Written under the direction of

Andrzej Ruszczyński

and approved by

_____

_____

_____

_____

Newark, New Jersey

May, 2020

ABSTRACT OF THE DISSERTATION

Methods of Temporal Differences for Risk-Averse Dynamic Programming and Learning

By Ümit Emre Köse

Dissertation Director:

Andrzej Ruszczyński

Stochastic sequential decision-making problems are generally modeled and solved as Markov decision processes. When the decision-makers are risk-averse, their risk-aversion can be incorporated into the model using dynamic risk-measures. Such risk-averse Markov decision processes can be theoretically solved by specialized dynamic programming methods. However, when the state space of the system becomes very large, then such methods become impractical.

We consider reinforcement learning with performance evaluated by a dynamic risk measure for Markov decision processes. We use a linear value function approximation scheme and construct a projected risk-averse dynamic programming equation that involves this scheme. We study the properties of this equation. To solve this equation, we propose risk-averse counterparts of the methods of temporal differences and we prove their convergence with probability one. We also perform an empirical study on a complex transportation problem where we demonstrate that the risk-averse methods of temporal differences outperform the well known risk-neutral methods in terms of average profit over time.

Acknowledgements

Firstly, I would like to express my utmost gratitude to my advisor, Dr. Andrzej Ruszczyński, for his help, guidance, wisdom, and patience throughout my doctoral study. My interest in my research area only increased under his supervision, and it has been an honor and privilege to be his student. I also owe my earnest gratitude to Dr. Darinka Dentcheva, Dr. Jonathan Eckstein, and Dr. Mert Gürbüzbalaban for taking the time to be in my committee to provide their valuable comments and remarks. I am sincerely thankful to Dr. Özlem Çavuş for introducing me to this research area in my study for a master's degree and suggesting me to further my education at Rutgers.

Of course, I would like to thank my parents, Harun Köse and Kadem Köse, and my brother, Berkay Köse for their understanding and continuous emotional support during my studies. I am also thankful to Astrid Noor for being my best friend throughout my study, and for her invaluable presence and company.

This research was funded by National Science Foundation under the award number 1907522. I must also acknowledge the Office of Advanced Research Computing at Rutgers, The State University of New Jersey for providing access to the Amarel cluster and associated research computing resources that have contributed to the results shown in this study.

Dedication

To my parents.

Table of Contents

Chapter 1

Preliminaries

1.1 Introduction

Markov decision processes (MDPs) are well known instruments in modeling and solving sequential decision making problems with stochasticity [5, 34, 57]. In the classical sense, the goal of an MDP is usually to minimize the expected cost over a finite planning horizon or expected discounted cost over an infinite horizon; there are models minimizing the expected undiscounted cost over an infinite horizon in problems having a Markov chain with transient nature [57]. MDPs have various applications such as in stochastic shortest path problems [7] and optimal stopping problems [18, 25, 26, 57].

Traditional MDP models, although effective for small to medium size problems, suffer from the curse of dimensionality in problems with large state space. Approximate dynamic programming approaches try to tackle the curse of dimensionality and provide an approximate solution of an MDP (see [54] for an overview). Such methods usually involve value function approximations, where the value of a state of the Markov process is approximated by a simple, usually linear, function of some selected features of the state [6].

Reinforcement learning methods [69, 54] involve simulation or observation of a Markov process to approximate the value function and learn the corresponding policies. The first studies attempted to emulate neural networks and biological learning processes, learning by trial and error [48, 29]. Some learning algorithms, such as Q-Learning [75, 76] and SARSA [61], follow this idea. One of core approaches in reinforcement learning is the method of temporal differences [68], known as TD(λ). It uses differences between the values of the approximate value function at successive states to improve the approximation, concurrently with the evolution of the system. TD(λ) is a continuum of algorithms depending on a parameter $\lambda \in [0,1]$ which is used to exponentially weight past observations. Consequently, related methods such as

Q(λ) [75, 51, 52, 60] and SARSA(λ) were developed [61, 60]. The methods of temporal differences have been proven to converge in the mean in [20] and almost surely by several studies, with different degrees of generality and precision [51, 21, 73, 36, 74].

We introduce risk models into temporal difference learning. In the extant literature, three basic approaches to risk aversion in MDPs have been employed: utility functions (see, e.g., [37, 38, 17, 22, 31, 4, 39]), mean–variance models (see. e.g., [78, 30, 46, 1, 13]), and entropic (exponential) models (see, e.g., [35, 47, 8, 19, 24, 43, 4]). Our research is rooted in the theory of dynamic measures of risk, which has been intensively developed in the last 15 years (see [66, 58, 59, 32, 14, 63, 3, 53, 41, 40, 15] and the references therein).

In [62], Markov dynamic risk measures were introduced, specially tailored for the MDPs. It allowed for the development of dynamic programming equations and corresponding solution methods, generalizing the well-known results for the expected value problems. These ideas were successfully extended to undiscounted problems in [12, 11], partially observable and history-dependent systems in [28, 27], and further generalized in [44, 67].

A number of works introduce models of risk into reinforcement learning: exponential utility functions [10, 9] and mean-variance models [71, 56]. Few later studies propose heuristic approaches involving coherent risk measures and their mean-risk counterparts [16, 70]; these studies employ policy gradients and use them in actor-critic type algorithms. Distributed policy gradient methods with risk measures were proposed in [45]. Model-related uncertainties are discussed in [72].

In this study, we use Markov risk measures of [62] in conjunction with linear approximations of the value function. We provide a projected risk-averse dynamic programming equation and the analysis of its properties. We develop the risk-averse methods of temporal differences in single-step and multi-step case and their convergence analyses. We demonstrate the performance of these methods on an empirical study involving a transportation problem.

1.2 Outline of the Dissertation

The first chapter is dedicated to provide a background and lay the preliminaries of the Markov decision processes, risk-neutral and risk-averse, and to briefly mention the risk-neutral methods of temporal differences. In Chapter 2, we develop the risk-averse projected dynamic programming equation with a linear value function approximation, in which we rely on a transition risk mapping instead of conditional expectation that is used in the risk-neutral case. We utilize the dual representation of the coherent measures of risk to develop the idea of distortion coefficient and use it to show the existence of a unique fixed point to the projected dynamic programming equation. In Chapter 3, we generalize the well-known TD(λ) algorithm [68] to the risk-averse case in order to approximately solve the projected risk-averse dynamic programming equation. Chapter 4 includes a discourse which shows that the risk-averse methods of temporal differences, for all selections of λ, converge to the fixed point of the projected risk-averse dynamic programming equation under mild conditions on the degree of risk-aversion inherent to the underlying coherent measure of risk. We compare risk-averse and risk-neutral methods of temporal differences in Chapter 5 on an empirical study involving a practical transportation problem and illustrate the ways in which the risk-averse methods outperform the risk-neutral ones. Chapter 6 includes our conclusion and further discussion.

1.3 Markov Decision Processes

Markov Decision Processes (MDPs) are well known tools to model and solve sequential decision processes under uncertainty [5, 34, 57]. In most of such problems a discrete time setting $t = 0, 1, 2 \ldots$ is assumed. Such a stochastic system, at a given time t, is at a state i_t within the state space $\mathscr{X}$ with n states, and depending on the action u_t taken within the action set $\mathscr{U}(i_t) \subset \mathscr{U}$ associated with state i_t, it moves to state $i_{t+1} \in \mathscr{X}$ with probability $P_{i_t, i_{t+1}}(u_t)$. The action space $\mathscr{U}$ can be interpreted as the space of all possible actions in all possible states. During a transition, a cost $c(i_t, u_t)$ depending on the state i_t and action u_t is incurred, where $c : \mathscr{X} \times \mathscr{U} \to \mathbb{R}$; the cost may also depend on the next state i_{t+1}. Markov decision processes can have a finite or infinite time horizon, and usually the objective is to minimize the total cost in the finite horizon case, and the discounted total cost in the infinite horizon case. The finite horizon case has the

objective

$$J_T = \mathbb{E}\Big[\sum_{t=0}^{T} c(i_t, u_t)\Big] \tag{1.1}$$

whereas in the infinite horizon case the objective is to minimize

$$J_\infty = \mathbb{E}\Big[\sum_{t=0}^{\infty} \alpha^{t-1} c(i_t, u_t)\Big], \quad \alpha \in (0,1). \tag{1.2}$$

The key consideration in a Markov decision process is to determine what is called a *policy*, that is, a rule determining actions based on the observations gained so far. In the models mentioned earlier, a Markov policy is optimal. A *Markov policy* is a collection of rules $\pi_t : \mathscr{X} \to \mathscr{U}$ that prescribe an action to take at time t for every state in $\mathscr{X}$. Thus, the goal is often to find a Markov policy that minimizes either (1.1) or (1.2) depending on whether the problem has a finite or infinite time horizon. Throughout the rest of the dissertation, we will mostly focus on the infinite horizon case, where the optimal policy is known to be stationary.

Once a Markov policy π is fixed, then the process becomes an ordinary Markov chain, since the decision is determined by the state. This Markov chain has transition probabilities $P_{ij}^{\pi} = P_{ij}(\pi(i))$, $i, j \in \mathscr{X}$, and cost function $c^{\pi}(i) = c(i, \pi(i))$, $i \in \mathscr{X}$. Since we are now dealing with an infinite horizon setting, it is possible to define a value function $v^{\pi} : \mathscr{X} \to \mathscr{V}$ associated with policy π where $\mathscr{V}$ is the space of all real functions that are defined over $\mathscr{X}$. This function stores the total expected discounted cost that will be observed starting from a state under policy π. Namely, it is defined as

$$v^{\pi}(i) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \alpha^t c(i_t, \pi(i_t)) \,\Big|\, i_0 = i\Big].$$

The value function defined in such a way satisfies what is called the *policy evaluation equation*

$$v^{\pi}(i) = c^{\pi}(i) + \alpha \sum_{j \in \mathscr{X}} P_{ij}^{\pi} v^{\pi}(j), \quad i \in \mathscr{X}.$$

This equation is linear, and if we treat v^{π} and c^{π} as vectors, we can write this equation compactly as

$$v^{\pi} = c^{\pi} + \alpha P^{\pi} v^{\pi}. \tag{1.3}$$

For a given policy π, the value function satisfying the policy evaluation equation, if the state space is small enough, can be found using traditional dynamic programming methods such as

value iteration [5]. Then the value function that is found can be used to improve upon the current policy, and then a new value function can be found which satisfies the policy evaluation equation under the new policy. This idea is called policy iteration [34]. In a policy iteration scheme, the premise is to be able to enumerate all of the states in the state space. For a given value function v^{π}, the new policy $\overline{\pi}$ is constructed with actions that satisfy the following

$$\overline{\pi}(i) \in \operatorname*{argmin}_{u \in \mathscr{U}(i)} \Big\{ c(i,u) + \alpha \sum_{j \in \mathscr{X}} P_{ij}(u)\, v^{\pi}(j) \Big\}, \quad i \in \mathscr{X}.$$

The traditional value iteration and policy iteration methods assume that we are able to enumerate the state space $\mathscr{X}$, and they become impractical when the state space is very large.

1.4 Risk-Neutral Approximate Dynamic Programming

The impracticality mentioned at the end of the Section 1.3 is commonly circumvented using what is called a *value function approximation*. In a setting where the state space is too large, one may not be able to express a value function or use it in the iterative methods previously mentioned. In this section we recall the well-known risk-neutral approximate dynamic programming approaches such that their contrast with the novel risk-averse methods are more pronounced.

1.4.1 Projected Dynamic Programming Equation

Instead of mapping every state $i \in \mathscr{X}$ to a real number using a value function, it is possible to assume that each state $i \in \mathscr{X}$ has a number of relevant *features* $\varphi_j(i) \in \mathbb{R}$, $j = 1, \ldots, m$, where $m \ll n$, and that the value $v^{\pi}(i)$ of a state can be approximated by a linear combination of its features:

$$v^{\pi}(i) \approx \widetilde{v}^{\pi}(i) = \sum_{j=1}^{m} r_j \varphi_j(i), \quad i \in \mathscr{X}. \tag{1.4}$$

From now on, we suppress the superscript π, because most of the considerations focus on evaluating a fixed policy. We define the matrix of the features of all states, namely

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}^{\top}(1) \\ \boldsymbol{\varphi}^{\top}(2) \\ \vdots \\ \boldsymbol{\varphi}^{\top}(n) \end{bmatrix}.$$

Now we can write our approximation as $v \approx \tilde{v} = \Phi r$. The goal is to emulate (1.3) with the approximate value function. However, we may observe that the right hand side of the equation, $c + \alpha P \Phi r$, may not be represented as a linear combination of the features. Therefore, we need to project this vector on the subspace spanned by the features, range(Φ). Accordingly, we define a projection operator, $L : \mathscr{V} \rightarrow \text{range}(\Phi)$, and formulate the projected dynamic programming equation:

$$\Phi r = L\big(c + \alpha \alpha P \Phi r\big). \tag{1.5}$$

It is assumed that the Markov system under policy π is ergodic, and we denote its vector of stationary probabilities by q. We define the projection operator using the following scalar product and the associated norm: $\langle v, w \rangle_q = \sum_{i=1}^{n} q_i v_i w_i$, $\|w\|_q^2 = \langle w, w \rangle_q$. Then

$$L(w) = \underset{z \in \text{range}(\Phi)}{\text{argmin}} \; ||z - w||_q, \quad w \in \mathscr{V}. \tag{1.6}$$

It is also noteworthy to mention that the projection operator $L(\cdot)$ is nonexpansive with respect to the norm $||\cdot||_q$.

1.4.2 Risk-Neutral Methods of Temporal Differences

The projected dynamic programming equation (1.5) implies that if r^* satisfies the aforementioned equation, it also satisfies the following, by the definition of the projection operator (1.6):

$$r^* = \underset{r}{\text{argmin}} \, ||\Phi r - (c + \alpha P \Phi r^*)||_q^2.$$

The optimality condition for the r^* that serves as the minimizing argument is that

$$\begin{aligned} \Phi^T Q[\Phi r - (c + \alpha P \Phi r^*] &= 0, \\ \Phi^T Q \Phi r - \Phi^T Q c - \alpha \Phi^T Q P \Phi r^* &= 0, \\ \Phi^T Q \Phi (I - \alpha P) \Phi r^* - \Phi^T Q c &= 0. \end{aligned} \tag{1.7}$$

Let $A = \Phi^T Q (I - \alpha P) \Phi$ and $b = \Phi^T Q c$. Then the equation (1.7) can be expressed simply as $A r^* = b$. It is possible to approximate A and b by simulation by running a trajectory of states

$\{i_t\}_{t=1}^{N+1}$. We can observe that by virtue of the ergodicity of the chain,

$$\Phi^T Q \Phi \approx \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \phi^T(i_t),$$
$$\Phi^T Q P \Phi \approx \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \mathbb{E}[\phi^T(i_{t+1}) | i_t].$$

In the risk-neutral method, it is possible to substitute the expected features of the next state with the observed features of the next state in the observed trajectory, because averaging over many transitions will recover the expected value. Then it follows that

$$\Phi^T Q P \Phi \approx \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \phi^T(i_{t+1}).$$

We also have

$$\Phi^T Q c \approx \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) c(i_t).$$

Then we obtain the following estimates:

$$A \approx A_N = \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \left[\phi^T(i_t) - \alpha \phi^T(i_{t+1}) \right]$$
$$b \approx b_N = \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) c(i_t).$$

For a given r, the residuals of the equation have the following form:

$$\begin{aligned} A_N r - b_N &= \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \phi(i_t)^T r - \alpha \phi(i_t) \phi^T(i_{t+1}) - \phi(i_t) c(i_t) \\ &= \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) \underbrace{\left[\phi^T(i_t) r - \alpha \phi^T(i_{t+1}) - c(i_t) \right]}_{d_t} \\ &= \frac{1}{N} \sum_{t=1}^{N} \phi(i_t) d_t \end{aligned}$$

where d_t is the *temporal difference*. The residual error can be corrected over a large trajectory incrementally using the temporal differences. This is the idea that the risk-neutral methods of temporal differences relies on. The temporal difference updates have the form

$$r_{t+1} = r_t - \gamma_t \phi(i_t) d_t,$$

where γ_t is a sufficiently small stepsize parameter with $\gamma_t \to 0$. This method is called the risk neutral TD(0) algorithm, and in its multi-step version, instead of the features of the current

state, an exponential weighting of all the past states are used to determine the update direction. This exponential weighting is done with an algorithmic discount parameter $\lambda \in [0,1]$, and the updates have the form

$$r_{t+1} = r_t - \gamma_t z_t d_t,$$
$$z_t = \alpha\lambda z_{t-1} + \phi(i_t);$$

usually z_0 is assumed to be zero for simplicity. The multi-step version is called the risk-neutral TD(λ) algorithm.

1.5 Risk-Averse Markov Decision Processes

Most real life problems involve risk-averse decision makers. Under risk-aversion, the considerations in Section 1.3 are no longer valid, because every random variable was taken into consideration with its excepted value alone. In [62], in a more general setting in a Polish space $\mathscr{X}$, Markov risk measures for cost evaluation in an MDP were introduced. In a finite-horizon setting, a Markov risk measure evaluates the sequence of discounted costs $\alpha^t c(i_t,u_t)$, $t = 0,1,2,\ldots,T$, under a Markov policy π, in a recursive way. Denoting by $\rho^{\pi}_{t,T}(i)$ the risk of the system starting from state i at time t, we have

$$\rho^{\pi}_{t,T}(i) = c^{\pi}_i + \alpha\sigma_i\big(P^{\pi}_i, \rho^{\pi}_{t+1,T}(\cdot)\big), \quad i \in \mathscr{X}, \quad t = 0,1,\ldots,T-1, \tag{1.8}$$

with $\rho^{\pi}_{T,T}(i) = c^{\pi}_i$, $i \in \mathscr{X}$. In equation (1.8), the operator $\sigma : \mathscr{X} \times \mathscr{P}(\mathscr{X}) \times \mathscr{V} \to \mathbb{R}$, where $\mathscr{P}(\mathscr{X})$ is the space of probability measures on $\mathscr{X}$ and $\mathscr{V}$ is the space of bounded functions on $\mathscr{X}$, is a *transition risk mapping*. It can be interpreted as risk-averse analog of the conditional expectation. Its first argument is the state i (which we write as a subscript). The second argument, the vector P^{π}_i, is the ith row of the matrix P^{π}: the probability distribution of the state following i under the policy π. The last argument, the function $\rho^{\pi}_{t+1,T}(\cdot)$, is the risk of running the system from the next state in the time interval from $t+1$ to T. The transition risk mapping is a special case of a *risk form*: a generalization of a risk measure introduced in [23] to accommodate the dependence of measures of risk on the underlying probability distribution. In the case of controlled Markov systems, this dependence is germane for the analysis.

As in [62], we assume that for each $i \in \mathscr{X}$ and each $P_i^\pi \in \mathscr{P}(\mathscr{X})$, the transition risk mapping $\sigma_i(p,\cdot)$, understood as a function its last argument, satisfies the axioms of a coherent measure of risk [2]. In the axioms below we suppress the argument P_i^π, focusing on the dependence on the third argument, a function of a state:

Convexity: $\sigma_i(\alpha v + (1-\alpha)w) \le \alpha\sigma_i(v) + (1-\alpha)\sigma_i(w)$, $\forall\, \alpha \in [0,1]$, $\forall\, v, w \in \mathscr{V}$;

Monotonicity: If $v \le w$ (componentwise) then $\sigma_i(v) \le \sigma_i(w)$;

Translation equivariance: $\sigma_i(v + \beta\mathbb{1}) = \sigma_i(v) + \beta$, for all $\beta \in \mathbb{R}$;

Positive homogeneity: $\sigma_i(\beta v) = \beta\sigma_i(v)$, for all $\beta \ge 0$.

Under these conditions, one can pass to the limit with $T \to \infty$ in (1.8) and prove the existence of an infinite-horizon discounted risk measure [62]

$$\rho^\pi_{0,\infty}(i) = \lim_{T\to\infty} \rho^\pi_{0,T}(i), \quad i \in \mathscr{X}.$$

We still denote its value at state i by $v^\pi(i)$; it will never lead to misunderstanding. The policy value $v^\pi(\cdot)$ satisfies the risk-averse policy evaluation equation:

$$v^\pi(i) = c^\pi(i) + \alpha\sigma_i\big(P_i^\pi, v^\pi(\cdot)\big), \quad i \in \mathscr{X}.$$

We introduce the space $\mathscr{Q}$ of transition kernels on $\mathscr{X}$, define a vector-valued *transition risk operator* $\sigma : \mathscr{Q} \times \mathscr{V} \to \mathscr{V}$, with components $\sigma_i(P_i^\pi, \cdot)$, $i \in \mathscr{X}$, and rewrite the last equation in a way similar to (1.3):

$$v^\pi = c^\pi + \alpha\sigma(P^\pi, v^\pi). \tag{1.9}$$

The only difference between (1.3) and (1.9) is that the matrix P^π has been replaced by a convex operator σ^π (which still depends on P^π). This operator, although convex, is highly nontrivial to work with compared to the linear operator P^π we observe under risk-neutrality. The risk-neutral case is a special case of (1.9) with $\sigma(P,v) = Pv$. References [62, 11, 12, 27] outline the theory, provide examples and applications.

Coherent risk measures admit a dual representation [64], which in our case can be stated as follows. For every $i \in \mathscr{X}$ a convex, closed and bounded set $A_i(P_i^\pi)$ of probability measures on $\mathscr{X}$ exists, such that

$$\sigma_i(P_i^\pi, v) = \max_{\mu \in A_i(P_i^\pi)} \langle \mu, v\rangle, \quad v \in \mathscr{V}. \tag{1.10}$$

In a risk-neutral case, the set $A_i(P_i^\pi) = \partial\sigma_i(P_i^\pi, 0)$ contains only one element, P_i^π, but in general it is larger and has P_i^π as one of its elements, provided we always have $\sigma_i(P_i^\pi, v) \geq P_i^\pi v$. The multifunction $A : \mathscr{X} \to \mathscr{P}(\mathscr{X}) \rightrightarrows \mathscr{P}(\mathscr{X})$ is called the *risk multikernel*. Every $\mu \in A_i(P_i^\pi)$ is absolutely continuous with respect to P_i^π.

Due to the nonlinearity inherent in transition risk mapping, the equation (1.9) can be solved by a specialized nonsmooth Newton's method [62], and the resulting value function v^π can be used in a policy improvement scheme, in which the new policy $\overline{\pi}$ can be constructed as

$$\overline{\pi}(i) \in \operatorname*{argmin}_{u \in \mathscr{U}(i)} \left\{ c(i,u) + \alpha\sigma_i\left(P_i^u, v^\pi(\cdot)\right) \right\}, \quad i \in \mathscr{X} \tag{1.11}$$

where P_i^u is the transition probability distribution from state i under action u. This new policy $\overline{\pi}$ can be used as input for the next value function, $v^{\overline{\pi}}$. This is the idea behind the risk-averse policy iteration method. However, this technique to evaluate and improve a policy still requires us to be able to enumerate all the states, and it becomes impractical when the dimension of the state space is very large.

Chapter 2

Projected Risk-Averse Dynamic Programming Equation

The impracticality of evaluating large MDPs using traditional dynamic programming methods persists in the risk-averse case as well. To circumvent this, we use the value approximation scheme defined in Section 1.4.1. We define the projected risk-averse dynamic programming equation as following:

$$\Phi r = L\big(c + \alpha\sigma(P, \Phi r)\big). \tag{2.1}$$

The difference from the risk-neutral projected dynamic progamming operation (1.5) is that we replace the conditional expectation with the transition risk mapping $\sigma(P,\cdot)$, which is a convex operator. Still following the expected value case, we assume that the Markov system under policy π is ergodic, and we denote its vector of stationary probabilities by q, and we rely on the projection operator defined as in (1.6).

The fundamental question is the existence and uniqueness of a solution of equation (2.1). This can be answered by establishing the contraction mapping property of the right hand side of (2.1):

$$\mathscr{D}(v) = L\big(c + \alpha\sigma(P, v)\big), \quad v \in \mathscr{V}, \tag{2.2}$$

which would imply the existence and uniqueness of a solution of the equation

$$v = \mathscr{D}v. \tag{2.3}$$

Crucial in this context is the *distortion coefficient* of the risk multikernel A:

$$\varkappa = \max\left\{ \frac{|\mu_{ij} - p_{ij}|}{p_{ij}} : \mu_i \in A_i(P_i^\pi),\ p_{ij} > 0,\ i, j \in \mathscr{X} \right\}.$$

By definition, $\varkappa \geq 0$, with the value 0 corresponding to the risk-neutral model. We also recall that for $p_{ij} = 0$ we always have $m_{ij} = 0$, for all $m_i \in A_i(P_i^\pi)$.

Lemma 1. *The transition risk operator satisfies for all $w, v \in \mathscr{V}$ the inequalities:*

$$\|\sigma(P,w) - \sigma(P,v)\|_q \le \sqrt{1+\varkappa}\,\|w-v\|_q, \tag{2.4}$$

and

$$\|\sigma(P,w) - \sigma(P,v) - P(w-v)\|_q \le \varkappa\,\|w-v\|_q. \tag{2.5}$$

Proof. For brevity, we omit the argument P of $\sigma(P,\cdot)$, because it is fixed. For every $i = 1,\dots,n$, by the mean value theorem for convex functions [77, 33], a point $\bar{v}^{(i)} = (1-\theta_i)v + \theta_i w$ exists, with $\theta_i \in [0,1]$, and a subgradient $m_i \in \partial\rho_i(\bar{v}^{(i)})$ exists, such that

$$\sigma_i(w) - \sigma_i(v) = \langle m_i, w-v\rangle.$$

Since the subdifferential $\partial\rho_i(\cdot) \subseteq A_i$, we have $m_i \in A_i$. Therefore, for a matrix M having m_i, $i = 1,\dots,n$, as its rows,

$$\sigma(w) - \sigma(v) = M(w-v). \tag{2.6}$$

As each m_i is a probability vector, Jensen's inequality with $h = w - v$, and the equation $q^\top P = q^\top$ yield

$$\begin{aligned}\|Mh\|_q^2 = \sum_{i\in\mathscr{X}} q_i\Big(\sum_{j\in\mathscr{X}} m_{ij}h_j\Big)^2 &\le \sum_{i\in\mathscr{X}} q_i \sum_{j\in\mathscr{X}} m_{ij}h_j^2 \\ &\le (1+\varkappa)\sum_{i\in\mathscr{X}} q_i \sum_{j\in\mathscr{X}} p_{ij}h_j^2 = (1+\varkappa)\sum_{j\in\mathscr{X}} q_j h_j^2 = (1+\varkappa)\|h\|_q^2. \end{aligned}\tag{2.7}$$

The last two relations imply (2.4). In a similar way, it follows from (2.6) that

$$\begin{aligned}\big\|\sigma(P,w) - \sigma(P,v) - P(w-v)\big\|_q^2 = \|(M-P)h\|_q^2 &\le \sum_{i\in\mathscr{X}} q_i\Big(\sum_{j\in\mathscr{X}} |m_{ij}-p_{ij}||h_j|\Big)^2 \\ &\le \varkappa^2 \sum_{i\in\mathscr{X}} q_i\Big(\sum_{j\in\mathscr{X}} p_{ij}|h_j|\Big)^2 \le \varkappa^2 \sum_{i\in\mathscr{X}} q_i \sum_{j\in\mathscr{X}} p_{ij}|h_j|^2 = \varkappa^2\,\|w-v\|_q^2,\end{aligned}$$

which is (2.5). □

Theorem 2. *If $\alpha\sqrt{1+\varkappa} < 1$ then the equation* (2.3) *has a unique solution v^*.*

Proof. We verify that the operator (2.2) is a contraction mapping in the norm $\|\cdot\|_q$. The orthogonal projection L is nonexpansive. The operator P is nonexpansive in the norm $\|\cdot\|_q$ as well (this is a special case of (2.7) with $M = P$ and $\varkappa = 0$). The transition risk operator $\sigma(\cdot)$ multiplied by α is a contraction by Lemma 1. The assertion follows now from the Banach contraction mapping theorem. □

If Φ has full column rank, equation (2.1) has a unique fixed point as well.

2.1 Simulation approach

The equation (2.1) cannot be solved using traditional dynamic programming methods due to the state space being too large. However, it is possible to find r that satisfies (2.1) by means of sampling and simulation, using a method analogous to the one described in Section 1.4.2. From (2.1), we infer that

$$r^* = \underset{r}{\operatorname{argmin}} ||\Phi r - (c + \alpha\sigma(\Phi r^*)||_q^2.$$

This implies for r the optimality condition

$$\begin{aligned} \Phi^T Q[\Phi r - (c + \alpha\rho(\Phi r^*)] &= 0 \\ \Phi^T Q\Phi r - \Phi^T Qc - \alpha\Phi^T Q\sigma(\Phi r^*) &= 0. \end{aligned} \tag{2.8}$$

We can also state the last equation using the dual representation of $\sigma(\cdot)$,

$$\Phi^T Q\Phi r - \Phi^T Qc - \alpha\Phi^T QM\Phi r^* = 0.$$

Let $A = \Phi^T Q\Phi$ and $b = \Phi^T Qc$. It is possible to approximate A and b by simulation by running a trajectory of states $\{i_t\}_{t=1}^N$. We know that

$$\begin{aligned} A \approx A_N &= \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)\phi^T(i_t) \\ b \approx b_N &= \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)c(i_t). \end{aligned}$$

However, the component $\Phi^T Q\sigma(\Phi r^*)$ in (2.8) is not straightforward to simulate. In the risk-neutral case described in Section 1.4.2, we simulated $\Phi^T QP\Phi r^*$, and we were able to substitute the expected feature vector of the next state with simply the next state on the trajectory. However, in order to assess risk, we need r, we cannot simply use features. At each state during the trajectory, we need to simulate, say K, next states and compute an empirical risk measure $\widetilde{\sigma}(\cdot)$ of the next state. For a given r, we can state

$$\Phi^T Q\rho(\Phi r) \approx \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)\widetilde{\sigma}(\phi(i_{t+1})r|i_t).$$

Now let

$$B_N r = \Phi^T Q M_N \Phi r = \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)\widetilde{\sigma}(\phi(i_{t+1})r|i_t).$$

Then for a given r, for the residuals of the equation, we can state

$$\begin{aligned}(A_N - \alpha B_N)r - b_N &= \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)\phi(i_t)^T r - \alpha\phi(i_t)\widetilde{\sigma}(\phi^T(i_{t+1})r|i_t) - \phi(i_t)c(i_t) \\ &= \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)\underbrace{\left[\phi^T(i_t)r - \alpha\widetilde{\sigma}(\phi^T(i_{t+1})r|i_t) - c(i_t)\right]}_{d_t} \\ &= \frac{1}{N}\sum_{t=1}^{N} \phi(i_t)d_t\end{aligned}$$

where d_t is the temporal difference. The residual error can be corrected over a large trajectory incrementally using the temporal differences. This idea forms the basis of the risk-averse methods of temporal differences which are explored in the following sections.

Chapter 3

The Risk-Averse Methods of Temporal Differences

3.1 The Single-Step Risk-Averse Methods of Temporal Differences

We propose to solve (2.1) by a risk-averse analog of the classical method of temporal differences [68]. We define v^* to be the solution of equation (2.3). We know that it exists and is unique, if $\alpha\sqrt{1+\varkappa} < 1$.

Consider the evolution of the system under policy π, resulting in a random trajectory of states i_t, $t = 0,1,2\ldots$. At each time t, we have an approximation r_t of a solution of the equation (2.1). Let $\mathscr{F}_t$ be the σ-algebra defined by all observations gathered up to time t.

The difference between the left and the right hand sides of equation (2.1) with coefficient values r_t and state i_t is the *risk-averse temporal difference*:

$$d_t = \varphi^\top(i_t) r_t - c(i_t) - \alpha\sigma_{i_t}(P_{i_t}, \Phi r_t), \quad t = 0,1,2,\ldots. \tag{3.1}$$

Evidently, it cannot be easily computed or observed; this would require the evaluation of the risk $\sigma_{i_t}(P_{i_t}, v)$ and thus consideration of *all* possible transitions from state i_t. Instead, we assume that we can observe a random estimate $\widetilde{\sigma}_{i_t}(P_{i_t}, \cdot)$, such that

$$\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t) = \sigma_{i_t}(P_{i_t}, \Phi r_t) + \xi_t, \quad t = 0,1,2,\ldots, \tag{3.2}$$

with some random errors ξ_t. The conditions on $\{\xi_t\}$ will be specified later. This allows us to define the observed risk-averse temporal differences,

$$\widetilde{d}_t = \varphi^\top(i_t) r_t - c(i_t) - \alpha\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t), \quad t = 0,1,2,\ldots, \tag{3.3}$$

and to construct the *risk-averse temporal difference method* as follows:

$$r_{t+1} = r_t - \gamma_t \varphi(i_t)\, \widetilde{d}_t, \quad t = 0,1,2,\ldots. \tag{3.4}$$

Before proceeding to the detailed convergence proof in the stochastic case, we analyze a deterministic model of the method, in which the errors ξ_t are ignored and the updates of the sequence $\{r_t\}$ are averaged over all states (with the distribution q). We define the operator:

$$U(r) = \mathbb{E}_{i\sim q}\big[\varphi(i)\big(\varphi^\top(i)r - c(i) - \alpha\sigma_i(P_i, \Phi r)\big)\big] = \Phi^\top Q\big[\Phi r - c - \alpha\,\sigma(P, \Phi r)\big]. \tag{3.5}$$

The deterministic analog of (3.3)–(3.4) reads:

$$\bar r_{t+1} = \bar r_t - \gamma U(r_t), \quad t = 0, 1, 2, \ldots, \quad \gamma > 0. \tag{3.6}$$

By the definition of the projection operator L, a point r^* is a solution of (2.1) if and only if

$$r^* = \operatorname*{argmin}_r \frac{1}{2}\big\|\Phi r - \big(c + \alpha\sigma(P, \Phi r^*)\big)\big\|_q^2.$$

This occurs if and only if r^* is a zero of $U(\cdot)$ and thus supports our idea of using the method (3.3)–(3.4).

Theorem 3. *If $\alpha\sqrt{1+\varkappa} < 1$, then $\gamma_0 > 0$ exists, such that for all $\gamma \in (0, \gamma_0)$ the algorithm* (3.6) *generates a sequence $\{\bar r_t\}$ convergent to a point r^* such that $U(r^*) = 0$.*

Proof. We shall show that for sufficiently small $\gamma > 0$ the operator $I - \gamma U$ is a contraction. For arbitrary r' and r'', we have

$$\begin{aligned}\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 &= \|r' - r''\|^2 \\ &- 2\gamma\langle r' - r'', \Phi^\top Q\Phi(r' - r'')\rangle + 2\gamma\alpha\big\langle r' - r'', \Phi^\top Q\big[\sigma(P, \Phi r') - \sigma(P, \Phi r'')\big]\big\rangle \\ &+ \gamma^2\Big\|\Phi^\top Q\Phi(r' - r'') - \alpha\Phi^\top Q\big[\sigma(P, \Phi r') - \sigma(P, \Phi r'')\big]\Big\|^2.\end{aligned}$$

The last term (with γ^2) can be bounded by $\gamma^2 C\|\Phi(r' - r'')\|_q^2$ where C is some constant. Then

$$\begin{aligned}\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 &\le \|r' - r''\|^2 - 2\gamma\|\Phi(r' - r'')\|_q^2 \\ &+ 2\gamma\alpha\big\langle\Phi(r' - r''), \sigma(\Phi r') - \sigma(\Phi r'')\big\rangle_q + \gamma^2 C\|\Phi(r' - r'')\|_q^2.\end{aligned}$$

The scalar product can be bounded by (2.4), and thus

$$\begin{aligned}&\big\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\big\|^2 \\ &\quad\le \|r' - r''\|^2 - 2\gamma\|\Phi(r' - r'')\|_q^2 + 2\gamma\alpha\sqrt{1+\varkappa}\|\Phi(r' - r'')\|_q^2 + \gamma^2 C\|\Phi(r' - r'')\|_q^2 \\ &\quad= \|r' - r''\|^2 - 2\gamma\left(1 - \alpha\sqrt{1+\varkappa} + \frac{\gamma C}{2}\right)\|\Phi(r' - r'')\|_q^2.\end{aligned}$$

Since $\alpha\sqrt{1+\varkappa} < 1$, then using $0 < \gamma < 2(1-\alpha\sqrt{1+\varkappa})/C$, we have

$$\left\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\right\|^2 \le \|r' - r''\|^2 - \gamma\beta\|\Phi(r' - r'')\|_q^2, \tag{3.7}$$

with some $\beta > 0$. In particular, setting $r' = \bar{r}_t$ and $r'' = r^*$ for a solution r^* of (2.1), we obtain the following relation between the successive iterates of the method (3.6):

$$\|\bar{r}_{t+1} - r^*\|^2 \le \|\bar{r}_t - r^*\|^2 - \gamma\beta\|\Phi(\bar{r}_t - r^*)\|_q^2. \tag{3.8}$$

This immediately proves that the sequence $\{\bar{r}_t\}$ is bounded and $\Phi\bar{r}_t \to \Phi r^*$. Every accumulation point $\hat{r}$ of $\{\bar{r}_t\}$ must be then a solution of equation (2.1). Substituting this accumulation point for r^* in the last inequality, we conclude that $\bar{r}_t \to \hat{r}$. □

If Φ has full column rank, the solution r^* is unique, because substituting another solution for $\bar{r}_t$ in (3.6) we obtain $r_{t+1} = r_t$, which leads to a contradiction in (3.8). The detailed proof in the stochastic case is provided in Section 4.1.

3.2 The Multi-Step Risk-Averse Methods of Temporal Differences

In the method discussed so far, the residuals are corrected by moving in the direction of the last feature vector $\varphi(i_t)$. Alternatively, we may use the weighted averages of all previous observations, where the highest weight is given to the most recent observation and the weights decrease exponentially as we look into the past observations. This idea is the core of the well-known TD(λ) algorithm [68]. We generalize it to the risk-averse case.

For a fixed policy π, we refer to v^π as v, and to P^π as P, for simplicity. The multi-step risk-averse method of temporal differences carries out the following iterations:

$$z_t = \lambda\alpha z_{t-1} + \varphi(i_t), \quad t = 0,1,2,\ldots, \tag{3.9}$$

$$r_{t+1} = r_t - \gamma_t z_t \widetilde{d}_t, \quad t = 0,1,2,\ldots \tag{3.10}$$

where $\lambda \in [0,1]$, and $\widetilde{d}_t$ is given by (3.3). For simplicity, z_{-1} is assumed to be the zero vector. In the risk-neutral case, when $\sigma_{i_t}(P_{i_t}, \Phi r_t) = P_{i_t}\Phi r_t$, the method reduces to the classical TD(λ).

Our convergence analysis will use some ideas from the analysis in the previous two sections, albeit in a form adapted to the version with exponentially averaged features. However, contrary

to the expected value setting, the method (3.9)–(3.10) will converge to a solution of an equation different from (2.3), but still relevant for our problem.

We start from a heuristic analysis of a deterministic counterpart of the method, to extract its drift. In the next section, we make all approximations precise, but we believe that this introduction is useful to decipher our detailed approach to follow. By direct calculation,

$$z_t = \sum_{k=0}^{t} (\lambda\alpha)^{t-k} \varphi(i_k), \tag{3.11}$$

and thus

$$z_t d_t = \Phi^\top \sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top \big(\Phi r_t - c - \alpha\sigma(P, \Phi r_t)\big).$$

Heuristically assuming that $r_t \approx r'$, we focus on the operator acting on the expected temporal differences. As each of the observed feature vectors $\varphi(i_k)$ affects all succeeding steps of the method, via the filter (3.9), we need to study the cumulative effect of many steps. We look, therefore, at the sums

$$G_T = \mathbb{E}\bigg[\sum_{t=0}^{T} \gamma_t \sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top\bigg].$$

Changing the order of summation and using the fact that $\{(\lambda\alpha)^{t-k}\}_{t\geq k}$ diminishes very fast, as compared to $\{\gamma_t\}_{t\geq k}$, we get

$$G_T = \mathbb{E}\bigg[\sum_{k=0}^{T}\sum_{t=k}^{T} \gamma_t (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top\bigg] \approx \mathbb{E}\bigg[\sum_{k=0}^{T} \gamma_k \sum_{t=k}^{T} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top\bigg].$$

Therefore

$$\begin{aligned}
G_T &\approx \mathbb{E}\bigg[\sum_{k=0}^{T} \gamma_k \sum_{t=k}^{T} (\lambda\alpha)^{t-k} e_{i_k} \mathbb{E}\big[e_{i_t}^\top \,\big|\, \mathscr{F}_k\big]\bigg] = \mathbb{E}\bigg[\sum_{k=0}^{T} \gamma_k \sum_{t=k}^{T} (\lambda\alpha)^{t-k} e_{i_k} e_{i_k}^\top P^{t-k}\bigg] \\
&= \sum_{k=0}^{T} \gamma_k \mathbb{E}[\mathrm{diag}(e_{i_k})] \sum_{t=k}^{T} (\lambda\alpha)^{t-k} P^{t-k} \approx \sum_{k=0}^{T} \gamma_k \mathbb{E}[\mathrm{diag}(e_{i_k})] \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} P^{t-k} \\
&\approx Q \sum_{k=0}^{T} \gamma_k \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} P^{t-k}.
\end{aligned}$$

The last approximations are possible because $\lambda\alpha \in [0,1)$ and $\mathbb{E}[\mathrm{diag}(e_{i_k})] \to q$ at an exponential rate. We now define the multi-step transition matrix,

$$\bar{P} = (1-\lambda\alpha) \sum_{\ell=0}^{\infty} (\lambda\alpha)^\ell P^\ell. \tag{3.12}$$

By construction, $\bar{P} \in \overline{\mathrm{conv}}\{I, P, P^2, \dots\}$. With these approximations, we can simply write

$$G_T \approx \frac{1}{1-\lambda\alpha} Q\bar{P} \sum_{k=0}^{T} \gamma_k.$$

Define the operators

$$\overline{U}(r) = \Phi^\top Q\bar{P}\big[\Phi r - c - \alpha\,\sigma(P,\Phi r)\big], \quad t = 0,1,2,\ldots \tag{3.13}$$

and consider the following deterministic counterpart of (3.9)–(3.10), with $\bar{\gamma} \sim \gamma_t/(1-\lambda\alpha)$:

$$r_{t+1} = r_t - \bar{\gamma}\overline{U}(r_t), \quad t = 0,1,2,\ldots, \quad \bar{\gamma} > 0. \tag{3.14}$$

Our intention is to show that for sufficiently small $\bar{\gamma}$ the method (3.14) converges to a point r^* such that $\overline{U}(r^*) = 0$. Such a point is also a solution of the following projected multi-step risk-averse dynamic programming equation:

$$L\bar{P}\Phi r = L\bar{P}\big(c + \alpha\sigma(P,\Phi r)\big), \tag{3.15}$$

where L is the projection operator defined in (1.6). The solutions of (3.15) differ from the solutions of (2.1), unlike in the risk-neutral case ($\varkappa = 0$). If we replace $\bar{P}$ with I, (3.15) reduces to (2.1).

Theorem 4. *If $\alpha(1+\varkappa) < 1$, then $\bar{\gamma}_0 > 0$ exists, such that for all $\bar{\gamma} \in (0,\bar{\gamma}_0)$ the algorithm* (3.14) *generates a sequence $\{r_t\}$ convergent to a point r^* such that $\overline{U}(r^*) = 0$.*

Proof. For two arbitrary points r' and r'' we have

$$\begin{aligned}\Big\|\big(r' - \bar{\gamma}\overline{U}(r')\big) - \big(r'' - \bar{\gamma}\overline{U}(r'')\big)\Big\|^2 \\ = \|r' - r''\|^2 \;\; + 2\bar{\gamma}\Big\langle r' - r'', \Phi^\top Q\bar{P}\big[-\Phi(r'-r'') + \alpha\sigma(P,\Phi r') - \alpha\sigma(P,\Phi r'')\big]\Big\rangle \\ + \bar{\gamma}^2\Big\|\Phi^\top Q\bar{P}\Phi(r'-r'') - \alpha\Phi^\top Q\bar{P}\big[\sigma(P,\Phi r') - \sigma(P,\Phi r'')\big]\Big\|^2.\end{aligned} \tag{3.16}$$

We focus on the scalar product in the middle of the right hand side of (3.16):

$$\begin{aligned}\Big\langle \Phi(r'-r''), \bar{P}\big[-\Phi(r'-r'') + \alpha\sigma(P,\Phi r') - \alpha\sigma(P,\Phi r'')\big]\Big\rangle_q \\ = \Big\langle \Phi(r'-r''), \bar{P}\big[-\Phi(r'-r'') + \alpha P\Phi(r'-r'')\big]\Big\rangle_q \\ + \alpha\Big\langle \Phi(r'-r''), \bar{P}\big[\sigma(P,\Phi r') - \sigma(P,\Phi r'') - P\Phi(r'-r'')\big]\Big\rangle_q.\end{aligned} \tag{3.17}$$

Setting $h = \Phi(r' - r'')$, we can estimate the first (quadratic) term on the right hand side of (3.17) by a calculation borrowed from [74, Lem. 8], with $h = \Phi(r' - r'')$:

$$\begin{aligned}
\Big\langle h, \bar{P}\big[-h + \alpha P h\big]\Big\rangle_q &= (1-\alpha\lambda)\Big\langle h, \sum_{\ell=0}^{\infty} (\alpha\lambda)^\ell P^\ell \big[-h + \alpha P h\big]\Big\rangle_q \\
&= (1-\alpha\lambda)(1-\lambda)\Big\langle h, \sum_{k=0}^{\infty} \lambda^k \sum_{\ell=0}^{k} \alpha^\ell P^\ell \big[-h + \alpha P h\big]\Big\rangle_q \\
&= (1-\alpha\lambda)(1-\lambda)\Big\langle h, \sum_{k=0}^{\infty} \lambda^k \big[\alpha^{k+1} P^{k+1} h - h\big]\Big\rangle_q \\
&= (1-\alpha\lambda)\Big\langle h, (1-\lambda)\sum_{k=0}^{\infty} \lambda^k \alpha^{k+1} P^{k+1} h - h\Big\rangle_q \\
&= (1-\alpha\lambda)\Big\langle h, \frac{\alpha(1-\lambda)}{1-\alpha\lambda} P\bar{P} h - h\Big\rangle_q \le (\alpha-1)\|h\|_q^2.
\end{aligned}$$

The last inequality is due to the fact that both P and $\bar{P}$ are nonexpansive in $\|\cdot\|_q$.

The second (nonsmooth) term on the right hand side of (3.17) can be estimated by (2.5), again with the use of the nonexpansiveness of $\bar{P}$:

$$\Big\langle \Phi(r' - r''), \bar{P}\big[\sigma(P, \Phi r') - \sigma(P, \Phi r'') - P\Phi(r' - r'')\big]\Big\rangle_q \le \varkappa \big\|\Phi(r' - r'')\big\|_q^2.$$

The last term on the right hand side of (3.16) (with γ^2) can be bounded by $\gamma^2 \bar{C}\|\Phi(r' - r'')\|_q^2$, where $\bar{C}$ is some constant. Integrating all these estimates into (3.16), we obtain the inequality

$$\big\|(I - \bar{\gamma}\bar{U})(r') - (I - \bar{\gamma}\bar{U})(r'')\big\|^2 \le \|r' - r''\|^2 - 2\bar{\gamma}\Big(1 - \alpha(1+\varkappa) - \frac{\bar{\gamma}\bar{C}}{2}\Big)\big\|\Phi(r' - r'')\big\|_q^2.$$

If $\alpha(1+\varkappa) < 1$, then using $0 < \bar{\gamma} < 2(1 - \alpha(1+\varkappa))/\bar{C}$, we obtain:

$$\big\|(I - \bar{\gamma}\bar{U})(r') - (I - \bar{\gamma}\bar{U})(r'')\big\|^2 \le \|r' - r''\|^2 - \bar{\gamma}\beta\|\Phi(r' - r'')\|_q^2, \tag{3.18}$$

with some $\beta > 0$. In particular, setting $r' = \bar{r}_t$ and $r'' = r^*$ for a solution r^* of (2.1), we obtain the following relation between successive iterates of the method (3.14):

$$\|r_{t+1} - r^*\|^2 \le \|r_t - r^*\|^2 - \bar{\gamma}\beta\|\Phi(r_t - r^*)\|_q^2. \tag{3.19}$$

This immediately proves that the sequence $\{r_t\}$ is bounded and $\Phi r_t \to \Phi r^*$. Every accumulation point $\hat{r}$ of $\{r_t\}$ must be then a solution of equation (3.15). Substituting this accumulation point for r^* in the last inequality, we conclude that the entire sequence $\{r_t\}$ is convergent to $\hat{r}$. □

The proof for the stochastic case is provided in Section 4.2.

Chapter 4

Convergence Analysis

4.1 Convergence Analysis of the Single-Step Risk-Averse Methods of Temporal Differences

We shall use the following result on convergence of deterministic nonmonotonic algorithms [49].

Theorem 5. *Let $Y^* \subset \mathbb{R}^m$. Suppose $\{r_t\} \subset \mathbb{R}^m$ is a bounded sequence which satisfies the following assumptions:*

A) If a subsequence $\{r_t\}_{t\in\mathcal{K}}$ converges to $r' \in Y^*$, then $\|r_{t+1} - r_t\| \to 0$, as $t \to \infty$, $t \in \mathcal{K}$;

B) If a subsequence $\{r_t\}_{t\in\mathcal{K}}$ converged to $r' \notin Y^*$, then $\varepsilon_0 > 0$ would exist such that for all $\varepsilon \in (0, \varepsilon_0]$ and for all $k \in \mathcal{K}$, the index $s(t,\varepsilon) = \min\{\ell \geq k : \|r_\ell - r_t\| > \varepsilon\}$ would be finite;

C) A continuous function $W : \mathbb{R}^m \to \mathbb{R}$ exists such that if $\{r_t\}_{t\in\mathcal{K}}$ converged to $r' \notin Y^*$ then $\varepsilon_1 > 0$ would exist such that for all $\varepsilon \in (0, \varepsilon_1]$ we would have

$$\limsup_{t\in\mathcal{K}} W(r_{s(t,\varepsilon)}) < W(r'),$$

where $s(t,\varepsilon)$ is defined in B);

D) The set $\{W(r) : r \in Y^*\}$ does not contain any segment of nonzero length.

Then the sequence $\{W(r_t)\}$ is convergent and all limit points of the sequence $\{r_t\}$ belong to Y^*.

We define the set of solutions of equation (2.1):

$$Y^* = \{r \in \mathbb{R}^m : \Phi r = v^*\},$$

where v^* is the unique solution of (2.3), provided $\alpha\sqrt{1+\kappa} < 1$ We shall show that the method (3.4) converges to Y^*, under the above-mentioned condition and some additional conditions on the stepsizes $\{\gamma_t\}$ and errors $\{\xi_t\}$.

We define $\mathscr{F}_t$ to be the σ-algebra generated by $\{i_0, r_0, \ldots, i_t, r_t\}$, $t = 0, 1, \ldots$, and make the following assumptions about the stepsize and error sequences. We allow the stepsizes to be random.

Assumption 1. *The sequence $\{\gamma_t\}$ is adapted to the filtration $\{\mathscr{F}_t\}$ and such that*

(i) $\gamma_t > 0$, $t = 0, 1, \ldots$, *and* $\lim_{t\to\infty} \gamma_t = 0 \quad a.s.;$

(ii) $\sum_{t=0}^{\infty} \gamma_t = \infty \quad a.s.;$

(iii) $\mathbb{E} \sum_{t=0}^{\infty} \gamma_t^2 < \infty;$

(iv) *For any* $\varepsilon > 0$, $\lim\limits_{t_0\to\infty} \sup\limits_{\{T:\sum_{t=t_0}^{T} \gamma_t \le \varepsilon\}} \sum\limits_{t=t_0}^{T} |\gamma_t - \gamma_{t+1}| = 0 \quad a.s.$

Assumption 2. *The sequence of errors $\{\xi_t\}_{t\ge 1}$ satisfies for $t = 0, 1, 2 \ldots$ the conditions*

(i) $\mathbb{E}[\xi_t \,|\, \mathscr{F}_t] = 0 \quad a.s.;$

(ii) $\mathbb{E}[\|\xi_t\|^2 \,|\, \mathscr{F}_t] \le \text{const} \quad a.s..$

First, we establish an important implication of the ergodicity of the chain. We write e_i for the ith unit vector in $\mathbb{R}^n$.

Lemma 6. *If the chain $\{i_t\}$ is ergodic with stationary distribution q and Assumption 1 is satisfied, then*

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma_t (e_{i_t} - q)}{\sum_{t=0}^{T} \gamma_t} = 0, \quad a.s., \tag{4.1}$$

and for any $\varepsilon > 0$,

$$\lim_{t_0\to\infty} \sup_{T\ge t_0} \frac{\sum_{t=t_0}^{T} \gamma_t (e_{i_t} - q)}{\max\left(\varepsilon, \sum_{t=t_0}^{T} \gamma_t\right)} = 0, \quad a.s.. \tag{4.2}$$

Proof. Due to the ergodicity of the chain, the vectors

$$\nu(i) = \mathbb{E}\left[\sum_{t=0}^{\infty} (e_{i_t} - q) \,\middle|\, i_0 = i\right], \quad i \in \mathscr{X},$$

are finite and satisfy the *Poisson equation*

$$\nu(i) = e_i - q + \sum_{j\in\mathscr{X}} P_{ij}\nu(j), \quad i \in \mathscr{X}. \tag{4.3}$$

Consider the sums $\sum_{t=0}^{T} \gamma_t (e_{i_t} - q)$. By the Poisson equation,

$$e_{i_t} - q = \nu(i_t) - \sum_{j\in\mathscr{X}} P_{i_t j}\nu(j) = \left[\nu(i_t) - \nu(i_{t+1})\right] + \left[\nu(i_{t+1}) - \sum_{j\in\mathscr{X}} P_{i_t j}\nu(j)\right]. \tag{4.4}$$

We consider the two components of the right hand side of (4.4), marked with brackets, separately. Due to Assumption 1, (i)—(iii), the series

$$\sum_{t=1}^{\infty} \gamma_t \Big[\nu(i_{t+1}) - \sum_{j \in \mathscr{X}} P_{i_t j} \nu(j) \Big] = \sum_{t=1}^{\infty} \gamma_t \big(\nu(i_{t+1}) - \mathbb{E}[\nu(i_{t+1}) \,|\, \mathscr{F}_t] \big)$$

is a convergent martingale. Therefore,

$$\lim_{T \to \infty} \frac{\sum_{t=0}^{T} \gamma_t \big(\nu(i_{t+1}) - \mathbb{E}_t[\nu(i_{t+1})] \big)}{\sum_{t=0}^{T} \gamma_t} = 0, \quad \text{a.s.}$$

We now focus on the sums

$$\sum_{t=0}^{T} \gamma_t \big[\nu(i_t) - \nu(i_{t+1}) \big] = \gamma_0 \nu(i_0) + \sum_{t=1}^{T} (\gamma_t - \gamma_{t-1}) \nu(i_t) - \gamma_T \nu(i_{T+1}).$$

Using Assumption 1(iv) and [65, Lem. A.3], we obtain (4.1)–(4.2). □

We can now prove the convergence of the method.

Theorem 7. *Suppose the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t}, \Phi r_t)$ satisfy* (3.2)*, Assumptions 1 and 2 are satisfied, and $\alpha\sqrt{1+\varkappa} < 1$. If the sequence $\{r_t\}$ is bounded with probability 1, then every accumulation point of the sequence $\{r_t\}$ is an element of Y^*, with probability 1.*

Proof. We use the global Lyapunov function:

$$W(r) = \min_{r^* \in Y^*} \|r - r^*\|^2. \tag{4.5}$$

The direction used in (3.4) at step t can be represented as

$$\varphi(i_t) \widetilde{d}_t = U(r_t) + \Delta_t, \tag{4.6}$$

with the operator $U(\cdot)$ defined in (3.5), and

$$\Delta_t = -\alpha \xi_t \varphi(i_t) + \Phi^{\top} \operatorname{diag}\big(e_{i_t} - q\big) \big[\Phi r_t - c - \alpha \sigma(P, \Phi r_t) \big]. \tag{4.7}$$

Our intention is to verify the conditions of Theorem 5 for almost all paths of the sequence $\{r_t\}$. For this purpose, we estimate the decrease of the function (4.5) in iteration t. For any $r^* \in Y^*$ we have:

$$\|r_{t+1} - r^*\|^2 = \|r_t - \gamma_t U(r_t) - r^*\|^2 - 2\gamma_t \langle \Delta_t, r_t - \gamma_t U(r_t) - r^* \rangle + \gamma_t^2 \|\Delta_t\|^2.$$

The term involving $U(r_t)$ was estimated in the derivation of (3.8). We obtain the inequality

$$\|r_{t+1}-r^*\|^2 \le \|r_t-r^*\|^2 - 2\gamma_t(1-\alpha\sqrt{1+\varkappa})\|\Phi(r_t-r^*)\|_q^2 - 2\gamma_t\langle\Delta_t, r_t-\gamma_t U(r_t)-r^*\rangle + C\gamma_t^2. \tag{4.8}$$

Now we can verify the conditions of Theorem 5 for almost all paths of the sequence $\{r_t\}$.

Condition A. Due to the boundedness of $\{r_t\}$ the sequence $\{U(r_t)\}$ is bounded as well. In view of (4.6), it is sufficient to verify that $\gamma_t\xi_t \to 0$. By Assumption 2(i), the sequence

$$S_T = \sum_{t=0}^{T}\gamma_t\xi_t, \quad T=0,1,2,\dots, \tag{4.9}$$

is a martingale. Due to Assumption 2(ii), $\mathbb{E}[S_T^2] \le \text{const}\cdot\mathbb{E}[\sum_{t=0}^{T}\gamma_t^2]$. In view of Assumption 1(ii), by virtue of the martingale convergence theorem, $\{S_T\}$ is convergent a.s., which yields $\lim_{t\to\infty}\gamma_t\xi_t = 0$.

Condition B. Suppose $r_k \to r' \notin Y^*$ for $k \in \mathscr{K}$ (on a certain path ω). If B were false, then for all $\varepsilon_0 > 0$ we could find $\varepsilon \in (0,\varepsilon_0]$ and $k \in \mathscr{K}$ such that $\|r_t - r_k\| \le \varepsilon$ for all $t \ge k$. Then for all $k_0 \in \mathscr{K}$, $k_0 \ge k$, we have $\|r_t - r_{k_0}\| \le 2\varepsilon$ for all $t \ge k_0$. Since r' is not optimal, we can choose $\varepsilon_0 > 0$ small enough, $k_0 \in \mathscr{K}$ large enough, and $\delta > 0$ small enough, so that $\|\Phi(r_t-r^*)\|_q^2 > \delta$ for all $t \ge k_0$. Then (4.8) yields

$$\begin{aligned}\|r_T-r^*\|^2 &\le \|r_{k_0}-r^*\|^2 \\ &+\left(-\delta(1-\alpha\sqrt{1+\varkappa}) + \frac{\sum_{t=k_0}^{T-1}\gamma_t\langle\Delta_t, r_t-\gamma_t U(r_t)-r^*\rangle}{\sum_{t=k_0}^{T-1}\gamma_t} + C\frac{\sum_{t=k_0}^{T-1}\gamma_t^2}{\sum_{t=k_0}^{T-1}\gamma_t}\right)\sum_{t=k_0}^{T-1}\gamma_t.\end{aligned} \tag{4.10}$$

We fix $r^* = \text{Proj}_{Y^*}(r_{k_0})$ and estimate the growth of the sums involving Δ_t. We write $\Delta_t = \Delta_t^{(1)} + \Delta_t^{(2)}$, where, in view of (4.7),

$$\Delta_t^{(1)} = -\alpha\xi_t\varphi(i_t), \quad \Delta_t^{(2)} = \Phi^\top \text{diag}\left(e_{i_t}-q\right)\left[\Phi r_t - c - \alpha\sigma(P,\Phi r_t)\right].$$

Since (4.9) is a convergent martingale and the terms $\langle\varphi(i_t), r_t-\gamma_t U(r_t)-r^*\rangle$ are bounded and $\mathscr{F}_t$-measurable, we have

$$\lim_{T\to\infty}\frac{\left|\sum_{t=k_0}^{T-1}\gamma_t\langle\Delta_t^{(1)}, r_t-\gamma_t U(r_t)-r^*\rangle\right|}{\sum_{t=k_0}^{T-1}\gamma_t} = 0.$$

To deal with the sum involving $\Delta_t^{(2)}$, observe that $\|r_t - r_{k_0}\| \le 2\varepsilon_0$ and thus

$$\begin{aligned}&\langle\Delta_t^{(2)}, r_t-\gamma_t U(r_t)-r^*\rangle = \\ &\quad\left\langle \text{diag}\left(e_{i_t}-q\right)\left[\Phi r_{k_0} - c - \alpha\sigma(P,\Phi r_{k_0})\right], \Phi(r_{k_0}-r^*)\right\rangle + h_t = \langle e_{i_t}-q, w\rangle + h_t,\end{aligned} \tag{4.11}$$

where $|h_t| \le C\varepsilon_0$ and w is a fixed vector (depending on k_0 only). It follows that

$$\left|\sum_{t=k_0}^{T-1} \gamma_t \langle \Delta_t^{(2)}, r_t - \gamma_t U(r_t) - r^* \rangle\right| \le C \left\|\sum_{t=k_0}^{T-1} \gamma_t (e_{i_t} - q)\right\| + C\varepsilon_0 \sum_{t=k_0}^{T-1} \gamma_t. \tag{4.12}$$

Dividing both sides of (4.12) by $\sum_{t=k_0}^{T-1} \gamma_t$ and using (4.1), we see that we can choose $\varepsilon > 0$ small enough and $k_0 \in \mathscr{K}$ large enough, so that the entire expression in parentheses in (4.10) is smaller than $-\delta(1-\alpha\sqrt{1+\varkappa})/2$, if T is large enough. But this yields $\|r_T - r^*\| \to -\infty$, as $T \to \infty$, a contradiction. Therefore, Condition B is satisfied.

Condition C. The inequality (4.10) remains valid for $T = s(k_0, \varepsilon)$. By the definition of $s(k_0, \varepsilon)$,

$$\left\| \sum_{t=k_0}^{T-1} \gamma_t (d_t + \xi_t) \right\| \ge \varepsilon.$$

By the convergence of (4.9), and the boundedness of $\{d_t\}$, a constant $C > 0$ exists such that for all sufficiently large k_0 and sufficiently small ε, we have

$$\sum_{t=k_0}^{T-1} \gamma_t \ge \varepsilon / C.$$

Using (4.2), by a similar argument as in the analysis of Condition B, we can choose $\varepsilon_1 > 0$ small enough that for all $k_0 \in \mathscr{K}$ large enough so that the entire expression in parentheses in (4.10) is smaller than $-\delta(1-\alpha\sqrt{1+\varkappa})/2$. Therefore, for all $\varepsilon \in (0, \varepsilon_1]$ and all sufficiently large $k_0 \in \mathscr{K}$

$$\|r_{s(k_0,\varepsilon)} - r^*\|^2 \le \|r_{k_0} - r^*\|^2 - \frac{\delta(1-\alpha\sqrt{1+\varkappa})\varepsilon}{2C}.$$

We fix $r^* = \mathrm{Proj}_{Y^*}(r_{k_0})$ on the right hand side, and obtain

$$W(r_{s(k_0,\varepsilon)}) \le \|r_{s(k_0,\varepsilon)} - r^*\|^2 \le W(r_{k_0}) - \frac{\delta(1-\alpha\sqrt{1+\varkappa})\varepsilon}{2C}.$$

Now, the limit with respect to $k_0 \to \infty$, $k_0 \in \mathscr{K}$, proves Condition C.

Condition D is satisfied trivially, because $W(r^*) \equiv 0$ for $r^* \in Y^*$. □

The only question remaining is the boundedness of the sequence $\{r_t\}$. It is a common issue in the analysis of stochastic approximation algorithms [42, §5.1]. In our case, no additional conditions and analysis are needed, because our Lyapunov function (4.5) is the squared distance to the optimal set. Therefore, a simple algorithmic modification: the projection on a bounded set Y intersecting with $\{r \in \mathbb{R}^m : \Phi r = v^*\}$, is sufficient to guarantee boundedness. The modified method (3.4) reads:

$$r_{t+1} = \mathrm{Proj}_Y\big(r_t - \gamma_t \varphi(i_t)\,\widetilde{d}_t\big), \quad t = 0, 1, 2, \ldots. \tag{4.13}$$

Now, $Y^* = \{r \in Y : \Phi r = v^*\}$ and we require that this set is nonempty. This modification does not affect our analysis in any meaningful way, because the projection is nonexpansive. In the proof of Theorem 3, we use the inequality

$$\left\|\mathrm{Proj}_Y(r' - \gamma U(r')) - \mathrm{Proj}_Y(r'' - \gamma U(r''))\right\|^2 \le \left\|(r' - \gamma U(r')) - (r'' - \gamma U(r''))\right\|^2$$

and proceed as before. In the proof of Theorem 7, we start from

$$\|r_{t+1} - r^*\|^2 = \left\|\mathrm{Proj}_Y\big(r_t - \gamma_t(U(r_t) + \Delta_t)\big) - r^*\right\|^2 \le \left\|r_t - \gamma_t(U(r_t) + \Delta_t) - r^*\right\|^2,$$

and then continue in the same way as before. We did not include projection into the method originally, because it obscures the presentation. In practice, we have not yet encountered any need for it.

4.2 Convergence Analysis of the Multi-Step Risk-Averse Methods of Temporal Differences

We now carry out a detailed analysis of the stochastic method (3.9)–(3.10).

Lemma 8. *For any array of uniformly bounded random variables* $\{A_{k,t}\}_{k\ge 0, t\ge 0}$

$$\lim_{T\to\infty} \frac{\sum_{k=0}^T \sum_{t=k}^T \gamma_t (\lambda\alpha)^{t-k} A_{k,t} - \sum_{k=0}^T \gamma_k \sum_{t=k}^\infty (\lambda\alpha)^{t-k} A_{k,t}}{\sum_{k=0}^T \gamma_k} = 0, \quad a.s.$$

Proof. Changing the order of summation twice, we obtain

$$\begin{aligned}
\sum_{k=0}^T \sum_{t=k+1}^T |\gamma_t - \gamma_k|(\lambda\alpha)^{t-k} &\le \sum_{k=0}^T \sum_{t=k+1}^T \sum_{\ell=k+1}^t |\gamma_\ell - \gamma_{\ell-1}|(\lambda\alpha)^{t-k} \\
&\le \frac{1}{1-\lambda\alpha} \sum_{k=0}^T \sum_{\ell=k+1}^T |\gamma_\ell - \gamma_{\ell-1}|(\lambda\alpha)^{\ell-k} = \frac{1}{1-\lambda\alpha} \sum_{\ell=1}^T |\gamma_\ell - \gamma_{\ell-1}| \sum_{k=0}^{\ell-1} (\lambda\alpha)^{\ell-k} \\
&\le \frac{\lambda\alpha}{(1-\lambda\alpha)^2} \sum_{\ell=1}^T |\gamma_\ell - \gamma_{\ell-1}|.
\end{aligned}$$

Therefore, with C being the uniform bound on $\|A_{k,t}\|$ and $\gamma_k^{\max} = \max_{t\ge k} \gamma_t$, we obtain

$$\begin{aligned}
&\left\| \sum_{k=0}^T \sum_{t=k}^T \gamma_t (\lambda\alpha)^{t-k} A_{k,t} - \sum_{k=0}^T \gamma_k \sum_{t=k}^\infty (\lambda\alpha)^{t-k} A_{k,t} \right\| \\
&\quad \le C \sum_{k=0}^T \sum_{t=k+1}^T |\gamma_t - \gamma_k|(\lambda\alpha)^{t-k} + C \sum_{k=0}^T \sum_{t=T+1}^\infty \gamma_t (\lambda\alpha)^{t-k} \\
&\quad \le \frac{C\lambda\alpha}{(1-\lambda\alpha)^2} \sum_{\ell=1}^T |\gamma_\ell - \gamma_{\ell-1}| + \frac{C\gamma_{T+1}^{\max}\lambda\alpha}{(1-\lambda\alpha)^2}.
\end{aligned}$$

Assumption 1(iv) and [65, Lem. A.3] imply the assertion. □

We need another auxiliary result, extending Lemma 6 to our case.

Lemma 9.

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma_t \left(\sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top - \frac{1}{1-\lambda\alpha} Q\bar{P} \right)}{\sum_{t=0}^{T} \gamma_t} = 0 \quad a.s., \tag{4.14}$$

and for any $\varepsilon > 0$*,*

$$\lim_{t_0\to\infty} \sup_{T\geq t_0} \frac{\sum_{t=t_0}^{T} \gamma_t \left(\sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top - \frac{1}{1-\lambda\alpha} Q\bar{P} \right)}{\max\left(\varepsilon, \sum_{t=t_0}^{T} \gamma_t\right)} = 0 \quad a.s.. \tag{4.15}$$

Proof. Consider the sums appearing in the numerator of (9):

$$\sum_{t=0}^{T} \gamma_t \sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top = \sum_{k=0}^{T} e_{i_k} \sum_{t=k}^{T} (\lambda\alpha)^{t-k} \gamma_t e_{i_t}^\top, \quad T = 1,2,\ldots.$$

In view of Lemma 8, it is sufficient to consider the sums

$$S_T = \sum_{k=0}^{T} \gamma_k e_{i_k} \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} e_{i_t}^\top, \quad T = 1,2,\ldots.$$

We transform the inner sum:

$$\begin{aligned}\sum_{t=k}^{\infty} ((\lambda\alpha))^{t-k} e_{i_t}^\top &= \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} \Big\{ \sum_{\ell=k+1}^{t} \left[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \right] + e_{i_k}^\top P^{t-k} \Big\} \\ &= \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} e_{i_k}^\top P^{t-k} + \sum_{\ell=k+1}^{\infty} \sum_{t=\ell}^{\infty} (\lambda\alpha)^{t-k} \left[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \right].\end{aligned}$$

We can thus write $S_T = S_T^{(1)} + S_T^{(2)}$, with

$$S_T^{(1)} = \sum_{k=0}^{T} \gamma_k e_{i_k} e_{i_k}^\top \sum_{t=k}^{\infty} (\lambda\alpha)^{t-k} P^{t-k} = \frac{1}{1-\lambda\alpha} \sum_{k=0}^{T} \gamma_k \operatorname{diag}(e_{i_k}) \bar{P}$$

and

$$\begin{aligned} S_T^{(2)} &= \sum_{k=0}^{T} \gamma_k e_{i_k} \sum_{\ell=k+1}^{\infty} \sum_{t=\ell}^{\infty} (\lambda\alpha)^{t-k} \left[e_{i_\ell}^\top P^{t-\ell} - e_{i_{\ell-1}}^\top P^{t-\ell+1} \right] \\ &= \frac{1}{1-\lambda\alpha} \sum_{k=0}^{T} \gamma_k e_{i_k} \sum_{\ell=k+1}^{\infty} (\lambda\alpha)^{\ell-k} \left[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P \right] \bar{P} \\ &= \frac{1}{1-\lambda\alpha} \sum_{\ell=1}^{\infty} \sum_{k=0}^{\min(T,\ell-1)} \gamma_k e_{i_k} (\lambda\alpha)^{\ell-k} \left[e_{i_\ell}^\top - e_{i_{\ell-1}}^\top P \right] \bar{P}. \end{aligned}$$

The second sum is a convergent martingale, because $\mathbb{E}\left[e_{i\ell}^\top \,\middle|\, \mathscr{F}_{\ell-1} \right] = e_{i_{\ell-1}}^\top P$. Therefore, it satisfies (4.14).

Applying Lemma 6 to $S_T^{(1)} - \frac{1}{1-\lambda\alpha} \sum_{k=0}^{T} \gamma_k \operatorname{diag}(q) \bar{P}$, we obtain both assertions. □

Now we can follow the arguments given in Section 4.1 and establish the convergence of the multi-step method.

Theorem 10. *Assume that $\alpha(1+\varkappa)<1$, the sequence $\{r_t\}$ is bounded with probability 1, and the random estimates $\widetilde{\sigma}_{i_t}(P_{i_t},\Phi r_t)$ satisfy* (3.2)*. Then, with probability 1, every accumulation point of the sequence $\{r_t\}$ generated by* (3.9)–(3.10) *is a solution of* (3.15)*.*

Proof. We represent the direction used in (3.10) at step t as

$$z_t\widetilde{d}_t=\frac{1}{1-\lambda\alpha}\overline{U}_t(r_t)+\Delta_t^{(1)}+\Delta_t^{(2)},$$

with the operator $\overline{U}_t(\cdot)$ defined in (3.13), and

$$\Delta_t^{(1)}=-\alpha z_t\xi_t,$$
$$\Delta_t^{(2)}=z_td_t-\frac{1}{1-\lambda\alpha}\overline{U}_t(r_t).$$

For any r^* solving (3.15), with $\bar{\gamma}_t=\gamma_t/(1-\lambda\alpha)$, we have

$$\big\|r_{t+1}-r^*\big\|^2=\big\|r_t-\bar{\gamma}_t\overline{U}_t(r_t)-r^*\big\|^2-2\bar{\gamma}_t\langle\Delta_t^{(1)}+\Delta_t^{(2)},r_t-\bar{\gamma}_t\overline{U}_t(r_t)-r^*\rangle+\bar{\gamma}_t^2\big\|\Delta_t^{(1)}+\Delta_t^{(2)}\big\|^2.$$

Our intention is to verify the conditions of Theorem 5 for almost all paths of the sequence $\{r_t\}$.

Condition A. The sequence $\{z_t\}$ is bounded by construction. Since the series (4.9) is a convergent martingale, we conclude that $\lim_{t\to\infty}\gamma_t z_t\widetilde{d}_t=0$.

Conditions B and C: We follow the proof of Theorem 7. The deterministic term involving $\overline{U}_t(r_t)$ can be estimated as in (3.19):

$$\big\|r_t-\bar{\gamma}_t\overline{U}_t(r_t)-r^*\big\|^2\le\big\|r_t-r^*\big\|^2-2\bar{\gamma}_t\big(1-\alpha(1+\varkappa)\big)\big\|\Phi(r_t-r^*)\big\|_q^2+C\bar{\gamma}_t^2.$$

Since $\{z_t\}$ and $\{r_t\}$ are bounded, Assumptions 1 and 2 imply that $\sum_{t=0}^{\infty}\bar{\gamma}_t\langle\Delta_t^{(1)},r_t-\bar{\gamma}_t\overline{U}_t(r_t)-r^*\rangle$ is a convergent martingale.

To analyze the second error term, $\Delta_t^{(2)}$, we observe that for a vector e_{i_k} having 1 at position i_k and zero otherwise, the formula (3.11) yields

$$\begin{aligned}z_td_t&=\sum_{k=0}^{t}(\lambda\alpha)^{t-k}\varphi(i_k)\big(\varphi^\top(i_t)r_t-c(i_t)-\alpha\sigma_{i_t}(P_{i_t},\Phi r_t)\big)\\&=\Phi^\top\Big(\sum_{k=0}^{t}(\lambda\alpha)^{t-k}e_{i_k}e_{i_t}^\top\Big)\big(\Phi r_t-c-\alpha\sigma(P,\Phi r_t)\big).\end{aligned}$$

Subtracting (3.13), we obtain

$$\Delta_t^{(2)} = \Phi^\top \Big(\sum_{k=0}^{t} (\lambda\alpha)^{t-k} e_{i_k} e_{i_t}^\top - \frac{1}{1-\lambda\alpha} Q\bar{P}_t \Big) \big[\Phi r_t - c - \alpha\sigma(P, \Phi r_t) \big].$$

By virtue of Lemma 9, for any $\varepsilon > 0$,

$$\lim_{T\to\infty} \frac{\sum_{t=0}^{T} \gamma_t \Delta_t^{(2)}}{\sum_{t=0}^{T} \gamma_t} = 0, \qquad \lim_{t_0\to\infty} \sup_{T\geq t_0} \frac{\sum_{t=t_0}^{T} \gamma_t \Delta_t^{(2)}}{\max\Big(\varepsilon, \sum_{t=t_0}^{T} \gamma_t\Big)} = 0 \quad \text{a.s..}$$

The remaining analysis is the same as in the proof of Theorem 7. We obtain an inequality corresponding to (4.10):

$$\begin{aligned} \|r_T - r^*\|^2 &\leq \|r_{k_0} - r^*\|^2 \\ &+ \left(-\delta(1-\alpha(1+\varkappa)) + \frac{\sum_{t=k_0}^{T-1} \gamma_t \langle \Delta_t^{(1)} + \Delta_t^{(2)}, r_t - \bar{\gamma}_t \overline{U}(r_t) - r^* \rangle}{\sum_{t=k_0}^{T-1} \bar{\gamma}_t} + C \frac{\sum_{t=k_0}^{T-1} \bar{\gamma}_t^2}{\sum_{t=k_0}^{T-1} \bar{\gamma}_t} \right) \sum_{t=k_0}^{T-1} \bar{\gamma}_t, \end{aligned}$$

with $\delta > 0$. This allows us to verify the conditions of Theorem 5 and prove our assertion following the last steps of the proof of Theorem 7 verbatim. □

It is worth mentioning that the convergence condition for the multi-step method: $\alpha(1+\varkappa) < 1$, is slightly stronger that the condition for the basic method: $\alpha\sqrt{1+\varkappa} < 1$.

Again, as in the case of the basic method, discussed in Section 4.1, the boundedness of the sequence $\{r_k\}$ is not an issue of concern, because it can be guaranteed by projection on a bounded set Y. The modified method has the following form:

$$r_{t+1} = \mathrm{Proj}_Y\big(r_t - \gamma_t z_t \widetilde{d}_t\big), \quad t = 0, 1, 2, \ldots. \tag{4.16}$$

We just need Y to have a nonempty intersection Y^* with the set of solutions of (3.15). Due to the nonexpansiveness of the projection operator, all our proofs remain unchanged with this modification, as discussed at the end of Section 4.1.

Chapter 5

Empirical Study

5.1 Sampling-Based Risk Estimation

We first discuss the issue of obtaining stochastic estimates $\widetilde{\sigma}_{i_t}(P_{i_t},\cdot)$ satisfying (3.2) and Assumption 2:

$$\mathbb{E}\big[\widetilde{\sigma}_{i_t}(P_{i_t},\Phi r_t)\big|\mathscr{F}_t\big]=\sigma_{i_t}(P_{i_t},\Phi r_t),\quad t=0,1,2,\ldots, \tag{5.1}$$

In the expected value case, where $\sigma_{i_t}(P_{i_t},\Phi r_t)=P_{i_t}\Phi r_t=\mathbb{E}\big[\varphi^\top(i_{t+1})r_t\,\big|\,\mathscr{F}_t\big]$, we could just use the approximation value at the next state observed, $\varphi^\top(i_{t+1})r_t$ as the stochastic estimate of the expected value function. However, due to the nonlinearity of a risk measure with respect to the probability measure P_{i_t}, such a straightforward approach is no longer possible.

Statistical estimation of measures of risk is a challenging problem, for which, so far, only solutions in special cases have been found [23]. To mitigate this problem, we propose to use a special class of transition risk mappings which are very convenient for statistical estimation. For a given transition risk mapping $\overline{\sigma}_i(P_i,v)$, we sample N conditionally independent transitions from the state i, resulting in states $j^1,\ldots,j^N$. This sample defines a random empirical distribution, $P_i^N=\frac{1}{N}\sum_{k=1}^N e_{j^k}$, where e_j is the jth unit vector in $\mathbb{R}^n$. Since the sample is finite, we can calculate the plug-in risk measure estimate,

$$\widetilde{\sigma}_i^N(P_i,v)=\overline{\sigma}_i(P_i^N,v), \tag{5.2}$$

by a closed-form expression. One can verify directly from the definition that the resulting *sample-based transition risk mapping*

$$\sigma_i^N(P_i,v)=\mathbb{E}\big[\overline{\sigma}_i(P_i^N,v)\big],$$

satisfies all conditions of a transition risk mapping given in Section 2, if $\overline{\sigma}_i(\cdot,\cdot)$ does. The expectation above is over all possible N-samples. Therefore, if we treat $\sigma_i^N(\cdot,\cdot)$ as the "true" risk

measure that we want to estimate, the plug-in formula (5.2) satisfies (3.2) and Assumption 2. In fact, for a broad class of measures of risk $\overline{\sigma}_i(P_i, v)$, we have a central limit result: $\overline{\sigma}_i(P_i^N, v)$ is convergent to $\overline{\sigma}_i(P_i, v)$ at the rate $1/\sqrt{N}$, and the error has an approximately normal distribution [23]. However, we do not rely on this result here, because we work with fixed N. In our experiments, the sample size $N = 4$ turned out to be sufficient, and even $N = 2$ would work well.

5.2 Post-Decision States

An alternative to the modeling approach detailed in Section 2 is to use post-decision states. As before, we consider a system with n states which belong to set $\mathscr{X}$, and each state has m features where $m << n$. At the pre-decision state i_t, we take decision $u_t \in \mathscr{U}(i_t)$, proceed to the post-decision state $\bar{i}_{t+1}$ and then we observe uncertainty W_{t+1}. The progression of the system is illustrated as below:

$$\dots i_{t-1} \xrightarrow{u_{t-1}} \bar{i}_t \xrightarrow{W_t} i_t \xrightarrow{u_t} \bar{i}_{t+1} \xrightarrow{W_{t+1}} i_{t+1} \dots, \quad t = 1, 2, \dots$$

Using the same linear approximate value function and feature matrix as in Section 2, we state the projected risk-averse dynamic programming equation based on post-decision states as

$$\varphi^\top(\bar{i}_t) r = L\Big(\sigma_{\bar{i}_t}\big(P_{\bar{i}_t, u_t},\ c(i_t, u_t) + \alpha \varphi^\top(\bar{i}_{t+1}) r\big)\Big),\ i_t, \bar{i}_{t+1} \in \mathscr{X}.$$

The dual representation (1.10) allows us to state the following equivalent equation

$$\varphi^\top(\bar{i}_t) r = L\Big(\big\langle M_{\bar{i}_t}(P_{\bar{i}_t}, \Phi r),\ c(i_t, u_t) + \alpha \varphi^\top(\bar{i}_{t+1}) r\big\rangle\Big),\ i_t, \bar{i}_{t+1} \in \mathscr{X}$$

When the process is modeled using the post decision states, the temporal difference updates are expressed as follows:

$$d_t = \varphi^\top(\bar{i}_t) r_t - \sigma_{\bar{i}_t}\big(P_{\bar{i}_t},\ c(i_t, u_t) + \alpha \varphi^\top(\bar{i}_{t+1}) r_t\big).$$

This approach is particularly useful when the action to take at i_t is determined by an optimization problem in which the risk measure of the approximate value of the next post decision state $\bar{i}_{t+1}$ is involved. It is far easier to evaluate the post decision state $\bar{i}_{t+1}$ compared to i_t, since in that case we can take the risk measure outside of the optimization problem instead. In the following section, we are dealing with a problem that exhibits such a feature, and we model it using post-decision states.

5.3 Numerical Results

We apply the risk-averse methods of temporal differences to a version of a transportation problem discussed in [55]. We have 200 vehicles that can be at $M = 50$ locations. Initially all vehicles are at the first location. At each time period t, a stochastic demand D_{ijt} for transportation from location i to location j occurs, $i, j = 1, \ldots, M$, $t = 1, 2, \ldots$. The demand in different time periods are independent and are observed according to the following distribution

$$D_{ijt} \sim \text{round}\big(\max\{0, N(0,8)\}\big), \quad i, j \in \mathscr{I}, i \neq j, \quad t = 1, 2, \ldots$$

The vehicles available at location i may be used to satisfy this demand. They may also be moved empty. The state x_t of the system at time t is the M-dimensional integer vector containing the numbers of vehicles at each location.

For simplicity, we assume that a vehicle can carry a unit demand, and the total demand at the location i at time t can be satisfied only if $x_{it} \geq \sum_{j=1}^{M} D_{ijt}$; otherwise, the demand may be only partially satisfied and the excess demand is lost. One can relocate the vehicles empty or loaded, and we denote the cost of moving a vehicle empty from location i to location j as c^e_{ij}. Since we stay in a cost minimization setting, we also denote the net negative profit of moving a vehicle loaded from location i to location j as c^ℓ_{ij}. Let u^e_{ijt} be the number of vehicles moved empty from location i to location j at time t and u^ℓ_{ijt} be the number of vehicles that are moved loaded. For simplicity, let us refer to the combination of u^e_t and u^ℓ_t as u_t and denote:

$$c^\top u_t = \sum_{i,j=1}^{M} \big(c^e_{ij} u^e_{ijt} + c^\ell_{ij} u^\ell_{ijt}\big).$$

In this problem, the control u_t is decided *after* the state x_t and the demand D_t are observed. The next state is a linear function of x_t and u_t:

$$x_{t+1} = x_t - A u_t,$$

where A can be written in an explicit way by counting the outgoing and incoming vehicles.

We denote by $\mathscr{U}(x_t, D_t)$ the set of decisions that can be taken at state x_t under demand D_t. Our approach allows us to evaluate a look-ahead policy defined by a simple linear programming problem:

$$u^\pi_t(x_t, D_t) = \operatorname*{argmin}_{u \in \mathscr{U}(x_t, D_t)} \Big\{ c^\top u + \alpha \pi^\top (x_t - Au) \Big\}. \tag{5.3}$$

Here, π is the vector of approximate next-state values fully defining the policy. In our case, the immediate cost $c^\top u_t$ depends on D_t, and thus the risk-averse policy evaluation equation (1.9) has the following form:

$$v^\pi(x) = \sigma\Big(P, c^\top u^\pi(x,D) + \alpha v^\pi\big(x - Au^\pi(x,D)\big)\Big),$$

with P denoting the distribution of the demand. Our objective is to evaluate the policy π and to improve it. As the size of the state space is enormous, we resort to linear approximations of form (1.4), using the state x as the feature vector: $\widetilde{v}(x_t) = x_t^\top r$. The approximate risk-averse dynamic programming equation (2.1) takes on the form:

$$r^\top x = \sigma\Big(P, c^\top u^\pi(x,D) + \alpha r^\top\big(x - Au^\pi(x,D)\big)\Big). \tag{5.4}$$

We omit the projection operator, because the feature space has full dimension. Thanks to that, the multi-step approximate risk-averse dynamic programming equation (3.15) coincides with (5.4), and all risk-averse methods with $\lambda \in [0,1]$ solve the same equation.

In fact, we can combine the learning and policy improvement in one process, known as the *optimistic approach*, in which we always use the current r_t as the vector π defining the policy. As the policy is already improved at every iteration of the trajectory, we are not in need of a policy iteration algorithm.

We tested the risk-averse and the risk-neutral TD(λ) methods under the same long simulated sequence of demand vectors. At every time t, we sampled $N = 4$ instances of the demand vectors, and for each instance, we computed the best decisions by (5.3), and the resulting states. Then we computed the empirical risk measure (5.2) of the approximate value of the next state, and we used it in the observed temporal difference calculation (3.3):

$$\widetilde{d}_t = r_t^\top x_t - \alpha\overline{\sigma}\Big(P^N, c^\top u^{r_t}(x_t,D) + \alpha r_t^\top\big(x_t - Au^{r_t}(x_t,D)\big)\Big).$$

We used the mean–semideviation risk measure [50] as $\overline{\sigma}(\cdot,\cdot)$, which can be calculated in closed form for an empirical distribution P^N with observed transition costs $w^{(1)},\dots,v^{(N)}$:

$$\overline{\sigma}(P^N, v) = \mu + \beta\frac{1}{N}\sum_{j=1}^{N}\max(0, w^{(j)} - \mu), \quad \mu = \frac{1}{N}\sum_{j=1}^{N} w^{(j)}, \quad \beta \in [0,1].$$

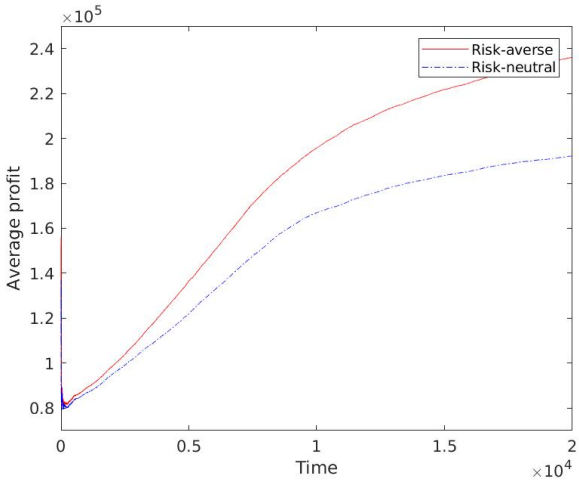
We used $\beta = 1$, $N = 4$, and $\alpha = 0.95$. We used stepsize $\gamma = 10^{-4}$ in the temporal difference updates. In the expected value model ($\beta = 0$), we also used $N = 4$ observations per stage, and

we averaged them, to make the comparison fair. The choice of $N = 4$ was due to the use of a four-core computer, on which the N transitions can be simulated and analyzed in parallel.

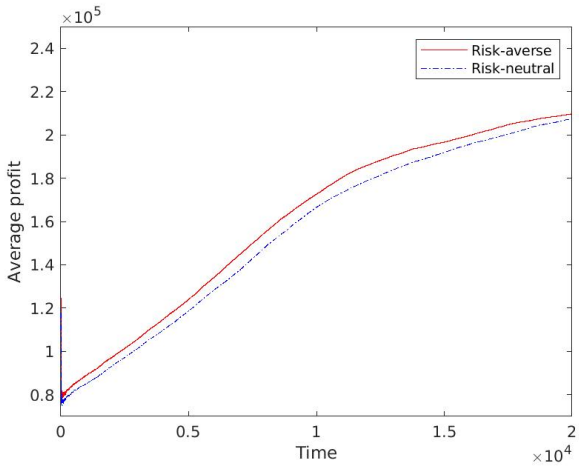
We compared the performance of the risk-averse and risk-neutral TD(λ) algorithms for $\lambda = 0$, 0.5, and 0.9, in terms of average profit per stage, on a trajectory with 20,000 decision stages. The results are depicted in Figure 5.1. We observe that the risk-averse algorithms outperform their risk-neutral counterparts in terms of the average profit in the long run. We also observe that the difference in performance is more significant when λ is closer to zero. It would appear that with risk-averse learning no additional advantage is gained by using $\lambda > 0$.

In addition to these results, we used 207 distinct trajectories, each with 200 decision stages, to compare the performance of the risk-averse and risk-neutral algorithms at the early training stages in terms of profit per stage. Figure 5.2 shows the empirical distribution function of the profit per stage of the risk-averse and risk-neutral algorithms at $t = 200$, for $\lambda = 0$, 0.5, and 0.9. The histograms of these findings can also be seen in Figure 5.3. The results demonstrate that in the early stages of learning ($t = 200$), the average profit of the risk-averse algorithm is more likely to be higher than that of the risk-neutral algorithm, and the difference is very pronounced for lower values of λ. The first order stochastic dominance relation between empirical distributions appears to exist.

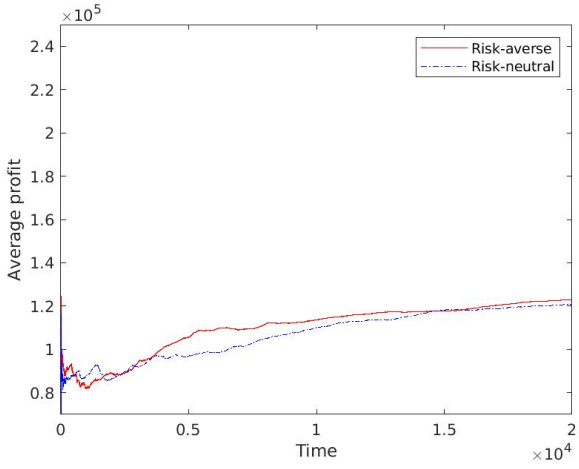
Although the risk-averse methods aim at optimizing the dynamic risk measure, rather than the expected value, they outperform the expected value model also in expectation. This may be due to the fact that the use of risk measures makes the method less sensitive to the imperfections of the value function approximation.



(a) $\lambda = 0$

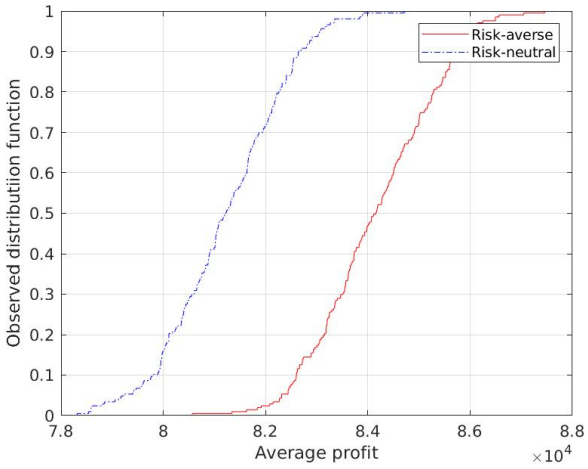


(b) $\lambda = 0.5$

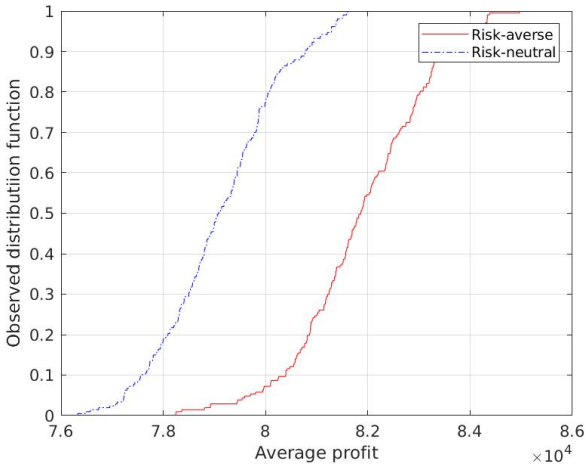


(c) $\lambda = 0.9$

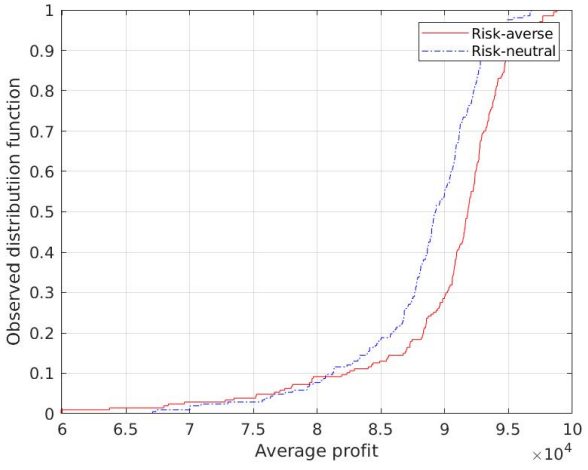
Figure 5.1: Evolution of the average profit per stage.



(a) $\lambda = 0$

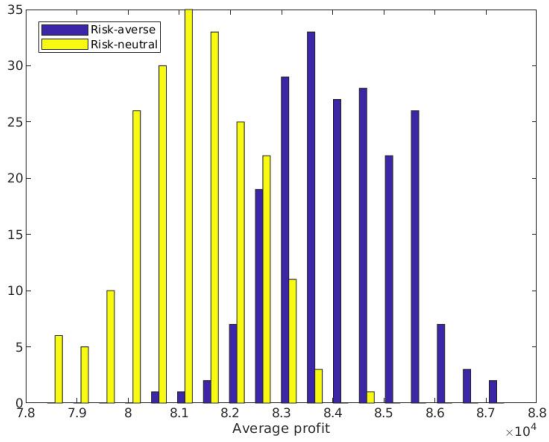


(b) $\lambda = 0.5$

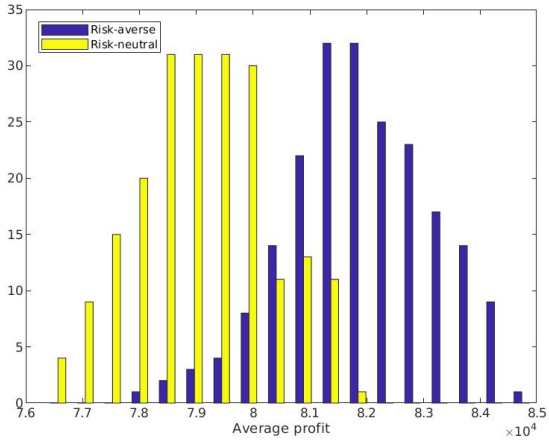


(c) $\lambda = 0.9$

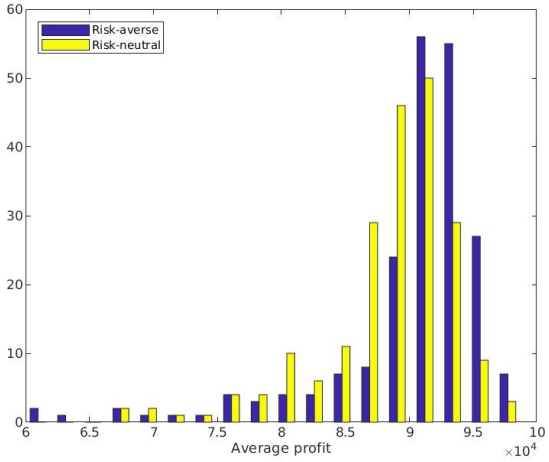
Figure 5.2: Empirical distribution of the average profit at $t = 200$.



(a) $\lambda = 0$



(b) $\lambda = 0.5$



(c) $\lambda = 0.9$

Figure 5.3: Histograms of the average profit at $t = 200$.

Chapter 6

Conclusion

The curse of dimensionality is inherent to most sequential decision making problems under uncertainty where risk-averse decision makers are involved. We state the risk-averse projected dynamic programming equation in which we rely on a linear value function approximation and employ transition risk mappings, which allows us to incorporate the theory of Markov risk measures [62] in an approximate dynamic programming scheme. We exploit the insight from the dual representation of coherent measures of risk to develop the concept of a distortion coefficient and we use it to establish that the risk-averse projected dynamic programming equation has a unique fixed point under mild conditions on the degree of risk-aversion.

In order to solve the risk-averse projected dynamic programming equation, we generalize the well-known TD(λ) algorithm [68] to the risk-averse case. We accomplish that by developing a new concept, namely risk-averse temporal difference, in which the observed approximate value of the next state on the trajectory is replaced by its empirical risk-measure. We analyze the risk-averse methods of temporal differences in two cases, where $\lambda = 0$ and where $\lambda \in (0, 1]$ and we establish the convergence of both algorithms. The multi-step method, where $\lambda \in (0, 1]$, requires slightly more strict conditions on the degree of risk-aversion.

In order to demonstrate the effectiveness of the risk-averse methods of temporal differences, we use it to learn the optimal allocation of vehicles in a transportation problem discussed in [55]. We run both the risk-averse methods of temporal differences and the well-known risk-neutral TD(λ) algorithm in tandem throughout very long trajectories of observed states. We observe that even when the algorithmic parameters don't strictly satisfy the convergence conditions, the risk-averse method outperforms the risk-neutral one for all selections of λ in the long run in terms of average profit. This difference is more significant when λ is lower. This is expected

because of the manner the data for the problem is generated. The data is highly variable throughout the decision stages and is time independent, thus, the advantages of the multi-step method aren't as pronounced compared to the basic method where $\lambda = 0$. Moreover, in order to compare the performance of the risk-averse and risk-neutral methods, we compare the distributions of average profit at the end of relatively short trajectories. We observe incontrovertible evidence indicating that the risk-averse methods stochastically dominate the risk-neutral ones in terms of accumulated profit.

References

[1] A. Arlotto, N. Gans, and J. M. Steele. Markov decision problems where means bound variances. *Operations Research*, 62(4):864–875, 2014.

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

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

[4] N. Bäuerle and U. Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2013.

[5] R. E. Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.

[6] R. E. Bellman, R. Kalaba, and B. Kotkin. Polynomial approximation – a new computational technique in dynamic programming. *Math. Comp.*, 17(8):155–161, 1963.

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

[8] T. Bielecki, D. Hernández-Hernández, and S. R. Pliska. Risk sensitive control of finite state Markov chains in discrete time, with applications to portfolio management. *Mathematical Methods of Operations Research*, 50(2):167–188, 1999.

[9] V. S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2):294–311, 2002.

[10] V.S. Borkar. A sensitivity formula for risk-sensitive cost and the actor-critic algorithm. *Systems & Control Letters*, 44(5):339 – 346, 2001.

[11] Ö. Çavuş and A. Ruszczyński. Computational methods for risk-averse undiscounted transient Markov models. *Operations Research*, 62(2):401–417, 2014.

[12] Ö. Çavuş and A. Ruszczyński. Risk-averse control of undiscounted transient Markov models. *SIAM Journal on Control and Optimization*, 52(6):3935–3966, 2014.

[13] Z. Chen, G. Li, and Y. Zhao. Time-consistent investment policies in Markovian markets: a case of mean-variance analysis. *J. Econom. Dynam. Control*, 40:293–316, 2014.

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

[15] P. Cheridito and M. Kupper. Composition of time-consistent dynamic monetary risk measures in discrete time. *International Journal of Theoretical and Applied Finance*, 14(01):137–162, 2011.

[16] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in neural information processing systems*, pages 3509–3517, 2014.

[17] K. J. Chung and M. J. Sobel. Discounted MDPs: distribution functions and exponential utility maximization. *SIAM*, 25:49–62, 1987.

[18] E. Cinlar. *An Introduction to Stochastic Processes*. Prentice-Hall, 1975.

[19] S. P. Coraluppi and S. I. Marcus. Risk-sensitive and minimax control of discrete-time, finite-state Markov decision processes. *Automatica*, 35(2):301–309, 1999.

[20] P. Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.

[21] P. Dayan and T. Sejnowski. TD(λ) converges with probability 1. *Machine Learning*, 14:295–301, 1994.

[22] E. V. Denardo and U. G. Rothblum. Optimal stopping, exponential utility, and linear programming. *Math. Programming*, 16(2):228–244, 1979.

[23] D. Dentcheva, S. Penev, and A. Ruszczyński. Statistical estimation of composite risk functionals and risk optimization problems. *Annals of the Institute of Statistical Mathematics*, 69(4):737–760, 2017.

[24] G. B. Di Masi and Ł. Stettner. Risk-sensitive control of discrete-time Markov processes with infinite horizon. *SIAM J. Control Optim.*, 38(1):61–78, 1999.

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

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

[27] J. Fan and A. Ruszczyński. Process-based risk measures and risk-averse control of discrete-time systems. *Mathematical Programming*, pages 1–28, 2018.

[28] J. Fan and A. Ruszczyński. Risk measurement and risk-averse control of partially observable discrete-time markov systems. *Mathematical Methods of Operations Research*, pages 1–24, 2018.

[29] B. G. Farley and W. A. Clark. Simulation of self-organizing systems by digital computer. *IRE Transactions on Information Theory*, 4:76–84, 1954.

[30] J. A. Filar, L. C. M. Kallenberg, and H.-M. Lee. Variance-penalized Markov decision processes. *Math. Oper. Res.*, 14(1):147–161, 1989.

[31] W. H. Fleming and S. J. Sheu. Optimal long term growth rate of expected utility of wealth. *The Annals of Applied Probability*, 9:871–903, 1999.

[32] H. Föllmer and I. Penner. Convex risk measures and the dynamics of their penalty functions. *Statistics & Decisions*, 24(1/2006):61–96, 2006.

[33] J.-B. Hiriart-Urruty. Mean value theorems in nonsmooth analysis. *Numerical Functional Analysis and Optimization*, 2(1):1–30, 1980.

[34] R. A. Howard. *Dynamic Programming and Markov Processes*. John Wiley & Sons, 1960.

[35] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Sci.*, 18:356–369, 1971/72.

[36] T. Jaakkola, M. I. Jordan, and S. P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.

[37] S. C. Jaquette. Markov decision processes with a new optimality criterion: discrete time. *Ann. Statist.*, 1:496–505, 1973.

[38] S. C. Jaquette. A utility criterion for Markov decision processes. *Management Sci.*, 23(1):43–49, 1975/76.

[39] A. Jaskiewicz, J. Matkowski, and A. S. Nowak. Persistently optimal policies in stochastic dynamic programming with generalized discounting. *Mathematics of Operations Research*, 38(1):108–121, 2013.

[40] A. Jobert and L. C. G. Rogers. Valuations and dynamic convex risk measures. *Mathematical Finance*, 18(1):1–22, 2008.

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

[42] H. Kushner and G. G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer, New York, 2003.

[43] S. Levitt and A. Ben-Israel. On modeling risk in Markov decision processes. In *Optimization and Related Topics (Ballarat/Melbourne, 1999)*, volume 47 of *Appl. Optim.*, pages 27–40. Kluwer Acad. Publ., Dordrecht, 2001.

[44] K. Lin and S. I. Marcus. Dynamic programming with non-convex risk-sensitive measures. In *American Control Conference (ACC), 2013*, pages 6778–6783. IEEE, 2013.

[45] W.-J. Ma, D. Dentcheva, and M. M. Zavlanos. Risk-averse sensor planning using distributed policy gradient. In *2017 American Control Conference (ACC)*, pages 4839–4844. IEEE, 2017.

[46] S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean-variance optimization in Markov decision processes. *European J. Oper. Res.*, 231(3):645–653, 2013.

[47] S. I. Marcus, E. Fernández-Gaucherand, D. Hernández-Hernández, S. Coraluppi, and P. Fard. Risk sensitive Markov decision processes. In *Systems and Control in the Twenty-First Century (St. Louis, MO, 1996)*, volume 22 of *Progr. Systems Control Theory*, pages 263–279. Birkhäuser, Boston, MA, 1997.

[48] M. L. Minsky. *Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain-Model Problem*. PhD thesis, Princeton University, 1954.

[49] E. A. Nurminski. Convergence conditions for nonlinear programming methods. *Kibernetika (Kiev)*, (6):79–81, 1972.

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

[51] J. Peng. *Efficient Dynamic Programming-Based Learning for Control*. PhD thesis, Northeastern University, 1993.

[52] J. Peng and R. J. Williams. Incremental multi-step Q-learning. *Proceedings of the Eleventh International Conference on Machine Learning*, pages 226–232, 1994.

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

[54] W. B. Powell. *Approximate Dynamic Programming - Solving the Curses of Dimensionality*. Wiley, 2011.

[55] W. B. Powell and H. Topaloglu. Approximate dynamic programming for large-scale resource allocation problems. In *Models, Methods, and Applications for Innovative Decision Making*, pages 123–147. INFORMS, 2006.

[56] L. A. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive reinforcement learning. *CoRR*, abs/1403.6530, 2014.

[57] M. L. Puterman. *Markov Decision Processes*. Wiley, 1994.

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

[59] B. Roorda, J. M. Schumacher, and J. Engwerda. Coherent acceptability measures in multiperiod models. *Mathematical Finance*, 15(4):589–612, 2005.

[60] G. A. Rummery. *Problem Solving with Reinforcement Learning*. PhD thesis, Cambridge University, 1995.

[61] G. A. Rummery and M. Niranjan. On-line Q-learning using connectionist systems. Technical report, Engineering Department, Cambridge University, 1994.

[62] A. Ruszczyński. Risk-averse dynamic programming for Markov decision processes. *Math. Program.*, 125(2, Ser. B):235–261, 2010.

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

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

[65] A. Ruszczyński and W. Syski. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12):1097–1105, 1983.

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

[67] Y. Shen, W. Stannat, and K. Obermayer. Risk-sensitive Markov control processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672, 2013.

[68] R. S. Sutton. Learning to predict by the method of temporal differences. *Machine Learning*, 3:9–44, 1988.

[69] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge: MIT press, 1998.

[70] A. Tamar, Y. Chow, M. Ghavamzadeh, and S. Mannor. Sequential decision making with coherent risk. *IEEE Transactions on Automatic Control*, 62(7):3323–3338, 2017.

[71] A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396, 2012.

[72] A. Tamar, S. Mannor, and H. Xu. Scaling up robust mdps using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.

[73] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16:185–202, 1994.

[74] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Fransactions on Automatic Control*, 42(5):674–690, 1997.

[75] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, 1989.

[76] C. J. C. H. Watkins and P. Dayan. Q - learning. *Machine Learning*, 8(3-4):279–292, 1992.

[77] L. L. Wegge. Mean value theorem for convex functions. *Journal of Mathematical Economics*, 1(2):207–208, 1974.

[78] D. J. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: a review. *J. Optim. Theory Appl.*, 56(1):1–29, 1988.