

**Robust, Risk-Sensitive, and Data-driven Control  
of Markov Decision Processes**

by

Yann Le Tallec

Submitted to the Sloan School of Management  
in partial fulfillment of the requirements for the degree of

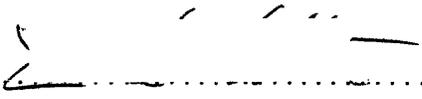
Doctor of Philosophy in Operations Research

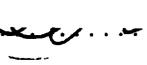
at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2007

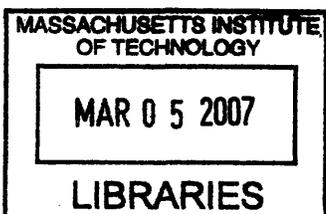
© Massachusetts Institute of Technology 2007. All rights reserved.

Author .....   
Sloan School of Management  
January 15, 2007

Certified by .....   
D. Simester  
Professor of Management Science  
Thesis Supervisor

Certified by .....   
J.N. Tsitsiklis  
Professor of Electrical Engineering and Computer Science  
Thesis Supervisor

Accepted by .....   
D. Bertsimas  
Co-director Operations Research Center,  
Professor of Operations Research



**ARCHIVES**



# Robust, Risk-Sensitive, and Data-driven Control of Markov Decision Processes

by

Yann Le Tallec

Submitted to the Sloan School of Management  
on January 15, 2007, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Operations Research

## Abstract

Markov Decision Processes (MDPs) model problems of sequential decision-making under uncertainty. They have been studied and applied extensively. Nonetheless, there are two major barriers that still hinder the applicability of MDPs to many more practical decision making problems:

- The decision maker is often lacking a reliable MDP model. Since the results obtained by dynamic programming are sensitive to the assumed MDP model, their relevance is challenged by model uncertainty.
- The structural and computational results of dynamic programming (which deals with expected performance) have been extended with only limited success to accommodate risk-sensitive decision makers.

In this thesis, we investigate two ways of dealing with uncertain MDPs and we develop a new connection between robust control of uncertain MDPs and risk-sensitive control of dynamical systems.

The first approach assumes a model of model uncertainty and formulates the control of uncertain MDPs as a problem of decision-making under (model) uncertainty. We establish that most formulations are at least NP-hard and thus suffer from the “curse of uncertainty.”

The worst-case control of MDPs with rectangular uncertainty sets is equivalent to a zero-sum game between the controller and nature. The structural and computational results for such games make this formulation appealing. By adding a penalty for unlikely parameters, we extend the formulation of worst-case control of uncertain MDPs and mitigate its conservativeness. We show a duality between the penalized worst-case control of uncertain MDPs with rectangular uncertainty and the minimization of a Markovian dynamically consistent convex risk measure of the sample cost. This notion of risk has desirable properties for multi-period decision making, including a new Markovian property that we introduce and motivate. This Markovian property is critical in establishing the equivalence between minimizing some risk

measure of the sample cost and solving a certain zero-sum Markov game between the decision maker and nature, and to tackling infinite-horizon problems.

An alternative approach to dealing with uncertain MDPs, which avoids the curse of uncertainty, is to exploit directly observational data. Specifically, we estimate the expected performance of any given policy (and its gradient with respect to certain policy parameters) from a training set comprising observed trajectories sampled under a known policy. We propose new value (and value gradient) estimators that are unbiased and have low training set to training set variance. We expect our approach to outperform competing approaches when there are few system observations compared to the underlying MDP size, as indicated by numerical experiments.

Thesis Supervisor: D. Simester

Title: Professor of Management Science

Thesis Supervisor: J.N. Tsitsiklis

Title: Professor of Electrical Engineering and Computer Science

## Acknowledgments

First, I would like to thank my thesis advisors John and Duncan who guided me to the completion of my dissertation. John and Duncan let me explore with great freedom some of my ideas, while providing complementary insights and questions. They were also instrumental in guaranteeing the convergence of my (random) research process. I am particularly grateful for John's continuous support and contribution to my research work.

I would like to thank Susan Murphy from the university of Michigan at Ann Arbor for her inspiration and advice for the chapter on data-driven approaches to MDPs. Also, I would like to thank my first advisor at MIT, Jeremie Gallien, for his support and his friendly guidance during my first two years.

I enjoyed MIT for its stimulating environment for learning and creativity, and for the exchange of ideas with an uncountable number of fellow students and remarkable faculty members. In particular, I would like to thank for their time and inspiration Profs. Dimitri Bertsekas, Daniela Pucci de Farias, Tommi Jaakkola, Pablo Parrilo, and Sanjoy Mitter.

The students of the Operations Research Center provided great companionship. At the ORC, I had memorable moments of engaging discussions on hazy topics and also many fun activities. I would like to particularly thank for their friendship Theo, Mike, Victor, Felipe, Guillaume, and the "inseparable" Katy, Margret and Juliane.

Volleyball introduced me to good friends in Cambridge, especially Al, Nikos, Jelena, Jeremy and Sonia, Matthias, and York. With them I could continue my favorite sport and foremost develop great friendships.

I was also very lucky to live with the company of good friends during my studies at MIT, thanks to my house-mates Balint, Leti, Taras, Matthieu, Marina, and Fabien.

An unexpected finding from my experience at MIT was to meet Tilke with whom I have had a fulfilling and harmonious relationship since we have met.

Last but not least, I would like to thank my family from France who I did not see often during my past five years at MIT. I wish I could have been closer to home

more frequently; nonetheless my family inhabited my mind throughout my PhD. I also missed my dear old friends, in particular Ben, Alex, Sam, Arnaud, Manu, Lulu, Benny, Philou, Sonia, Julien, Romain, Delphine, and my partners in adventure David, Diane, Arnaud, Damien, Lal, Francois, Severine. Despite the separation in time and space and too few get together, I am grateful for their continued friendship.

This research was partially supported by the National Science Foundation under grant DMI-0322823.

# Contents

- 1 Introduction** **15**
  
- 2 Computational complexity of control of uncertain Markov Decision Processes** **19**
  - 2.1 Introduction . . . . . 19
    - 2.1.1 Motivation . . . . . 19
    - 2.1.2 Contributions and literature review . . . . . 20
    - 2.1.3 Chapter structure . . . . . 21
    - 2.1.4 Computational complexity theory . . . . . 21
  - 2.2 Uncertain Markov Decision Processes . . . . . 22
    - 2.2.1 Markov Decision Processes . . . . . 22
    - 2.2.2 Different descriptions of parametric uncertainty . . . . . 26
    - 2.2.3 Connection with decision theory . . . . . 28
  - 2.3 History-dependent control of uncertain Markov decision processes . . 31
    - 2.3.1 Random uncertainty . . . . . 32
    - 2.3.2 Worst-case uncertainty . . . . . 35
    - 2.3.3 Worst-case regret . . . . . 36
  - 2.4 Stationary control of uncertain MDPs: the worst-case uncertainty . . 37
    - 2.4.1 General uncertainty . . . . . 38
    - 2.4.2 Only two possible MDP models . . . . . 39
    - 2.4.3 Rectangular uncertainty . . . . . 43
  - 2.5 Stationary control of uncertain MDPs: the case of random uncertainty 48
    - 2.5.1 General uncertainty . . . . . 49

2.5.2	Only two possible MDP models . . . . .	49
2.5.3	Rectangular uncertainty . . . . .	49
2.6	Stationary control of uncertain MDPs: the worst-case regret case . .	56
2.6.1	General uncertainty set . . . . .	56
2.6.2	Two possible MDP models . . . . .	57
2.7	Conclusion . . . . .	58
<b>3</b>	<b>Risk-averse and robust control of Markov Decision Processes</b>	<b>61</b>
3.1	Introduction . . . . .	61
3.1.1	MDP control background and motivations . . . . .	61
3.1.2	Literature review . . . . .	62
3.1.3	Chapter contributions . . . . .	65
3.1.4	Chapter structure . . . . .	66
3.2	Convex and coherent risk measures . . . . .	67
3.2.1	Definition of convex and coherent risk measures . . . . .	67
3.2.2	Examples of coherent and convex risk measures, and connection with expected utility . . . . .	69
3.2.3	Representation of convex risk measures . . . . .	70
3.3	Risk-averse control of dynamical systems . . . . .	77
3.3.1	Model description and notation . . . . .	78
3.3.2	Comments on our model choice . . . . .	83
3.3.3	Properties of multi-period risk . . . . .	85
3.3.4	A Markov game induced by a risk measure . . . . .	90
3.4	Risk minimization over a finite horizon . . . . .	93
3.5	Risk minimization over an infinite horizon . . . . .	100
3.5.1	Coherent risk measure of discounted sample cost . . . . .	100
3.5.2	Convex risk measure of undiscounted sample cost . . . . .	102
3.5.3	An illustration: minimization of exponential utility function .	106
3.6	From robust control to risk minimization . . . . .	110

3.6.1	A multi-period risk measure induced by uncertainty sets and penalty functions . . . . .	111
3.6.2	Connection with robust control formulation of MDPs . . . . .	117
3.6.3	From single-period risk measures to dynamically consistent Markovian multi-period risk measures . . . . .	120
3.7	Conclusion . . . . .	123
3.8	Glossary . . . . .	124
3.8.1	Definitions . . . . .	124
3.8.2	Notations . . . . .	125
<b>4</b>	<b>Data-driven approach to Markov Decision Processes</b>	<b>127</b>
4.1	Introduction . . . . .	127
4.1.1	Motivating examples . . . . .	128
4.1.2	Literature review . . . . .	132
4.1.3	Contributions . . . . .	135
4.1.4	Chapter structure . . . . .	136
4.2	Problem formulation . . . . .	136
4.2.1	Probabilistic model of the system . . . . .	137
4.2.2	Observed data . . . . .	138
4.2.3	Estimation problem . . . . .	145
4.3	Method of control variates, innovations of a random variable and geometry . . . . .	146
4.3.1	Method of control variates . . . . .	146
4.3.2	Innovations of a random variable . . . . .	148
4.3.3	Minimum variance estimator: the unconstrained case . . . . .	152
4.3.4	Minimum variance estimator: the constrained case . . . . .	154
4.4	Estimation of the value of a policy . . . . .	157
4.4.1	Unconstrained value estimator . . . . .	160
4.4.2	Constrained value estimator . . . . .	169
4.4.3	Algorithms . . . . .	171

4.4.4	Numerical experiments . . . . .	175
4.5	Estimation of the gradient of policy value . . . . .	183
4.5.1	Characterization of optimal gradient estimators . . . . .	185
4.5.2	Comparison with the baseline approach . . . . .	190
4.5.3	Comparison with actor-critic approaches . . . . .	193
4.6	Conclusion . . . . .	194
4.6.1	Key findings . . . . .	194
4.6.2	Discussion of the motivating examples . . . . .	195
4.6.3	Concluding remarks . . . . .	197
<b>5</b>	<b>Concluding remarks</b>	<b>199</b>

# List of Figures

2-1	Reduction of DECISION TREE to control of uncertain MDP . . . . .	34
2-2	Graphical representation of an uncertain MDP with optimal policy that is randomized . . . . .	55
2-3	Reduction of PARTITION to control of uncertain MDP . . . . .	58
3-1	Graphical representation of a toy dynamical system . . . . .	83
3-2	Graphical representation of a market MDP . . . . .	118
4-1	Graphical representation of a line MDP . . . . .	178
4-2	Estimation error of different estimators for a line MDP . . . . .	179
4-3	Estimation error of different estimators for a line MDP with little data	180
4-4	Estimation error of different estimators for an inventory problem . . .	182
4-5	Estimation error of different estimators for an inventory problem with poor approximation architecture . . . . .	183



# List of Tables

2.1	Complexity of the history-dependent control of uncertain MDPs . . .	32
2.2	Complexity of the worst-case control of uncertain MDPs . . . . .	37
2.3	Complexity of the average-case control of uncertain MDPs . . . . .	48
2.4	Complexity of the worst-case regret control of uncertain MDPs . . . .	56



# Chapter 1

## Introduction

Decision problems can be abstracted mathematically in terms of a set  $\mathcal{U}$  of options to choose from and a mapping that associates a performance vector to each option  $u \in \mathcal{U}$ . The performance vector captures all the evaluation criteria that the decision maker considers relevant. In general, the performance vector has multiple conflicting dimensions, making the choice of a decision subtle. For example, a manufacturer might have to trade-off the quality and the production cost of its products. In addition, when the performance of a decision depends on the realization of an uncertain exogenous factor  $\omega$ , the performance becomes a function of both the decision  $u \in \mathcal{U}$  and of the uncertain variable  $\omega$ . Notwithstanding, in most cases the decision maker can rank the attractiveness of all possible performance vectors. Specifically, we will assume that the decision maker has a “cost function”  $c$  (which could also be interpreted as the negative of its utility function) that maps each possible decision to a real number, with preferred decisions corresponding to lower cost. Given such a model for a decision problem, the identification of the best decision amounts to solving the minimization problem

$$\inf_{u \in \mathcal{U}} c(u).$$

At a high level, this thesis investigates how the aforementioned approach needs to be adapted when no model is available for the problem of interest, or when the model

is uncertain. Specifically, we will focus on problems of sequential decision making under uncertainty. These are decision problems with two salient characteristics:

- the decision maker interacts with the system over multiple “time periods,” making a new decision at each period,
- the decision maker’s decision at time  $t$  can depend on some observations of the system up to time  $t$ . In particular, it can depend on the realization of some uncertain variables (feedback control).

Since the actions of the decision maker are “feedback policies,” the decision space  $\mathcal{U}$  is a policy space, which can be huge, both in terms of cardinality and dimension. This thesis focuses on an important class of problems of sequential decision making under uncertainty, called Markov Decision Processes (MDPs - cf. Subsection 2.2.1 for a definition). An MDP model describes the state dynamics of a system and the associated cost over a time horizon. At each time period, the controller observes the state of the system and chooses an action. Conditional on this choice, the system moves to a new (possibly random) state and the controller incurs a cost. The objective of the controller is to minimize its expected cost.

In this thesis we investigate three main questions about the control of uncertain MDPs:

- (a) the computational complexity of different formulations for the optimal control of uncertain MDPs,
- (b) the decision-theoretic justification of the worst-case control of MDPs and the duality between risk-sensitive and robust control of MDPs, which allows us to define a broad class of tractable risk criteria for MDP control, and
- (c) the “efficient” estimation of the value of different policies from system trajectories.

In the next paragraphs, we outline the content of the three chapters dedicated to each of these questions, but we leave the literature review and the precise statement of contributions to the introduction section of each chapter.

In the next chapter, we study the computational complexity of different formulations for the optimal control of MDPs when the model is uncertain. When the controller has an MDP model for the system of interest (say, with a finite state-action space  $\mathcal{X}\mathcal{A}$  over a finite time horizon  $T$ ), it is well-known that a cost-minimizing policy can be chosen among the deterministic policies whose action at a given time depends only on the current system state. Hence, provided the model is known, the Markov property allows the controller to restrict, without loss of optimality, its attention to the space of “Markovian” deterministic policies, which has cardinality  $|\mathcal{A}|^{T \cdot |\mathcal{X}|}$ . In addition, dynamic programming methods allow us to find a minimum cost policy in this exponentially large decision space with only order  $O(T|\mathcal{X}|^2|\mathcal{A}|)$  arithmetic operations. In applications where the state-action space is very large, the computational complexity of solving an MDP to optimality becomes intractable — this limitation of the use of MDPs is known as the “curse of dimensionality.” In Chapter 2, we consider the case where the MDP model is not perfectly known and show that most formulations for the control of uncertain MDPs suffer from a worse complexity curse, namely the “curse of uncertainty.” Most formulations are NP-hard, and even PSPACE-hard in situations where learning takes place. We draw a comprehensive picture for the complexity of the control of uncertain MDPs by analyzing many different formulations. This picture is important for a modeler who can trade-off the complexity of a particular formulation with its fit to a particular application.

In Chapter 3, we make connections between risk-sensitive control of MDPs, zero-sum Markov games, and the worst-case formulation of uncertain MDPs. Specifically, we define the concept of a Markovian dynamically consistent convex risk measure of the uncertain outcome of a dynamic state space model (which is a notion of preference on uncertain trajectories, not on uncertain models, with desirable decision-theoretic properties). We prove that the minimization of such risk measures amounts to solving a zero-sum Markov game, which is also equivalent to optimizing an uncertain MDP for the worst-case model with “state-rectangular uncertainty set”. Our notion of risk not only justifies this common formulation for the control of uncertain MDPs from a decision theory perspective, but it also motivates a new worst-case formulation, where

each possible model comes with a penalty (for example capturing its unlikelihood) so that the worst-case model balances its penalty and its disadvantage to the controller. Thus, the penalized worst-case formulation to the control of uncertain MDPs yields less conservative policies for the decision maker than the worst-case approach. Furthermore, the penalized worst-case control of MDPs generates a broad class of tractable risk criteria for the risk-sensitive control of MDPs.

Finally, Chapter 4 of this thesis takes a data-driven approach to the control of MDPs. Instead of estimating a model and then evaluating or optimizing the estimated MDP model, we bypass model estimation to avoid the curse of uncertainty by exploiting system trajectories observed under a known sampling policy to estimate directly the performance of other policies. We derive estimators for the value and the value gradient that are guaranteed to be unbiased and that have lower variance than competing approaches from the literature, and we illustrate the advantages of our approach with numerical experiments.

# Chapter 2

## Computational complexity of control of uncertain Markov Decision Processes

### 2.1 Introduction

#### 2.1.1 Motivation

Markov decision processes (MDPs) are a versatile class of models for controlled discrete-time stochastic dynamical systems [11, 12]. They have been extensively studied in operations research and applied in many different domains such as operations management, network management, marketing, or robotics.

An optimal policy in a known MDP can be computed efficiently by dynamic programming techniques. For large-scale problems, approximate solution methods are available [14]. Still, the applicability of this approach is limited by the sensitivity of the optimal solution to parametric uncertainty. In particular, an optimal policy for a given set of model parameters might perform poorly on a system with slightly different parameter values, as illustrated recently in [86] and [54]. In this chapter, we are interested in the influence of parametric uncertainty on the expected policy performance in MDPs, but not so much in the internal randomness of the sample cost

along system trajectories, as in the literature on risk-sensitive control. Specifically, we study different formulations to find a policy that performs “consistently” well in an uncertain MDP, with an emphasis on their computational complexity.

### 2.1.2 Contributions and literature review

In this chapter, we study different formulations of the robust control of uncertain MDPs from a computational complexity perspective. The controller can optimize different objectives: the finite-horizon cost, the infinite-horizon discounted cost, or the expected average cost.

We assume that the uncertain parameters are constant so that the controller could learn about their value along the way and exploit this knowledge with history-dependent policies. Hence, the control problem can be related to partially observable Markov decision processes (POMDP) or Markov games with imperfect information. The complexity of POMDPs has been studied first in [58] and some refinements were obtained in [47]. Globally, these active learning problems are much harder (PSPACE-hard) than the optimal control of a given MDP with perfectly known parameters.

This huge complexity increase together with other arguments (cf. Subsection 2.2.3) makes the history-dependent control of uncertain MDPs often inappropriate. Therefore, most of this chapter deals with the simpler Markovian policies.

Two recent papers [54] and [38], building on the theory of zero-sum Markov games with perfect information, proposed a robust control formulation based on the worst-case formulation. This formulation provides a strong performance guarantee in face of parametric uncertainty. Furthermore, stationary robust policies can usually be computed at almost no extra cost compared to the nominal problem. However, this analysis requires a so-called rectangularity assumption on the uncertain parameter; otherwise it does not apply. We show that without this assumption the problem is NP-hard, even when there are only two possible values for the uncertain parameters.

Since the approach in [54, 38] can be very conservative, because the policy is tailored to the worst-case value of the parameters, we also investigate two alternative formulations, well-motivated by decision theory: one based on expected utility

and another based on the worst-case regret. We establish a comprehensive table of complexity results for the three formulations, with different assumptions on the system’s uncertain parameters and on the controller’s objective. Most of the resulting problems are “intractable,” making by contrast the worst-case formulation under state-rectangularity computationally attractive.

In the special case where the controller minimizes its average cost with respect to a (generalized) state-rectangular uncertainty over the set of Markovian policies, we establish modified Bellman’s equations provided that the system does not come back to state that it has left in the past (Proposition 2.5.3).

### 2.1.3 Chapter structure

We end this introduction with a note on complexity theory. The next section is devoted to the definition of uncertain MDPs. Section 2.3 studies the complexity of history-dependent control of uncertain MDPs for different decision models. The rest of the chapter is then focused on stationary control and is organized by the controller’s objective function: Section 2.4 deals with the worst-case performance, Section 2.5 studies the average performance for different parameter values, and Section 2.6 analyzes the worst-case regret objective.

### 2.1.4 Computational complexity theory

In this thesis, we assume that the reader is already familiar with computational complexity theory (cf. [26] for an introduction). Our model of computation is a Turing machine (However, we will only count arithmetic operations, as opposed to bit operations, in our size and running time estimates). The complexity classes we will need are those consisting of decision problems solvable in polynomial time (**P**), nondeterministic polynomial time (**NP**), and polynomial space (**PSPACE**) as a function of the instance size. Recall that  $\mathbf{P} \subset \mathbf{NP} \subset \mathbf{PSPACE}$ .

While we are ultimately interested in finding an optimal policy, we will focus on the question whether the minimal cost is below a given bound, in order to fit the

complexity framework of binary decision problems.

## 2.2 Uncertain Markov Decision Processes

### 2.2.1 Markov Decision Processes

#### System description and notations

Markov Decision Processes are controlled discrete-time stochastic dynamical systems [11]. In this chapter, which is focused on computational complexity issues, we will only consider an MDP with a finite state space  $\mathcal{X}$ ,  $|\mathcal{X}| = n$ , where in all states  $x \in \mathcal{X}$ , there is a finite action set  $\mathcal{A}(x)$  available to the controller,  $|\mathcal{A}(x)| \leq r$ . We will use the notations  $\mathcal{XA}$  for the state-action space and  $\Delta$  for the probability simplex over  $\mathcal{X}$ . Given that the system is in state  $x \in \mathcal{X}$  and that the controller chooses action  $a \in \mathcal{A}(x)$ , the system moves at the next stage to state  $y \in \mathcal{X}$ , with probability  $p(y; x, a)$ , and incurs an expected immediate cost  $c(x, a)$ . The notations  $x^\pi(t)$ ,  $a^\pi(t)$  and  $c^\pi(t)$  refer, respectively, to the realizations of the state of the system, the controller's action and the incurred cost at time  $t$ , under a given policy  $\pi$ . Denote by  $\nu_0$  the distribution over  $\mathcal{X}$  of the state at the beginning of the time horizon.

The parameters of an MDP are the expected immediate cost  $c(x, a)$  and the one-step transition probability vectors  $p(\cdot; x, a)$  for all state-action pairs  $(x, a) \in \mathcal{XA}$ . The family of transition probabilities vectors  $P = (p(\cdot; x, a))_{(x,a) \in \mathcal{XA}}$  is called the *transition kernel* of the MDP. Later, we will use the notation  $p(\cdot; x, \cdot) \in \Delta^{\mathcal{A}(x)}$  where  $x \in \mathcal{X}$  for the vector  $(p(\cdot; x, a))_{a \in \mathcal{A}(x)}$  of transition probabilities for all actions available in state  $x$ .

For finite-horizon problems, the specification of an MDP instance includes generally different cost parameters and transition probabilities for each time point, whereas we assume that these parameters are independent of time in infinite-horizon problems. The respective description lengths are then  $O(n^2rT)$  and  $O(n^2r)$ .

## Optimal control problem

The controller's interaction with the dynamical system is specified by a "feedback" policy. Let  $I_t$  be the information available to the controller at time  $t$ . It comprises at least the current state but could include the time index and the past trajectory. A randomized policy  $\pi$  of the controller is a mapping from the information space to the distributions over the action space such that  $\pi(I_t)$  gives non-zero probability only to actions available in the current state, i.e.,  $a^\pi(t) \in \mathcal{A}(x^\pi(t))$  almost surely. A deterministic policy is a randomized policy that picks an action as a deterministic function of  $I_t$ . The policy is Markovian (resp. stationary) if it depends only on the state and time index  $(x, t)$  (resp. the state  $x$ ). Denote by  $\Pi_{h,r}$ ,  $\Pi_{m,r}$  and  $\Pi_{s,r}$  the space of randomized history-dependent, Markovian and stationary policies, respectively. Similarly, let  $\Pi_{h,d}$ ,  $\Pi_{m,d}$ , and  $\Pi_{s,d}$  be the space of deterministic history-dependent, Markovian, and stationary policies.

The controller could minimize different objective functions over a given policy space. In this chapter, we will consider as objective

1. The finite-horizon cost,  $E \left[ \sum_{t=0}^T c^\pi(t) | \nu_0 \right]$ , where  $T \in \mathbb{N}$  is the horizon length.
2. The infinite-horizon cost:
  - (a) Discounted cost,  $E \left[ \sum_{t=0}^{\infty} \alpha^t c^\pi(t) | \nu_0 \right]$ , where  $\alpha \in [0, 1)$  is a constant discount factor.
  - (b) Expected average cost,  $\limsup_{T \rightarrow \infty} \frac{1}{T} E \left[ \sum_{t=0}^T c^\pi(t) | \nu_0 \right]$ .

## Solution of nominal MDPs

When the transition kernel and the expected immediate cost parameters of an MDP are known, the controller can use classical dynamic programming techniques to minimize the cost, over the set of history-dependent policies  $\Pi_{h,r}$ .

For finite horizon problems, an optimal policy in  $\Pi_{h,r}$  can be taken deterministic Markovian (and may depend on the time index). An optimal policy can be computed by the dynamic programming recursion in  $O(rn^2T)$  arithmetic operations, whereas

the instance size describing the parameters for each time is also  $O(rn^2T)$ . If the parameters are time-invariant, the finite horizon control problem can be described with only  $O(rn^2 + \log T)$  bits, but the evaluation of a stationary policy can still be done in polynomial time as shown in the next proposition.

**Proposition 2.2.1.** *Computing the value function of a time-invariant Markovian policy in a time-independent finite horizon problem can be done in  $O(n^3 \log T)$  time.*

*Proof.* For notational convenience, assume  $T = 2^K - 1$ . Let  $P \in \mathbb{R}^{n \times n}$  be the constant transition kernel under a stationary policy and  $c \in \mathbb{R}^n$  be the associated expected immediate cost. The value function of the policy is

$$V = \sum_{t=0}^{2^K-1} P^t c = \left( I + P^{2^{K-1}} \right) \left( \sum_{t=0}^{2^{K-1}-1} P^t \right) c = \prod_{k=0}^{K-1} (I + P^{2^k}) c.$$

Given  $P^{2^k}$  for  $k = 0, \dots, K-1$ , the value function can be computed in  $O(n^3 \log T)$ . The powers  $P^{2^k}$ ,  $k = 0, \dots, K-1$ , can also be computed in  $O(n^3 \log T)$  time.  $\square$

**Remark 2.2.2.** *When a finite-horizon MDP control problem is specified in terms of a time-invariant transition kernel, the optimal control problem is P-hard, and could be NP-hard (it is not known whether this is the case). For the finite-horizon discounted cost problem, Tseng gave in [97] an exact polynomial-time algorithm to compute the optimal expected cost of an MDP, provided that the same MDP over an infinite horizon has a unique optimal stationary policy. But the complexity of this algorithm is proportional to  $O(1/(1-\alpha))$  so that it is polynomial only when the discount factor  $\alpha$  is bounded away from one.*

For infinite horizon problems, the stationarity of the environment makes the complexity of the nominal MDP control relatively simpler, even though the instance description length is only  $O(rn^2)$ .

For discounted cost problems with a discount factor  $\alpha \in [0, 1)$ , the minimal expected cost starting from state  $x$ ,  $V^*(x)$ , over the set of history-dependent policies

satisfies the Bellman equations [12], i.e.,

$$V^*(x) = \min_{a \in \mathcal{A}(x)} \left[ c(x, a) + \alpha \sum_{y \in \mathcal{X}} p(y; x, a) V^*(y) \right], \quad x \in \mathcal{X}.$$

Without loss of optimality, the controller can pick a policy in the set of deterministic stationary Markovian policies  $\Pi_{s,d}$  that is greedy with respect to  $V^*$ . It is well-known that the optimal value  $V^*$  can be found by solving the following linear program with  $n$  variables and at most  $nr$  constraints, where  $c_i > 0$ ,  $i = 1, \dots, n$ , are arbitrary,

$$\begin{aligned} \max_{V \in \mathbb{R}^n} \quad & \sum_{i=1}^n c_i V(i) \\ \text{s.t.} \quad & V(i) \leq c(x, a) + \alpha \sum_{y \in \mathcal{X}} p(y; x, a) V(y), \quad (x, a) \in \mathcal{XA}. \end{aligned}$$

Since linear programs are solvable in polynomial time, the optimal value of discounted cost problem can also be computed in polynomial time.

For expected average cost control problems with finite state and action spaces, results similar to the ones for discounted cost problems hold under mild assumptions. For example, when all stationary policies are unichain (cf. p. 204 in [12]), the minimal average cost  $\lambda^*$  satisfies the Bellman equations,

$$\lambda^* + h^*(x) = \min_{a \in \mathcal{A}(x)} c(x, a) + \sum_{y \in \mathcal{X}} p(y; x, a) h^*(y), \quad x \in \mathcal{X}.$$

The controller can pick an optimal policy that is deterministic, stationary, and Markovian. Moreover, the optimal average cost  $\lambda^*$  is independent of the initial state distribution  $\nu_0$ , and can be found in polynomial-time as an optimal solution of the following linear program

$$\begin{aligned} \max_{\lambda, h \in \mathbb{R}^n} \quad & \lambda \\ \text{s.t.} \quad & \lambda + h(x) \leq c(x, a) + \sum_{y \in \mathcal{X}} p(y; x, a) h(y), \quad (x, a) \in \mathcal{XA}. \end{aligned}$$

In all cases, it is worthwhile observing that the standard dynamic programming technique provides polynomial-time algorithms to verify whether the cost of a policy is less than a given bound, for any fixed Markovian policy in finite horizon problems, and any fixed stationary Markovian policy for infinite-horizon problems.

## 2.2.2 Different descriptions of parametric uncertainty

In this chapter, we will focus on uncertainty about the transition kernel

$$P = (p(\cdot; x, a))_{(x,a) \in \mathcal{X}\mathcal{A}}$$

because uncertainty on cost parameters can be reduced to uncertainty on the transition kernel. Moreover, when the transition kernel is known, the problem under a worst-case criterion essentially reduces to a constrained MDP (cf. [2] and the references therein), which has been studied extensively and can be solved efficiently. In contrast, uncertainty on the transition kernel is harder to deal with.

We assume that the uncertain transition kernel  $P$  lies in a known *uncertainty set*  $\mathcal{P} \subset \Delta^{\mathcal{X}\mathcal{A}}$ . Since we are interested in computational complexity issues, we will assume that  $\mathcal{P}$  is a finite set. To avoid notational ambiguity, we will sometimes index  $\mathcal{P}$  by a model index set  $\mathcal{M}$ . For example, the transition probability vector from state-action pair  $(x, a)$  in the MDP model  $m \in \mathcal{M}$  will be denoted  $p_m(\cdot; x, a)$ . Each parameter value defines a nominal MDP model, which is often called scenario or environment in the sequel. By default, the instance description of an uncertain control problem includes as input the specification of  $|\mathcal{M}|$  nominal MDPs.

In some contexts, we will endow the uncertainty set  $\mathcal{P}$  with a probability measure  $q = (q_m)_{m \in \mathcal{M}}$ . The expected cost of a policy (where the expectation is taken with respect to the stochastic trajectory realizations for a given MDP model  $P$ ) becomes a random variable with respect to the uncertain parameter  $P \in \mathcal{P}$ .

Observe that the uncertainty set  $\mathcal{P}$  or the distribution of the uncertain parameter  $q$  can encode some dependencies between the uncertain parameters at different state-action pairs. Indeed, for a set  $\mathcal{P} \subset \Delta^{\mathcal{X}\mathcal{A}}$ , define the sets  $\mathcal{P}_x = \{p(\cdot; x, \cdot), p \in \mathcal{P}\}$  for

all states  $x \in \mathcal{X}$  and  $\mathcal{P}_{(x,a)} = \{p(\cdot; x, a), p \in \mathcal{P}\}$  for all state-action pairs  $(x, a) \in \mathcal{XA}$ . Clearly,  $\mathcal{P} \subset \prod_{x \in \mathcal{X}} \mathcal{P}_x \subset \prod_{(x,a) \in \mathcal{XA}} \mathcal{P}_{(x,a)}$ . But these inclusions can all be proper with the elements in the difference “ruled out by dependencies” across states or state-action pairs.

A *state-rectangular uncertainty* imposes some independence across states of the uncertain parameters. It plays an important role when it induces a principle of optimality that decomposes the problem state by state (cf. Subsection 2.4.3). When nature picks the worst possible parameter  $P \in \mathcal{P}$ , the rectangularity assumption has been made explicit in [38] and [54].

**Definition 2.2.3.** *When the time horizon is infinite, the uncertainty set  $\mathcal{P}$  is state-rectangular if there are sets  $\mathcal{P}_x \subset \Delta^{\mathcal{A}(x)}$ ,  $x \in \mathcal{X}$ , such that  $\mathcal{P} = \prod_{x \in \mathcal{X}} \mathcal{P}_x$ ; equivalently,*

$$\mathcal{P} = \{P \in \Delta^{\mathcal{XA}} \mid P = (p(\cdot; x, \cdot))_{x \in \mathcal{X}} \text{ and } p(\cdot; x, \cdot) \in \mathcal{P}_x, \forall x \in \mathcal{X}\}.$$

*Similarly, the uncertainty set  $\mathcal{P}$  is state-action rectangular if there are sets  $\mathcal{P}_{(x,a)} \subset \Delta$ , for all  $(x, a) \in \mathcal{XA}$ , such that  $\mathcal{P} = \prod_{(x,a) \in \mathcal{XA}} \mathcal{P}_{(x,a)}$ ; equivalently,*

$$\mathcal{P} = \{P \in \Delta^{\mathcal{XA}} \mid P = (p(\cdot; x, a))_{(x,a) \in \mathcal{XA}} \text{ and } p(\cdot; x, a) \in \mathcal{P}_{(x,a)}, \forall (x, a) \in \mathcal{XA}\}.$$

*When the time horizon  $T$  is finite, the uncertainty set  $\mathcal{P}$  is state-rectangular if there are sets  $\mathcal{P}_{t,x} \subset \Delta^{\mathcal{A}(x)}$ ,  $x \in \mathcal{X}$ ,  $t = 1, \dots, T$ , such that  $\mathcal{P} = \prod_{t,x \in \mathcal{X}} \mathcal{P}_{t,x}$ ; equivalently,*

$$\mathcal{P} = \{P \in \Delta^{T\mathcal{XA}} \mid P = (p(\cdot; t, x, \cdot))_{x \in \mathcal{X}} \text{ and } p(\cdot; t, x, \cdot) \in \mathcal{P}_{t,x}, \forall x \in \mathcal{X}, t = 1, \dots, T\}.$$

*Similarly, the uncertainty set  $\mathcal{P}$  is state-action rectangular if there are sets  $\mathcal{P}_{(t,x,a)} \subset \Delta$ , for  $t = 1, \dots, T$  and all  $(x, a) \in \mathcal{XA}$ , such that  $\mathcal{P} = \prod_{t,x,a} \mathcal{P}_{(t,x,a)}$ .*

*In particular, with this convention, state-rectangularity of the uncertainty set  $\mathcal{P}$  of a finite-horizon problem implies that the set  $\mathcal{P}$  factors along the time dimension — a property that we will refer to as time-rectangularity.*

When the uncertainty set  $\mathcal{P}$  is endowed with a probability measure  $q$ , the concept

of rectangularity can be extended so that the observation of an uncertain parameter at a state is not informative on its value at other states.

**Definition 2.2.4.**

- a) A random uncertainty  $q$  is state-rectangular if it has a state-rectangular support  $\mathcal{P} \subset \Delta^{\mathcal{X}\mathcal{A}}$  and if there are probability distributions  $q_x$  on  $\mathcal{P}_x$  for all  $x \in \mathcal{X}$  such that for all  $P \in \mathcal{P}$ ,

$$q(P) = \prod_{x \in \mathcal{X}} q_x(p(\cdot; x, \cdot)).$$

- b) Similarly, a random uncertainty  $q$  is state-action-rectangular if its support  $\mathcal{P}$  is state-action-rectangular and if there exist probability distributions  $q_{(x,a)}$  on  $\mathcal{P}_{(x,a)}$  for all state-action pairs  $(x, a) \in \mathcal{X}\mathcal{A}$  such that for all  $P \in \mathcal{P}$ ,

$$q(P) = \prod_{(x,a) \in \mathcal{X}\mathcal{A}} q_{(x,a)}(p(\cdot; x, a)).$$

**Remark 2.2.5.** When we assume that the uncertainty is rectangular, we will also assume that the problem instance is given in factored form. For example, if the uncertainty is state-rectangular,  $\mathcal{P}$  (resp.  $q$ ) is described by  $\mathcal{P}_x$ ,  $x \in \mathcal{X}$  (resp.  $q_x$ ,  $x \in \mathcal{X}$ ). Thus, the number of uncertain scenarios in  $\mathcal{P}$  is exponential in the instance description length.

### 2.2.3 Connection with decision theory

We will focus on three formulations for the control of uncertain MDPs. They address different needs, and correspond to different approaches in decision theory.

#### Random uncertainty

When the uncertain parameter is endowed with a probability distribution  $q$ , a natural decision-theoretic framework is the one of expected utility as laid out by Von Neumann and Morgenstern [53].

Recall that there are two levels of randomness. For a fixed MDP model  $m \in \mathcal{M}$ , the cost along a realized trajectory of the associated MDP under policy  $\pi$  is a random variable with expectation denoted by  $C_m(\pi)$ . Utility-based decision theory says that a controller with utility function  $u$  and subjective probability measure  $q$  on the uncertain MDP model would choose a policy  $\pi$  that maximizes his expected utility. In this chapter, we restrict our attention to a risk-neutral decision-maker, that is, with a utility function  $u(x) = x$ . Nonetheless, this work serves as a basis to analyze other utility functions, in particular risk-averse criteria. Formally, the decision maker solves

$$\sup_{\pi} \sum_{m \in \mathcal{M}} q_m C_m(\pi).$$

Furthermore, observe that a decision-maker might be essentially risk-neutral with respect to the trajectory realization but risk-averse with respect to the uncertain parameters. For example, if the controller is interacting with a large population of identical but independent systems, the randomness of the trajectory averages out but the parametric uncertainty does not.

### **Worst-case uncertainty**

In the tradition of robust continuous-time control and more recently robust optimization [92], decision robustness is interpreted as a performance guarantee for the worst-case parameter in the uncertainty set. In that case, the controller plays a zero-sum game with nature. As in most of the formulations considered in the literature, nature observes the controller's policy when choosing the uncertain parameter. Thus, the decision maker solves

$$\inf_{\pi} \max_{m \in \mathcal{M}} C_m(\pi). \tag{2.2.1}$$

The worst-case formulation (2.2.1) relates to risk-averse decision theory through the recent, yet popular, notion of coherent risk measure [3]. In this framework the controller seeks to minimize his risk over his available positions. Here, if the risk is defined as the worst-case expectation over  $\mathcal{M}$ , which is a coherent risk measure,

then (2.2.1) corresponds to a risk minimization problem. The notion of coherent risk measures is motivated axiomatically and generalizes to some extent decision theory based on expected utility. The reader should refer to Chapter 3 for more details.

### Maximum regret

The last robust formulation that we consider is based on the maximum regret decision theory introduced by Savage [79] as an alternative to the utility-based theory. It also offers a principled way to mitigate the potential conservativeness of the worst-case approach mentioned above. Indeed, the worst-case formulation will only consider the worst possible scenario and possibly behave very poorly in all other, more favorable, cases, whereas the maximum regret formulation adjusts for the potential of each scenario.

In our setting, the regret of a given policy in a certain environment is the difference between the expected cost of the policy in that environment and the minimal cost achievable in that environment if the controller knew which environment it was facing. Let  $C_m^*$  be the minimal cost for the MDP model  $m \in \mathcal{M}$ . Minimizing the worst-case regret amounts to solving

$$\inf_{\pi} \max_{m \in \mathcal{M}} (C_m(\pi) - C_m^*).$$

As mentioned earlier, the advantage of this formulation compared to the worst-case approach (2.2.1) is that it takes into account the intrinsic potential of a given environment  $m$  through  $C_m^*$ . As a result, the attention of the decision maker is not restricted to the worst environment, but spans all the possible scenarios.

### Why restrict the policy space?

History-dependent policies are studied (briefly) in Section 2.3, but the core of this chapter is devoted to Markovian control of uncertain MDPs. When the time horizon is infinite, we will further restrict our attention to stationary policies. There are several motivations for giving up the opportunity to learn the unknown parameters with history-dependent policies.

1. A history-dependent policy can be hard to compute and potentially impossible to implement. The former point is formalized in Section 2.3, while the latter relates to possible practical limitations or interests of the decision-maker. For example, the past information available to the controller may be limited. In a dynamic pricing problem, it might be hard or costly to adjust the prices, whereas optimal history-dependent policies tend to take relatively large and irregular actions (cf. p. 478 in [5]) .
2. The value of active learning is limited by the time needed to exploit new findings or by the cost of gathering more information given a possibly extensive preexisting knowledge of the system. In these cases, stationary policies can perform almost as well as history-dependent policies and are simpler.

The restriction of the policy space to deterministic policies can also be justified in many contexts. For example, in social sciences where the controller interacts with people, deterministic policies can be better accepted since they are more predictable, transparent, and fair.

## 2.3 History-dependent control of uncertain Markov decision processes

In this section, the controller picks a policy in the set of history-dependent policies, which is a very large decision set. For simplicity of exposition, we present the finite-horizon case here, but most of the results presented in this section generalize to the infinite horizon discounted and average cost problems.

Since the time horizon is finite, nature can in general choose the uncertain parameters independently for each time step, but her choice is fixed at the beginning of the time horizon. Her choice is either random or adversarial, knowing the controller's policy. In both cases, the controller has the opportunity to learn about nature's choice and exploit new information by using a history-dependent policy.

The results of this section are summarized in Table 2.1. For conciseness, the proofs are presented only for deterministic policies, but they extend readily to randomized policies.

Besides, we do not study the case where there are a fixed number of uncertain parameters in this section.

Uncertainty	Random uncertainty	Worst-case	Maximum regret
General case	PSPACE-hard	PSPACE-hard	PSPACE-hard
State-rectangularity	NP-hard	zero-sum Markov games	?

Table 2.1: Summary of the complexity of finite-horizon history-dependent control of uncertain MDPs

### 2.3.1 Random uncertainty

Here we assume that nature picks a scenario in a finite set  $\mathcal{M}$  according to a distribution  $q$  known to the controller. However, the controller does not observe the chosen scenario. A risk-neutral controller wants to solve

$$\min_{\pi \in \Pi_{h,d}} \sum_{m \in \mathcal{M}} q_m C_m(\pi). \quad (2.3.1)$$

#### The case of general random uncertainty

When no assumption is made on the uncertainty distribution  $q$ , the active learning problem (2.3.1) is a particular type of partially observable Markov decision process (POMDP). The computational complexity of POMDP problems has been studied first in [58] and then refined in [47], which delineates how the difficulty to represent history-dependent policies affects the complexity of solving a POMDP. The analysis of the former paper applies to our special type of POMDPs.

For our special type of POMDPs and a uniform distribution  $q$  over the set of scenarios  $\mathcal{M}$  (i.e.,  $q_m = 1/|\mathcal{M}|$ ), Papadimitriou and Tsitsiklis show in Theorem 6 of [58] that the PSPACE-complete quantified satisfiability problem (QSAT) can be

reduced to the finite horizon control problem. In addition, when the time horizon  $T$  satisfies  $T \leq |\mathcal{X}|$ , the problem is in PSPACE.

QSAT amounts to checking whether an expression  $\exists x_1 \forall x_2 \exists \dots F(x_1, \dots, x_n)$  is true, where  $F$  is a Boolean expression in conjunctive normal form with three literals per clause. The reduction used in [58] builds a scenario for each logical clause of  $F$  such that a history-dependent policy is mapped to a quantified logical assignment. A clause is satisfied if and only if the policy achieves zero expected cost in the corresponding scenario. Otherwise, the controller pays a cost of 1. Consequently,  $F$  is a “yes” instance of QSAT if and only if there is a history-dependent policy achieving zero expected cost.

It is clear that the problem remains PSPACE-hard even for the more general case of a non-uniform distribution  $q$ .

### Rectangular random uncertainty

Even when the probability distribution  $q$  of the uncertain scenario is state-rectangular, the decision problem associated with (2.3.1) is not tractable. This result should be contrasted with the case of worst-case control studied in Subsection 2.3.2.

First, observe that, even when the probability distribution  $q$  is state-rectangular, the problem (2.3.1) does not reduce to a standard MDP. Indeed, when the system comes back to a state  $x$  with uncertain parameter  $p(\cdot; x, a)$ , the uncertain parameter is required to be the same since the uncertain parameters are sampled once at the beginning of the time horizon. In particular, the uncertainty is not time-rectangular.

**Theorem 2.3.1.** *The risk-neutral history-dependent (deterministic or not) finite-horizon Markovian control of multiple MDPs under the state-rectangularity assumption (cf. Definition 2.2.4), even with four states, is NP-hard.*

*Proof.* We reduce the NP-complete problem DECISION TREE in [26] (p. 282) to the control problem (2.3.1).

INSTANCE: Finite set  $\mathcal{M}$  of hypotheses, collection  $\mathcal{T} = \{T_1, \dots, T_r\}$  of binary tests, that is  $T_i : \mathcal{M} \mapsto \{0, 1\}$ , positive integer  $K$ .

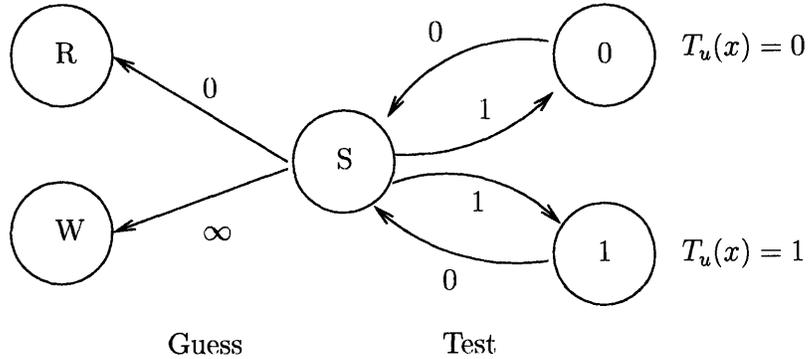


Figure 2-1: The reduction of DECISION TREE to control of uncertain MDP by history-dependent policies.

QUESTION: Is there a decision tree for  $\mathcal{M}$  using the tests in  $\mathcal{T}$  that has total external path length  $K$  or less? The total external path length of a decision tree is the sum over all leaves of the number of edges on the path from the root to the leaf. Hence, it is  $|\mathcal{M}|$  times the expected number of questions from  $\mathcal{T}$  to identify with certainty a random, uniformly chosen element of  $\mathcal{M}$ , according to a given decision tree.

Given an instance of DECISION TREE, we build a four-state uncertain MDP with  $|\mathcal{M}|$  equiprobable scenarios (hypotheses) defined as follows. Let  $m \in \mathcal{M}$  be the true scenario initially unknown to the controller. In the start node  $S$ , the controller can choose a test, say  $T_u$ ,  $u = 1, \dots, r$ , with a cost of 1, that leads him to state 0 if  $T_u(m) = 0$  and 1 otherwise, or make a guess  $y \in \mathcal{M}$  about the hypothesis he is facing. If the guess is right, he moves at no cost to an absorbing state  $R$ ; otherwise he pays a large cost  $C > |\mathcal{M}|r$  and goes to state 1. From both states 0 and 1, the system moves with certainty back to state  $S$  at no cost. The system is illustrated in Figure 2-1. Observe that the dynamics are uncertain only out of state  $S$ . Therefore, the uncertain parameter is state-rectangular, but not state-action-rectangular.

From the standard theory of POMDPs, deterministic history-dependent policies are optimal, so that the result for randomized policies follows from the one for deterministic policies. Hence, we assume that the controller can only pick a deterministic policy.

Furthermore, it is sufficient to consider a horizon of length  $T = 2r + 1$  since by that time, the controller would have had the opportunity to use all the tests and identify the right hypothesis. The expected cost of a deterministic policy, averaged over the  $|\mathcal{M}|$  possible scenarios, is less than or equal to  $r$  if the corresponding policy induces a decision tree over  $\mathcal{M}$ . When a deterministic policy corresponds to a decision tree, its expected cost is equal to  $1/|\mathcal{M}|$  times the total external path length of the associated decision tree. As a result, there is a decision tree with external path length of  $K$  or less if and only if there is a history-dependent deterministic policy for the uncertain MDP described above that achieves an expected cost of  $K/|\mathcal{M}|$  or less.  $\square$

### 2.3.2 Worst-case uncertainty

In this subsection, nature observes the controller's policy and picks the worst possible parameter in the uncertainty set  $\mathcal{M}$ . Formally, the controller solves

$$\inf_{\pi \in \Pi_{h,r}} \max_{m \in \mathcal{M}} C_m(\pi). \quad (2.3.2)$$

For general uncertainty set  $\mathcal{M}$ , checking whether the optimal value of (2.3.2) is below a threshold is PSPACE-hard. Indeed, the aforementioned argument of [58] generalizes straightforwardly to this problem, because the average of nonnegative costs is zero if and only if their maximum is zero.

However, when the uncertainty set  $\mathcal{M}$  is state-rectangular, Problem (2.3.2) reduces to a zero-sum sequential Markov game [82] (Recall that for finite-horizon problems, state-rectangularity implies time-rectangularity). These games have been extensively studied. For the finite-horizon and the infinite-horizon discounted cases, Bellman-Shapley equations hold, deterministic Markovian policies are optimal, and the value of the game can be computed by value iteration. As a result, the finite-horizon control problem can be solved in polynomial time when the problem instance comprises the description of the transition kernels at all times. The infinite-horizon discounted case can be efficiently approximated by value iteration. These results hold

even though the uncertainty set is exponentially large in the instance’s description length when it is state-rectangular (cf. Remark 2.2.5).

The consequences of the equivalence between worst-case control with state-rectangular uncertainty set and zero-sum Markov games will be studied in more detail in Subsection 2.4.3.

### 2.3.3 Worst-case regret

When the uncertainty set is general, the same argument as in the previous subsection shows that minimizing the worst-case regret of an uncertain policy is PSPACE-hard. Indeed, the minimal maximum regret achieved by a history-dependent policy in the uncertain MDP defined in [58] is less than or equal to zero if and only if the corresponding QSAT instance is satisfiable.

However, it is not clear whether rectangular uncertainty set makes the control problem any easier, as it was the case for the worst-case performance, because the connection with zero-sum Markov games does not hold when dealing with worst-case regret.

We finally note that the analysis in this section can be adapted to the control problem of minimizing expected regret, because it is equivalent to minimizing expected cost. When the random uncertainty is general, the problem is PSPACE-hard, and when the uncertainty set is state-rectangular, the problem is at least NP-hard.

In this section, we analyzed the complexity of history-dependent control of uncertain MDPs. When learning takes place, the control problems are intractable. These results are little surprising since the decision set  $\Pi_{h,r}$  is huge. As a result of this analysis and of the motivations reviewed at the end of Section 2.2 for using simpler policy space, we restrict the controller to pick Markovian policies in the rest of this chapter.

## 2.4 Stationary control of uncertain MDPs: the worst-case uncertainty

In contrast to the previous section on history-dependent control of uncertain MDPs, the controller now has to select a Markovian policy. Moreover, the chosen Markovian policy needs to be stationary when the time horizon is infinite. The controller wants to minimize its cost, but nature knows the policy and chooses the worst possible parameter in a given uncertainty set. Hence, for finite-horizon problems, the controller solves

$$\inf_{\pi \in \Pi_{m,r}} \sup_{m \in \mathcal{M}} C_m(\pi), \quad (2.4.1)$$

and for infinite-horizon problems, it solves

$$\inf_{\pi \in \Pi_{s,r}} \sup_{m \in \mathcal{M}} C_m(\pi). \quad (2.4.2)$$

The complexity of deciding whether the above infima are less than a threshold is summarized in the following table, whether the controller uses deterministic or randomized policies.

Uncertainty	Finite-horizon	Discounted cost	Average cost
General	NP-complete	NP-complete	NP-complete
2 MDPs	NP-complete	NP-complete	NP-complete
State-rectangular	Polynomial-time	Markov game	Markov game

Table 2.2: Summary of the complexity of Markovian control of uncertain MDPs in the worst-case scenario

When the uncertainty set is state-rectangular, this problem can be cast as a zero-sum Markov game with perfect information, as observed recently in [54] and [38]. Hence, the worst-case control of uncertain MDPs builds on a large literature on Markov games, starting from the seminal paper by Shapley [82]. However, the previous works on robust control of MDPs do not make clear the whole complexity picture. When the uncertainty set is not rectangular – for example when the uncertain parameters are collectively influenced by a possibly small number of causes – the worst-case

control problem is NP-complete.

### 2.4.1 General uncertainty

When the uncertainty set is general, Corollary 2 in [58] can be adapted to show that the worst-case control of uncertain MDPs is NP-complete. The reduction is analogous to the one used for the worst-case control with history-dependent policies.

Recall from the previous section that any instance of QSAT can be reduced to the worst-case control of an MDP with history-dependent policies on a finite-horizon. Intuitively, each clause was mapped to an MDP model. A policy  $\pi$  achieving zero expected cost on a given MDP model corresponds to a satisfying quantified assignment for which the dependencies of policy  $\pi$  on the history captures the dependencies of a given Boolean variable on other variables in the satisfying assignment of the QSAT instance. Here, we restrict the policy space to Markovian stationary policies which cannot capture the variable quantification of QSAT instances. Nonetheless, QSAT instances without universal quantifiers are instances of the satisfiability problem (SAT), which is NP-complete.

Hence, the reduction of QSAT also reduces SAT to the worst-case control of uncertain MDPs. To any SAT instance we can associate an uncertain MDP with one scenario per clause, on which stationary deterministic policies map to a SAT variable assignment. The expected cost of a stationary policy on a clause/scenario in this reduction is zero if and only if the corresponding assignment of the boolean variables makes this clause true. If the variable assignment associated with policy  $\pi$  does not make a SAT clause true, the cost of  $\pi$  on the MDP scenario corresponding to the unsatisfied clause is one. As a result, the minimal expected cost of a stationary policy on the considered uncertain MDP is zero if and only if the SAT instance is satisfiable.

This argument extends straightforwardly to randomized policies and the infinite-horizon setting so that the finite-horizon, infinite-horizon discounted, and average cost problems are all NP-complete under general uncertainty for both deterministic and randomized control.

## 2.4.2 Only two possible MDP models

The result of previous subsection can be strengthened to apply to the case of a fixed number of values for the uncertain parameter, even to the case of only two possible parameter values.

**Theorem 2.4.1.** *Even when there are only two possible values for the uncertain parameter, deciding whether there is a deterministic or randomized Markovian and stationary policy that achieves a worst-case infinite-horizon discounted or average cost less than a threshold for an uncertain MDP is NP-complete.*

*Under the same condition, deciding whether there is a deterministic or randomized Markovian policy that achieves a worst-case finite-horizon cost less than a threshold is NP-hard. When the instance's description comprises one transition kernel per stage, it is NP-complete.*

*Proof.* We will reduce the NP-complete problem PATH WITH FORBIDDEN PAIRS ([GT54], p. 203 in [26]) to the control problem of uncertain MDPs with only two scenarios. The problem statement for PATH WITH FORBIDDEN PAIRS is as follows. Let  $G = (V, A)$  be a directed acyclic graph with specified vertices  $s, t \in V$ . Let  $v = |V|$  and  $a = |A|$ . Let  $K = \{\{e_1, e'_1\}, \dots, \{e_n, e'_n\}\}$  be a list of pairs of vertices in  $V$ . The question is whether there exists a directed path from  $s$  to  $t$  in  $G$  that contains at most one vertex from each pair in  $K$  (such a path will be called an admissible path). This problem is NP-complete even when  $G$  has no nodes with in- or out-degree exceeding 2 and the pairs in  $K$  are disjoint.

By a polynomial time preprocessing of the graph  $G$ , we can delete all nodes in  $V$  that are not connected to the destination node  $t$  by a directed path. As a result, we will assume that all nodes in  $V$  are connected to node  $t$  by a directed path. We will also assume that there exists a directed path from  $s$  to  $e_i$  for  $i = 1, \dots, n$ .

By another polynomial-time preprocessing of the constraint set  $K$ , we can check whether there exists a directed path from  $e_i$  to  $e'_i$  or conversely. If there is none, we can delete the constraint  $\{e_i, e'_i\}$ . If there is a path, say from  $e_i$  to  $e'_i$ , there is no reverse path since  $G$  is acyclic. Hence, we can assume without loss of generality that

the pairs in  $K$  are ordered so that the prime refers to the second vertex along  $G$ .

From an instance of PATH WITH FORBIDDEN PAIRS, we will build an uncertain MDP whose cost will be zero if and only if there is an admissible path in the graph  $G$ .

First, let us show that the infinite-horizon stationary control of an uncertain MDP is NP-complete. Consider an uncertain MDP on the state space  $\mathcal{X} = V \cup \{e\}$ , initialized at state  $s$ . The first MDP model, “Normal”, is induced by the directed graph  $G$ . From each node  $i \in V$ , one can move to a child node of  $i$  at no cost. In addition, at the first node  $e_i$  of each constrained pair  $(e_i, e'_i) \in K$ , one can exit at zero-cost to the absorbing state  $e$ . In state  $e$ , the system loops and incurs a cost of 1 per stage. In the reduction for worst-case discounted cost control, the destination node  $t$  is zero-cost absorbing, while in the average cost problem, the system moves with probability one and zero cost to state  $s$ .

The second MDP model “Test” is similar to the Normal MDP, but differs in three ways. First, the system goes with zero cost and probability  $1/n$  to each of the first node  $e_i$  of the constrained pairs in  $K$ . Second, when the system reaches the second node  $e'_i$  of a constrained pair  $(e_i, e'_i) \in K$ , it moves with probability one to the first one,  $e_i$ , and incurs a cost 1. Third, the recurring cost in node  $e$  is zero.

If there is an admissible path from the source  $s$  to the destination  $t$ , the policy  $\pi$  that follows this path at the states on the path and exits to state  $e$  at the states out of the path incurs a worst-case cost of zero.

Now, assume that there is no admissible path from  $s$  to  $t$ . Let  $\mathcal{S}_\pi \subset V$  be the set of states that are visited with positive probability in the Normal MDP under the stationary Markovian policy  $\pi$ . A stationary policy  $\pi$  that has zero worst-case expected cost cannot use the action “exit” at any state in  $\mathcal{S}_\pi$ , because otherwise its cost in the Normal MDP would be positive. Since  $G$  is acyclic, the set  $\mathcal{S}_\pi$  contains a path from  $s$  to  $t$ . On the other hand, since we assumed that there is no admissible path from  $s$  to  $t$ , this path has to contain a forbidden pair, say  $(e_1, e'_1)$ . In the Test MDP, the system starts by moving to state  $e_1$  with probability  $1/n$ , will reach with positive probability  $e'_k$  for some  $k$  ( $k$  could be different from 1), and incur a cost of 1

when moving from  $e'_k$  to  $e_k$ . Hence, the expected cost of policy  $\pi$  is strictly positive in the Test MDP.

This concludes the proof that the worst-case (average or  $\alpha$ -discounted,  $0 \leq \alpha < 1$ ) cost of any stationary policy in our uncertain MDP is positive if there is no admissible path. On the other hand, it is possible to check in polynomial-time whether a given stationary policy has a zero worst-case cost. Hence, this problem is NP-complete.

We need to adapt the above reduction for infinite-horizon problems in order to deal with the worst-case control of uncertain MDP when the time horizon  $T$  is finite, because we allow in this case non-stationary policies.

Now, in the Normal and Test MDPs, when the system is at the source state  $s$  and a child node  $i$  is selected, it stays in state  $s$  with probability  $1/2$  and moves to  $i$  with probability  $1/2$ . Let the time horizon  $T$  be  $|V|$ .

Assume that there is no constrained path from  $s$  to  $t$ . Consider a Markovian policy  $\pi$  achieving zero cost in the Normal MDP. With positive probability, the system never loops during the time horizon in the Normal MDP. Since the policy  $\pi$  has zero expected cost in the Normal MDP, it reaches the destination node  $t$  through both nodes of a constrained pair, say  $(e_1, e'_1) \in \mathcal{S}_\pi$ . Let  $T_\pi \in [1, T] \cup \{+\infty\}$  be the random time when the system arrives in state  $e_1$  in the Normal MDP under policy  $\pi$ , where  $T_\pi = +\infty$  if  $e_1$  is not reached during  $[1, T]$ . Let  $t \in [1, T]$  be the shortest time such that  $T_\pi = t$  with positive probability.

In the Test MDP, there is a positive probability that the system loops  $t - 1$  times in node  $s$ , then moves to node  $e_1$ , and that afterwards the system never remains at the same state for two consecutive times. In such cases, the system is in a situation indistinguishable from the Normal MDP where the policy  $\pi$  does not use the action “exit”. As a result, the system will eventually visit a state  $e'_k$  before time  $T - 1$  and incur at least a cost of 1 when moving from state  $e'_k$  to  $e_k$  in the Test MDP.

Since one can check in polynomial time that a Markovian policy has zero cost on a finite horizon (provided the instance description comprises a transition kernel per stage), the worst-case control of uncertain MDPs over a finite time horizon is NP-complete.  $\square$

Even when the uncertainty is only on the cost parameters, the *deterministic* worst-case control of MDPs with only two parameter values is NP-complete.

**Theorem 2.4.2.** *Even when there are only two possible values for the uncertain cost parameter, deciding whether there is a deterministic Markovian policy that achieves a worst-case finite-horizon, infinite-horizon discounted or average cost less than a threshold for an uncertain MDP is NP-complete.*

*Proof.* We reduce the NP-complete problem SHORTEST WEIGHT-CONSTRAINED PATH ([ND30] in [26]) to the worst-case control problem.

INSTANCE: Directed graph  $G = (V, E)$ , length  $l(e) \in \mathbb{Z}^+$  and weight  $w(e) \in \mathbb{Z}^+$ , specified vertices  $s, t \in V$ , positive integers  $K, W$ .

QUESTION: Is there a simple path in  $G$  from  $s$  to  $t$  with total weight  $W$  or less and total length  $K$  or less?

Let us start with the finite horizon version and let the horizon  $T = |V|$ . Consider two MDPs with state space  $V$  and initialized in  $s$ . From any node  $v \in V$ , the controller can pick an edge  $e \in E$  starting from  $v$  and the system moves with probability one to the child node. The first MDP cost is the natural length  $l$  and the other is the weight  $w$ . When system reaches the zero-cost terminal node  $t$ , the controller incurs a cost of  $-K$  or  $-W$ . We have a YES instance of the short weight-constrained path problem if and only if the minimal worst-case cost of this finite horizon uncertain MDP is less than 0.

To deal with the average cost problem, we let the system go from  $t$  to  $s$  with probability one at no cost.

In both cases, one can check in polynomial time whether a stationary policy achieves a cost less than a fixed bound. Therefore, these uncertain MDP control problems are NP-complete.

This reduction can be adapted to uncertain  $\alpha$ -discounted cost problems. Consider a deterministic policy that defines a simple path from  $s$  to  $t$  but with a length exceeding  $K$ . Then its length is at least  $(K+1)$  since all edge lengths are integer-valued, and the corresponding discounted cost is at least  $\alpha^{|V|}(K+1)$ . Let us pick a discount factor

$\alpha$  close enough to one such that  $\alpha^{|V|}(K+1) > K > \alpha^{|V|}K$  and  $\alpha^{|V|}(L+1) > L > \alpha^{|V|}L$ . Then, there is a YES instance of the short weight-constrained path problem if and only if there is a deterministic policy with non-positive worst-case cost.  $\square$

This reduction using only cost uncertainty does not work with randomized policies; in fact the problem becomes the well studied problem of constrained MDPs (cf. [2] and references therein). Intuitively, when the dynamic structure of the MDP is certain and randomized policies are allowed, the set of achievable steady-state probabilities is described by a polyhedron, and multiple expected cost constraints can be handled efficiently by linear programming.

### 2.4.3 Rectangular uncertainty

When the uncertainty set is state-rectangular, the worst-case control problem reduces to a sequential zero-sum Markov game between the controller and nature. Hence, the worst-case control of MDPs is “tractable,” even though there are exponentially many uncertain scenarios.

A sequential zero-sum Markov game is a zero-sum game between two players, one trying to maximize an objective (the maximizer; here, nature) and the other trying to minimize it (the minimizer; here, the controller). In our case, the sequence of events in the game is as follows:

- At time  $t$ , the controller observes the system in state  $x \in \mathcal{X}$  and chooses an action  $a \in \mathcal{A}(x)$ .
- Nature observes  $(x, a) \in \mathcal{XA}$  and chooses a parameter  $p(\cdot; x, a)$  in the uncertainty set  $\mathcal{P}(x)$ , which determines the distribution of the new state  $y \in \mathcal{X}$  at time  $t + 1$  given  $(x, a)$ .
- The new state and an associated immediate cost are realized according to  $p(\cdot; x, a)$ .

This is equivalent to the situation where the controller chooses a non-stationary Markovian policy  $\pi$  and nature chooses a non-stationary policy  $\mu$  such that the prob-

ability  $\mu(p|x, a)$  of choosing parameter  $p$  in the state-action pair  $(x, a)$  depends only on the current state-action pair. The objective in our case is the expected cost of the policy  $\pi$  chosen by the controller, evaluated against the nature policy  $\mu$ .

For finite-horizon and infinite-horizon discounted cost problems, Shapley [82] showed that the associated game has a value, i.e.

$$\min_{\pi \in \Pi_{h,r}} \max_{m \in \mathcal{M}} C_m(\pi) = \max_{m \in \mathcal{M}} \min_{\pi \in \Pi_{h,r}} C_m(\pi).$$

For finite-horizon problems, the value  $V^*(1, x)$  of the game initialized in state  $x$  at time 1 can be computed by the recursion

$$\begin{aligned} V^*(T, x) &= \min_{a \in \mathcal{A}(x)} c(x, a), \quad x \in \mathcal{X}, \\ V^*(t, x) &= \min_{a \in \mathcal{A}(x)} \sup_{p(\cdot; x, a) \in \mathcal{P}_x} \left[ c(x, a) + \sum_{y \in \mathcal{X}} p(y; x, a) V^*(t+1, y) \right], \quad x \in \mathcal{X}, \quad t = 1, \dots, T-1. \end{aligned}$$

For infinite-horizon discounted cost problem, the value of the game satisfies Shapley's equations

$$V^*(x) = \min_{a \in \mathcal{A}(x)} \sup_{p(\cdot; x, a) \in \mathcal{P}_x} \left[ c(x, a) + \alpha \sum_{y \in \mathcal{X}} p(y; x, a) V^*(y) \right], \quad x \in \mathcal{X}. \quad (2.4.3)$$

The respective optimal policies for the controller and nature are the ones that achieve the minimum and maximum in (2.4.3). They are deterministic Markovian (and stationary for infinite-horizon games) independently of the policy space  $\Pi_h$ ,  $\Pi_m$  or  $\Pi_s$  for the controller and nature.

The rest of this subsection looks in more detail into the structure of the zero-sum Markov games associated with the worst-case control of uncertain MDPs. Although we can state some insightful properties, the complexity picture of zero-sum sequential Markov games is not completely clear yet.

For a state  $x \in \mathcal{X}$ , an action  $a \in \mathcal{A}(x)$ , and a transition probability vector  $p(\cdot; x, \cdot) \in \mathcal{P}_x$ , define the mapping  $\mathbf{T}_{x,a,p}$  from  $\mathbb{R}^n$  to  $\mathbb{R}$  by  $\mathbf{T}_{x,a,p}V = c(x, a) +$

$\alpha \sum_{y \in \mathcal{X}} p(y; x, a) V(y)$ . Define the operator  $\mathbf{T}$  on  $\mathbb{R}^n$  by

$$(\mathbf{T}V)(x) = \min_{a \in \mathcal{A}(x)} \sup_{p \in \mathcal{P}_x} \mathbf{T}_{x,a,p} V.$$

Shapley's equations state that the value of the game  $V^*$  is a fixed point of the operator  $\mathbf{T}$ . We will also need the operator  $\mathbf{T}_{\pi,\mu}$  on  $\mathbb{R}^n$  defined for a fixed controller's policy  $\pi$  and nature's policy  $\mu$  by

$$(\mathbf{T}_{\pi,\mu} V)(x) = \sum_{a \in \mathcal{A}(x)} \pi(a|x) \sum_{p \in \mathcal{P}_x} \mu(p|x, a) \mathbf{T}_{x,a,p} V.$$

The operator  $\mathbf{T}$  is an  $\alpha$ -contraction for the sup-norm on  $\mathbb{R}^n$ . Hence, the value  $V^*$  is the unique fixed point of  $\mathbf{T}$  in  $\mathbb{R}^n$ . Moreover, the sequence  $(V_t)$  produced by the value iteration algorithm,  $V_{t+1} = \mathbf{T}V_t$ , converges geometrically, at rate  $\alpha$ , to  $V^*$ , for any starting point  $V_0$ .

Although the sequence  $(V_t)$  usually keeps changing, we will see that the policy that is greedy with respect to  $V_t$  becomes constant after a polynomial number of iterations, for a **fixed discount factor**  $\alpha \in [0, 1)$ .

Let  $\delta \in \mathbb{N}$  be the accuracy of the rational data, i.e., the smallest natural number such that  $\delta\alpha$  and  $\delta p(y; x, a)$  are integer-valued for all  $y, x, a$ .

**Lemma 2.4.3.** *Assume that all immediate expected costs  $c(x, a)$  are integer-valued and that there exists  $\bar{c} < +\infty$  such that  $\max_{(x,a) \in \mathcal{X}\mathcal{A}} |c(x, a)| \leq \bar{c}$ . Then there is a smallest positive integer  $t^*$  such that for all  $t \geq t^*$ ,  $\mathbf{T}V_t = \mathbf{T}_{\pi,\mu} V_t$  implies  $\mathbf{T}V^* = \mathbf{T}_{\pi,\mu} V^*$ . Furthermore,  $t^* \leq \hat{t}$  with*

$$\hat{t} = \left\lceil \log \left( \frac{2\delta^{2n}(1+\alpha)^n}{(\|V_0\|_\infty + \bar{c}/(1-\alpha))} \right) / \log(1/\alpha) \right\rceil.$$

*Proof.* The proof uses an analogous argument to Lemma 1 in [97]. It is reproduced here (with a minor strengthening) for completeness and with an improved bound  $\hat{t}$ .

Since  $\mathbf{T}$  is an  $\alpha$ -contraction for the sup-norm on  $\mathbb{R}^n$ , after  $t = \lceil \log(\epsilon/\|V_0 - V^*\|_\infty) / \log(\alpha) \rceil$  iterations, we will have  $\|V_t - V^*\|_\infty \leq \epsilon$ , for all  $\epsilon > 0$ . We will

conclude the proof by showing that when  $\|V - V^*\|_\infty \leq \epsilon$  with  $\epsilon \leq 1/(2\delta^{2n}(1 + \alpha)^n)$ , a deterministic policy greedy with respect to  $V$  is optimal (in [97], the analogous expression involves  $1/(2\delta^{2n}n^n)$ ).

When  $\max_{(x,a) \in \mathcal{X}\mathcal{A}} |c(x, a)| \leq \bar{c} < +\infty$ , it is well-known that the value  $V^*$  satisfies  $\|V^*\|_\infty \leq \bar{c}/(1 - \alpha)$ . The bound  $\hat{t}$  is obtained by replacing  $\|V_0 - V^*\|_\infty$  in the previously obtained bound  $\lceil \log(\epsilon/\|V_0 - V^*\|_\infty)/\log(\alpha) \rceil$  by  $\|V_0\|_\infty + \bar{c}/(1 - \alpha)$ .

Let  $\pi^*$  and  $\mu^*$  be respectively optimal deterministic stationary policies for the controller and nature, which induce a Markov chain on  $\mathcal{X}$  with a transition probability matrix denoted by  $Q$ . The game value  $V^*$  satisfies  $V^* = \delta^2[\delta^2(I - \alpha Q)]^{-1}c$ , where  $\delta^2(I - \alpha Q)$  is a  $n \times n$  matrix with integer-valued entries. By Cramer's rule, its inverse is a rational matrix with a maximum denominator equal to the  $\det[\delta^2(I - \alpha Q)] = \delta^{2n} \det(I - \alpha Q)$ . Denote  $\chi = |\det(I - \alpha Q)| \in \mathbb{N}/\delta^{2n}$ , the absolute the value of the determinant of  $(I - \alpha Q)$ . We can write  $V^*(x) = \delta^2 W(x)/(\delta^{2n}\chi)$ , where  $W \in \mathbb{Z}^n$  is integer-valued.

If  $x \in \mathcal{X}$ ,  $a \in \mathcal{A}(x)$  and  $p(\cdot; x, a) \in \mathcal{P}_{x,a}$  are such that  $\mathbf{T}_{x,a,p(\cdot; x, a)} V^* \neq V^*(x)$ , then

$$\begin{aligned} \mathbf{T}_{x,a,p(\cdot; x, a)} V^* &= c(x, a) + \alpha \sum_{y \in \mathcal{X}} p(y; x, a) V^*(y) \\ &= \frac{c(x, a)\delta^{2n}\chi + \alpha \sum_{y \in \mathcal{X}} p(y; x, a)\delta^2 W(y)}{\delta^{2n}\chi} \neq \frac{\delta^2 W(x)}{\delta^{2n}\chi} = V^*(x). \end{aligned}$$

The numerators are integer-valued. Consequently, the two fractions differ by at least  $1/(\delta^{2n}\chi)$ .

Since  $Q$  is a stochastic matrix, the magnitude of its eigenvalues is less than or equal to one by Perron-Frobenius theory. Hence,  $\chi \leq (1 + \alpha)^n$ . As a result, the condition  $\|V_t - V^*\|_\infty \leq 1/2\delta^{2n}(1 + \alpha)^n$  implies that  $\|V_t - V^*\|_\infty < 1/2\delta^{2n}\chi$ . For the sake of contradiction, assume there is a deterministic policy that is greedy with respect to  $V_t$  with  $t \geq \hat{t}$  and that is not optimal. It follows that we have for some  $x, a$  and  $p(\cdot; x, a)$  that  $\mathbf{T}_{x,a,p(\cdot; x, a)} V^* \neq V^*(x)$ , and the difference is at least  $1/\delta^{2n}\chi$ . Since  $t \geq \hat{t}$ , we showed at the beginning of the proof that  $\|V_t - V^*\|_\infty \leq 1/(2\delta^{2n}(1 + \alpha)^n)$ .

Then we have the following contradiction:

$$\begin{aligned}
|\mathbf{T}_{x,a,p(\cdot;x,a)}V_t - V^*(x)| &= |\alpha \sum_{y \in \mathcal{X}} p(y; x, a)(V_t(y) - V^*(y)) + \mathbf{T}_{x,a,p(\cdot;x,a)}V^* - V^*(x)| \\
&\geq |\mathbf{T}_{x,a,p(\cdot;x,a)}V^* - V^*(x)| - |\sum_{y \in \mathcal{X}} p(y; x, a)(V_t(y) - V^*(y))| \\
&\geq 1/\delta^{2n}\chi - 1/2\delta^{2n}\chi \\
&\geq 1/2\delta^{2n}\chi \\
&> \|V_t - V^*\|_\infty
\end{aligned}$$

But, since  $\mathbf{T}$  is a contraction,  $|\mathbf{T}V_t(x) - V^*(x)| \leq \alpha \|V_t - V^*\|_\infty$ . This concludes the proof.  $\square$

The expression for  $\hat{t}$  should be compared to the instance description length, which is  $O(rn^2 \log \delta)$ . Hence, when  $\alpha$  is fixed, an optimal policy for the controller and nature can be computed in polynomial-time. By fixing these policies, the Markov game becomes a Markov chain for which we can compute in polynomial time the expected cost by solving a system of linear equations (cf. Subsection 2.2.1). However, the complexity of zero-sum sequential Markov game is still unknown when the discount factor  $\alpha$  is free.

Sequential games under an average-cost criterion are more subtle than discounted games, and a comprehensive analysis of these games is beyond the scope of this dissertation. In some cases, they have a value that satisfies an average-cost Shapley equation, but there are many caveats and pitfalls. In any case, average-cost games cannot be easier than  $\alpha$ -discounted games since we can reformulate the latter into the former by a well-known random restart procedure (e.g., p. 5 in [20]).

In this section, we established that most worst-case formulations are NP-hard. When the uncertainty set is state-rectangular, the worst-case formulation is equivalent to a zero-sum Markov game between controller and nature. These game enjoy attractive structural properties such as Shapley's equations and computational algorithms such as value iteration, which makes this formulation for the control of

uncertain MDPs appealing.

## 2.5 Stationary control of uncertain MDPs: the case of random uncertainty

In this section, we assume that the uncertain transition kernel  $P \in \mathcal{P}$  is sampled according to a known distribution  $q$ . However, the realization of the uncertain parameter  $P \in \mathcal{P}$  is not observed by the controller, who wants to minimize his expected cost over the uncertain parameters. For finite-horizon problems, the controller solves

$$\inf_{\pi \in \Pi_{m,r}} \sum_{m \in \mathcal{M}} q_m C_m(\pi), \quad (2.5.1)$$

and for infinite-horizon problems it solves

$$\inf_{\pi \in \Pi_{s,r}} \sum_{m \in \mathcal{M}} q_m C_m(\pi). \quad (2.5.2)$$

We study the complexity of deciding whether the above infima are below a given threshold for a general uncertainty set and for a fixed number of scenarios. Inspired by the positive results for the worst-case control with rectangular uncertainty, we also investigate the case of random rectangular uncertainty.

Our findings are summarized in the following table, for the case where the controller uses deterministic or randomized policies.

When the uncertainty set is state-rectangular, even policy evaluation is NP-hard.

Uncertainty type	Finite-horizon	Discounted cost	Average cost
General	NP-complete	NP-complete	NP-complete
2 MDPs	NP-complete	NP-complete	NP-complete
State(-action) rectangular	NP-hard	NP-hard	NP-hard

Table 2.3: Summary of the complexity of Markovian control of the average performance of uncertain MDPs

However, in the case where the uncertainty set is state-rectangular and that the

system does not come back to already left states, we prove that modified Bellman's equations hold and yield an optimal deterministic or randomized policy.

### 2.5.1 General uncertainty

When the uncertainty set is general and  $q_m = 1/|\mathcal{M}|$ , Corollary 2 in [58] shows that the control of uncertain MDP with random uncertainty (2.5.2) is NP-complete.

### 2.5.2 Only two possible MDP models

The stationary control of an uncertain MDP with random uncertainty is NP-complete, even when there are only two possible scenarios.

**Theorem 2.5.1.** *Even when there are only two possible parameter values in an uncertain MDP, deciding whether there is a deterministic or randomized stationary policy that achieves an expected cost less than a threshold is NP-complete for the infinite-horizon discounted and average cost problems.*

*Under the same conditions, deciding whether there is a Markovian policy that achieves a finite-horizon expected cost less than a threshold is NP-complete.*

*Proof.* The proof is an immediate adaptation of the proof of Theorem 2.4.1 since the maximum of two non-negative costs is zero if and only if their average is zero.  $\square$

### 2.5.3 Rectangular uncertainty

When the random uncertainty is state-action rectangular, the *evaluation* of a stationary policy is NP-hard. But it is not known to be in NP when the uncertainty set is state-rectangular (or state-action rectangular), and therefore, exponentially large in the instance's description length.

**Theorem 2.5.2.** *When the random uncertainty is state- (or state-action-) rectangular, deciding whether a given, deterministic or randomized, stationary policy achieves an expected cost lower than a given threshold is NP-hard for the finite horizon case, the infinite horizon discounted and average cost case.*

*Proof.* This proof uses a reduction from NETWORK RELIABILITY, which is NP-hard [26]. The corresponding uncertainty set is state- and state-action-rectangular so that the two corresponding control problems are NP-hard.

The statement of NETWORK RELIABILITY is as follow.

INSTANCE: Undirected graph  $G = (V, E)$ , two vertices  $s, t \in V$ , a rational failure probability  $p(e) \in [0, 1]$  for each edge  $e \in E$ , a positive rational number  $q \leq 1$ . Let  $v = |V|$  and  $m = |E|$ .

QUESTION: Assuming edge failures are independent of one another, is the probability that there is at least one path from  $s$  to  $t$  containing no failed edge larger than or equal to  $q$ ?

With an instance of this problem, let us associate the uncertain MDP with states  $V$  and initial state  $s$ . At each state  $v \in V - \{t\}$ , the system is reset to  $s$  with probability  $\rho \in (0, 1)$ , or an edge  $e = (v, v')$  is picked randomly uniformly at no cost. If the edge  $e$  is not failed, the system moves with probability one to  $v'$ ; otherwise the system is reset to  $s$ . When the system reaches  $t$ , it occurs a cost of 1 and is absorbed. A scenario is defined by the status of all edges. Since failures occur independently, this induces a state- (and state-action-) rectangular uncertainty.

Let us start with the analysis of the average cost problem. Here, there is a recurring cost of one in state  $t$ . If there is a simple path containing no failed edge from  $s$  to  $t$ , the system will eventually reach  $t$ . Hence, the average cost in such a scenario is 1. If there is no path between  $s$  and  $t$ , the cost is zero. Hence, the average cost over all scenarios will be greater than or equal to  $q$  if and only if the probability that there is a path from  $s$  to  $t$  without failed edge is greater than or equal to  $q$ .

The finite horizon and infinite horizon discounted cases require a finer analysis. Let  $N$  be the random number of transitions starting from  $s$  before  $t$  is reached. When there is no simple path from  $s$  to  $t$  without a failed edge,  $N = +\infty$  with probability one and the expected cost is zero.

Now, consider a case where there is a path from  $s$  to  $t$  without a failed edge. Starting from any state that is connected to state  $t$ , a path leading to  $t$  is followed with probability at least  $p = (1 - \rho)^{v-1}(1/m)^{v-1}$ . Indeed, there is such a path of

length less than  $v$ , and the correct edge is selected each time independently with probability at least  $1/m$ . Moreover,  $P(N \geq kv) \leq (1 - p)^k$  for  $k \geq 1$ . Since the system starts from state  $s$ , which we assumed connected to  $t$ , all the states on the trajectory are also connected to  $t$ . Hence, considering the trajectory on the time intervals  $(rv, (r + 1)v)$ ,  $r = 0, \dots, k - 1$ , we have  $P(N \geq kv) \leq (1 - p)^k$ , and  $P(N \geq kv) \rightarrow 0$  as  $k \rightarrow +\infty$ .

Let  $\delta$  be the product of the denominators of  $p(e)$  for all edges  $e \in E$ , and let  $\underline{q}$  be the largest rational with denominator  $\delta$  that is less than  $q$ ,  $q \geq \underline{q}$ . Choose the time horizon length  $T = kv$  with  $k$  the smallest integer such that  $(1 - p)^k \rho^{k-1} < \max(q - \underline{q}, 1/\delta)$ . Such a time horizon  $T$  is polynomial.

If the equality  $q = \underline{q}$  holds, the expected cost incurred by the system on the horizon  $T$  is at least  $\underline{q} - 1/\delta$  if and only if the probability that there is a path from  $s$  to  $t$  without failed edge is at least  $q$ . Similarly, if  $q > \underline{q}$ , the expected cost incurred by the system on the horizon  $T$  is at least  $\underline{q}$  if and only if the probability that there is a path from  $s$  to  $t$  without failed edge is at least  $q$ .

Now, let us deal with the infinite horizon discounted cost problem with discount factor  $\alpha \in [0, 1)$ . When a path exists, the expectation  $E[\alpha^N]$  goes to one from below as  $\alpha \rightarrow 1$ . Choosing  $\alpha$  close enough to one so that  $1 - E[\alpha^N] < \max(q - \underline{q}, 1/\delta)$  yields a statement analogous to the finite horizon case.  $\square$

### **A tractable special case under rectangular random uncertainty**

As we observed in Subsection 2.3.1, an uncertain MDP with state-rectangular random uncertainty does not reduce to an MDP because the system may come back to states visited earlier. When the trajectory of the system does not come back to states that have already been left (but loops are allowed), an uncertain MDP with rectangular uncertainty is tractable. This situation will illustrate a context where randomization can improve the controller's performance, in contrast to the optimal control of known MDPs.

**Proposition 2.5.3.** *Assume that the expected immediate cost of all transitions is bounded uniformly by  $B$  and that the random uncertainty's probability distribution  $q$*

is state-rectangular, i.e., the uncertain parameters  $(p(\cdot; x, \cdot), c(x, \cdot))$  at state  $x \in \mathcal{X}$  take value  $(p_k(\cdot; x, \cdot), c_k(x, \cdot))$  with probability  $q_x(k)$  for  $k = 1, \dots, K(x)$ , independently of the value of the uncertain parameters at other states. Let  $p_k^\pi(y; x) = \sum_{a \in \mathcal{A}(x)} \pi(a|x) p_k(y; x, a)$  and  $c_k^\pi(x) = \sum_{a \in \mathcal{A}(x)} \pi(a|x) c_k(x, a, y)$  be respectively the probability of moving to state  $y$  and the immediate cost, given that the system is in state  $x$  under a stationary policy  $\pi \in \Pi_{s,r}$  and the  $k$ th uncertain parameter in state  $x$  is realized.

If in all scenarios and under all policies the system does not come back to states that have already been left, i.e.,  $\pi \in \Pi_{s,r}$ ,  $x^\pi(t) \neq x^\pi(t+1) \Rightarrow x^\pi(\tau) \neq x^\pi(t), \forall \tau > t$  with probability one, then the following holds:

- (a) The expected cost  $V^\pi(x) = E[\sum_{t \geq 1} \alpha^t c^\pi(t) | x(0) = x]$  incurred by a controller following a stationary policy  $\pi \in \Pi_{s,r}$ , starting from state  $x \in \mathcal{X}$ , is the unique solution of the modified Bellman's equations

$$V^\pi(x) = \sum_{k=1}^{K(x)} q_x(k) \frac{1}{1 - \alpha p_k^\pi(x; x)} \left[ c_k^\pi(x) + \alpha \sum_{y \neq x} \frac{p_k^\pi(y; x)}{1 - p_k^\pi(x; x)} V^\pi(y) \right], \quad x \in \mathcal{X}. \quad (2.5.3)$$

- (b) For all states  $x \in \mathcal{X}$ , the optimal expected cost  $V^*(x) = \inf_{\pi \in \Pi_{s,r}} V^\pi(x)$  satisfies the modified Bellman's equations

$$V^*(x) = \inf_{\pi(\cdot|x)} \sum_{k=1}^{K(x)} q_x(k) \frac{1}{1 - \alpha p_k^\pi(x; x)} \left[ c_k^\pi(x) + \alpha \sum_{y \neq x} \frac{p_k^\pi(y; x)}{1 - p_k^\pi(x; x)} V^*(y) \right], \quad x \in \mathcal{X}. \quad (2.5.4)$$

- (c) The stationary policy that uses the randomization that achieves the minimum in (2.5.4) is optimal among the stationary Markovian policies.

Similarly, the deterministic stationary policy that achieves the minimum over the action choice in Bellman equations (2.5.4) is optimal among the deterministic stationary policies.

*Proof.* First, let us make some preliminary observations.

Let  $\Gamma = (\mathcal{X}, \mathcal{E})$  be the directed graph on the state space. The set of edges  $\mathcal{E}$  contains all the directed pairs of states  $(x, y)$  such that there is a model  $m \in \mathcal{M}$  and an action  $a \in \mathcal{A}(x)$  leading from  $x$  to  $y$  with positive probability in model  $m$ , i.e.,  $\mathcal{E} = \{(x, y) \in \mathcal{X}^2 \mid \exists m \in \mathcal{M}, a \in \mathcal{A}(x) : p_m(y; x, a) > 0\}$ .

Under the proposition's assumptions, it is easy to see that the directed graph  $\Gamma$  cannot contain loops of length more than one. Aside from possible loops from one state to itself, the graph  $\Gamma$  is essentially acyclic. As a result, it is possible to index the states by integers such that the state index increases along the system trajectories with probability one.

Since all the costs are bounded by  $B < +\infty$  and the discount factor  $\alpha$  satisfies  $0 < \alpha < 1$ , all the expectations appearing in the proposition are well-defined and finite.

(a) Let us fix a stationary policy  $\pi \in \Pi_{s,r}$ . By conditioning on the uncertain model parameter in state  $x$ , we have  $V^\pi(x) = \sum_{k=1}^{K(x)} q_x(k) E[\sum_{t \geq 1} \alpha^t c^\pi(t) \mid x(0) = x, k]$ .

If the state  $x$  is not a leaf node under policy  $\pi$  and uncertain parameter  $k \in K(x)$ , i.e.,  $p_k^\pi(x; x) < 1$ , let  $T_{k,x}^\pi \geq 0$  be the random number of loops in state  $x$  before the system moves to another state. Observe that the random variable  $T_{k,x}^\pi$  is geometrically distributed, i.e.,  $P(T_{k,x}^\pi = t) = p_k^\pi(x; x)^t (1 - p_k^\pi(x; x))$ , and hence “memoryless”. Furthermore,

$$\begin{aligned} E \left[ \sum_{t=0}^{T_{k,x}^\pi} \alpha^t \right] &= E \left[ \frac{1 - \alpha^{T_{k,x}^\pi + 1}}{1 - \alpha} \right] \\ &= \frac{1}{1 - \alpha p_k^\pi(x; x)}, \end{aligned}$$

and

$$E \left[ \sum_{t=0}^{T_{k,x}^\pi} \alpha^t c^\pi(t) \right] = c_k^\pi(x) \frac{1}{1 - \alpha p_k^\pi(x; x)}.$$

Given  $k$ , the random new state  $Y$  visited after state  $x$  is  $y$ , with probability  $p_k^\pi(y; x)/(1 - p_k^\pi(x; x))$ , independently of  $T_{k,x}^\pi$ .

If the state  $x$  is absorbing under policy  $\pi$  and parameter  $k$ , i.e.,  $p_k^\pi(x; x) = 1$ , then

$T_{k,x}^\pi = +\infty$  with probability one, and there is no node subsequently visited. In this case, we let  $\frac{p_k^\pi(y;x)}{1-p_k^\pi(x;x)} = 0$ , for  $y \neq x$ .

By the law of iterated expectations, we have

$$\begin{aligned}
V^\pi(x) &= \sum_{k=1}^{K(x)} q_x(k) \left( E \left[ \sum_{t=0}^{T_{k,x}^\pi} c^\pi(t) \mid x(0) = x, k \right] + E \left[ \sum_{t \geq T_{k,x}^\pi} c^\pi(t) \mid x(0) = x, k \right] \right) \\
&= \sum_{k=1}^{K(x)} q_x(k) \left\{ c_k^\pi(x) \frac{1}{1 - \alpha p_k^\pi(x;x)} + \sum_{y \neq x} \frac{p_k^\pi(y;x)}{(1 - p_k^\pi(x;x))} E \left[ \sum_{t \geq T_{k,x}^\pi + 1} \alpha^t c^\pi(t) \mid x(0) = x, k, Y = y \right] \right\} \\
&= \sum_{k=1}^{K(x)} q_x(k) \frac{1}{1 - \alpha p_k^\pi(x;x)} \left\{ c_k^\pi(x) + \alpha \sum_{y \neq x} \frac{p_k^\pi(y;x)}{1 - p_k^\pi(x;x)} E \left[ \sum_{t \geq 0} \alpha^t c^\pi(t + T_{k,x}^\pi + 1) \mid x, k, Y = y \right] \right\} \\
&= \sum_{k=1}^{K(x)} q_x(k) \frac{1}{1 - \alpha p_k^\pi(x;x)} \left\{ c_k^\pi(x) + \alpha \sum_{y \neq x} \frac{p_k^\pi(y;x)}{1 - p_k^\pi(x;x)} E \left[ \sum_{t \geq 0} \alpha^t c^\pi(t) \mid x(0) = y \right] \right\}.
\end{aligned}$$

The last equality follows from the rectangularity of the uncertain parameter distribution and the Markov property of the system trajectory given each parameter value.

This establishes that the Bellman's equations (2.5.3) hold for any fixed stationary policy  $\pi \in \Pi_{s,r}$ .

(b-c) By induction on the states in the order of decreasing index, there is a unique solution  $V^*$  to the Bellman's equations (2.5.4). Furthermore, using Equation (2.5.3), it follows that a stationary policy, which is greedy with respect to  $V^*$  is optimal among the stationary policies.

A similar argument takes care of deterministic policies. □

In contrast to standard dynamic programming, the optimal policy might be strictly randomized as illustrated in the following example. Consider an uncertain MDP with three states: one starting node  $S$  and the two absorbing ones  $E$  and  $G$  with zero cost. The controller chooses either action  $NM$ , which leads with certainty and no cost to state  $E$ , or choose action  $M$ . Under action  $M$ , it is uncertain whether the system will stay in  $S$  with a cost of 1 (scenario 1) or move to state  $G$  with a cost of  $-2$  (scenario 2). The model is illustrated in Figure 2-2. The controller thinks that both models are

equally likely and needs to pick a stationary policy to minimize its discounted cost with  $\alpha = 0.8$ .

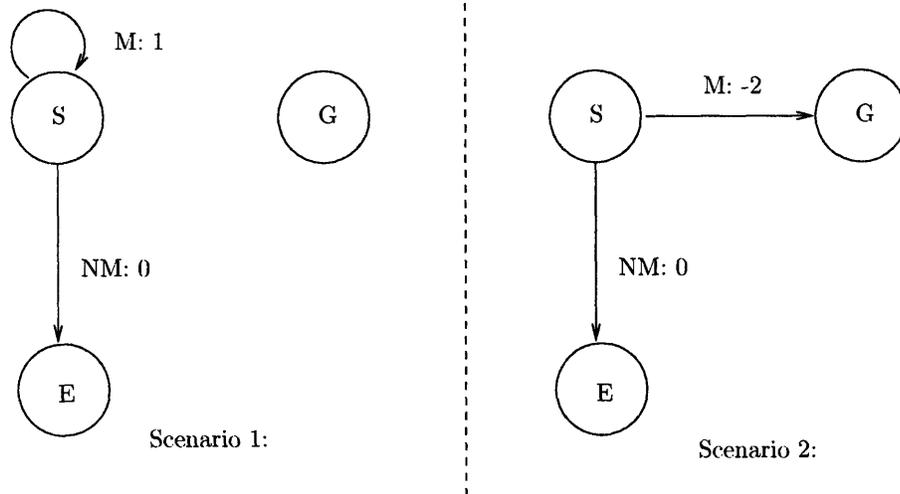


Figure 2-2: Simple problem where the optimal stationary policy is randomized.

The value function in  $E$  and  $G$  are both zero, i.e.,  $V^*(E) = V^*(G) = 0$ . Let  $\pi \in [0, 1]$  be the probability of choosing  $M$  in  $S$ . The Bellman equation from Proposition 2.5.3 for state  $S$  becomes

$$V^*(S) = \min_{\pi \in [0,1]} \frac{1}{2} \cdot \frac{\pi}{1 - \alpha\pi} - \pi$$

The policy that never mails has an expected cost of 0, whereas the policy that always mails has an expected cost of 1.5. The optimal stationary policy corresponds to  $\pi^* = \frac{1 - \sqrt{1/2}}{\alpha} \approx 0.3611$  and yields an expected cost of  $-0.1072$ , which is better than both deterministic policies.

In this section, we established that most formulations for the control of uncertain MDPs are at least NP-hard, with the exception of a special case where the principle of optimality is preserved.

## 2.6 Stationary control of uncertain MDPs: the worst-case regret case

In this section, the controller minimizes its worst-case regret, namely for finite-horizon problems, it solves

$$\inf_{\pi \in \Pi_{m,r}} \sup_{m \in \mathcal{M}} [C_m(\pi) - C_m(\pi^*(m))], \quad (2.6.1)$$

and for infinite-horizon problems, it solves

$$\inf_{\pi \in \Pi_{s,r}} \sup_{m \in \mathcal{M}} [C_m(\pi) - C_m(\pi^*(m))], \quad (2.6.2)$$

where  $\pi^*(m)$  is an optimal policy when scenario  $m$  is known to have occurred.

We analyze the complexity of deciding whether these infima are below a given threshold. Our results in this section are summarized in the following table, when the controller uses either deterministic or randomized policies. But the question whether rectangular uncertainty sets makes the worst-case regret minimization problem easier to solve remains open.

Uncertainty type	Finite-horizon	Discounted cost	Average cost
General	NP-complete	NP-complete	NP-complete
2 MDPs	NP-complete	NP-complete	NP-complete

Table 2.4: Summary of the complexity of worst-case regret control

### 2.6.1 General uncertainty set

When the uncertainty set is general, the argument of Subsection 2.4.1 showing that the worst-case control of uncertain MDPs is NP-complete can be tailored to uncertain MDP control problems based on worst-case regret. The finite-horizon and infinite-horizon discounted and average cost problems are NP-complete whether the controller uses randomized or deterministic policies.

## 2.6.2 Two possible MDP models

Even when there are only two possible scenarios, the robust control formulations based on worst-case regret are NP-complete.

**Theorem 2.6.1.** *Even when there are only two possible parameter values for an uncertain MDP, deciding whether there is a deterministic or randomized stationary policy that achieves a worst-case regret less than a threshold is NP-complete for the infinite horizon discounted and average cost problems.*

*Under the same conditions, deciding whether there is a Markovian policy that achieves a finite-horizon worst-case regret less than a threshold is NP-complete.*

*Proof.* The proof is an immediate adaptation of the proof of Theorem 2.4.1. Indeed, if the controller knows whether it is facing MDP Normal or Test, it can achieve zero cost. Hence, the worst-case regret is equal to the worst-case cost in the reduction used in the proof of Theorem 2.4.1.  $\square$

Even when the uncertainty is only on the cost parameter, the deterministic control problem is NP-complete.

**Theorem 2.6.2.** *Even when there are only two possible values for the uncertain cost parameter (and no uncertainty on the MDP dynamics  $P$ ), deciding whether there is a deterministic Markovian policy that achieves a worst-case regret on an uncertain MDP less than a given threshold is NP-complete in the finite horizon case, infinite horizon discounted, and average cost cases.*

*Proof.* First, we will reduce the NP-complete PARTITION problem (Problem [SP12], p. 223 in [26]) to a finite horizon worst-case regret problem.

INSTANCE: Finite set  $\mathcal{A}$  and size  $s(a) \in \mathbb{N}$  for each  $a \in \mathcal{A}$ .

PROBLEM: Is there a subset  $\mathcal{A}' \subset \mathcal{A}$  such that  $\sum_{a \in \mathcal{A}'} s(a) = \sum_{a \in \mathcal{A} - \mathcal{A}'} s(a) = K/2$ , where  $K = \sum_{a \in \mathcal{A}} s(a)$ ?

Given a partition instance defined as above with  $\mathcal{A} = \{a_1, \dots, a_n\}$ , let the time horizon be  $T = n$  and construct an uncertain MDP with  $n + 1$  states as illustrated in Figure 2-3. There is one state per element in  $\mathcal{A}$ , also noted  $a_i$ , with two available

actions  $S$  (element  $a_i$  Selected in  $\mathcal{A}'$ ) or  $NS$  ( $a_i$  Not Selected), plus a terminal state  $t$ . The system starts in state  $a_1$ . Under any of the two actions  $S$  or  $NS$ , the system moves from state  $a_i$ ,  $i = 1, \dots, n-1$  to state  $a_{i+1}$  and from  $a_n$  to  $t$ . Any deterministic stationary policy  $\pi$  defines a set  $\mathcal{A}' \subset \mathcal{A}$  by  $\{a \in \mathcal{A} \mid \pi(a) = S\}$ .

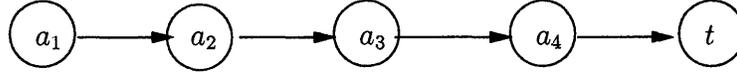


Figure 2-3: An illustration of the reduction of PARTITION with four elements to regret-based control of an uncertain MDP.

There are two possible MDP models. The first one has cost 0 under action  $NS$  and cost  $s(a)$  when taken in state  $a$ . In the second cost model, we exchange the role of the actions  $S$  and  $NS$ .

When the cost parameters are known to the controller, the optimal cost is always zero. Therefore, the worst-case regret is equal to the worst-case cost, which is  $\max \{ \sum_{a \in \mathcal{A}'} s(a), \sum_{a \in \mathcal{A} - \mathcal{A}'} s(a) \}$ . Hence, the maximum regret of a deterministic stationary policy is less than or equal to  $K/2$  if and only if the associated set  $\mathcal{A}'$  satisfies  $\sum_{a \in \mathcal{A}'} s(a) = \sum_{a \in \mathcal{A} - \mathcal{A}'} s(a)$ .

This argument can easily be adapted to infinite-horizon discounted cost and average cost problems. □

## 2.7 Conclusion

In this chapter, we classified several formulations for the optimal control of uncertain MDPs into the computational complexity hierarchy. In particular, we identified those that are solvable deterministically in polynomial time (in **P**), non-deterministically in polynomial time (in **NP**), and in polynomial space **PSPACE**. By studying different variations of the optimal control problems, we could pinpoint more accurately the source of complexity. Our results are summarized in Tables 2.1, 2.2, 2.3, and 2.6.

Most of the problems we considered are at least NP-hard. As a general rule of thumb, a problem is intractable as soon as the principle of optimality of dynamic

programming is lost.

Our findings shed some light on the modeling of uncertain MDPs. In the absence of strong motivation for coping with the complexity of harder models, the more tractable models bring the advantage of being computationally amenable, while capturing uncertainty. For example, the worst-case control of uncertain MDPs with rectangular uncertainty sets is a convenient model. On the downside, it can yield conservative controls.

The complexity of some problems remains unknown, including:

- the control of uncertain MDPs by history-dependent policies when the uncertainty set has only two (or a fixed number of) elements,
- the control of uncertain MDPs with random state-action-rectangular uncertainty by history-dependent policies,
- the worst-case control of infinite-horizon MDPs with rectangular uncertainty (or equivalently zero-sum sequential Markov games) when the objective is the expected average cost or discounted cost,
- the worst-case regret minimization when the uncertainty is state-rectangular or state-action-rectangular, either when the controller uses history-dependent policies, or only Markovian policies.



# Chapter 3

## Risk-averse and robust control of Markov Decision Processes

### 3.1 Introduction

#### 3.1.1 MDP control background and motivations

Markov Decision Processes (MDPs) [12] have been extensively studied in the control and operations research literature. They model decision making processes in discrete-time stochastic dynamical systems where the controller has the opportunity to reevaluate his decisions as information is revealed. MDPs have been used in many operational applications such as inventory control [94], pricing [59], network routing, customized marketing [86], and more.

Still, some important limitations hinder the applicability of MDP models. MDP models generally focus on the expected performance of a system in spite of the strong interest in risk-sensitive decision making, because the latter is theoretically and computationally challenging. Only exponential utility maximization has been identified to yield a tractable risk-sensitive control problem [36, 61].

Another important limitation to the outreach of MDP models in real-world applications is the sensitivity of the recommended policies on the input parameters together with the difficulty to accurately specify the parameters of these models.

Similar concerns have already been raised in the context of continuous-time control and more recently optimization (e.g., [92, 93, 90, 91, 27, 28, 15]). When dealing with MDPs, one needs to specify the distribution of random costs and state transitions, and there are typically many such parameters. Oftentimes, the parameters are estimated from data, for example system behaviors observed in the past, resulting in uncertainty about their values. These challenges raise the problem of making the solution of an MDP robust to uncertainty in its parameters, so that the recommended policies perform well in practice.

In this chapter, we will address the two aforementioned challenges to the normative application of MDP models in practice. On the one hand, we will propose a risk-sensitive framework for making decisions in an MDP that is computationally tractable. On the other hand, we will introduce a new formulation for robust control of an MDP, which is motivated from the perspective of risk-based decision theory and which is computationally tractable.

### 3.1.2 Literature review

Before getting into the details of our contributions, let us review (not exhaustively) three streams of research relevant to our work, namely utility-based risk-sensitive MDP control, multi-period convex risk measures, and worst-case control of MDPs.

**Risk-sensitive control of MDPs**, based on expected utility maximization, has been introduced in the seminal paper of Howard and Matheson [36], which minimizes the expectation of the exponential of the undiscounted sample cost of terminating MDPs. Recently, the same problem has been analyzed more comprehensively in [61]. Under some general assumptions, structural properties of the optimal solution are established and algorithms to compute the optimal expected utility and an optimal policy are provided. Also assuming an exponential utility function, the papers [74, 34] rely on Donsker-Varadhan large deviation results to study the average cost problem. They establish the equivalence of the risk-sensitive average cost problem with a related zero-sum game. The reader can refer to the review [46] for more details and references. Considering discounted cost, Coraluppi and Marcus [18] relate the *high-risk* limit

of the risk-sensitive control problem to a zero-sum game. Such games enjoy nice structural properties and are computationally “tractable” [82]. Although exponential utility maximization in an MDP often amounts to solving a zero-sum sequential game, not all worst-case control problems can be interpreted in terms of a risk-sensitive control problem with exponential utility function. Besides, when general utilities are considered, Bellman’s principle of optimality is lost on the original state space. Sometimes, the addition of continuous state variables allows the minimization of non-exponential utilities, but the computational complexity of the risk-sensitive MDP control problem increases significantly.

The notion of a **coherent or convex risk measure** [3, 25] captures the preferences of a decision maker over positions with uncertain outcomes (and not only the position’s “variability”). Convex risk measures lead to a different approach to risk-sensitive control than expected utility theory, and the present work will build on that concept. The axiomatic definition of coherent and convex risk measures is now widely accepted in the financial literature. The key property of convex risk measures is that they can be represented as worst-case penalized expectations. As a result, minimizing such risk measures amounts to solving a zero-sum game between the decision maker and nature. Originally introduced for single-period positions, coherent risk measures have been extended to multi-period settings, e.g., in [67, 4, 72, 76, 22]. The work [4] takes a different perspective than the other papers and our work. Intuitively, its authors map a discrete-time stochastic process to a random variable for which they can use single-period coherent risk measures. Hence, the coherence assumption for the multi-period risk measure has a meaning unique to their approach. Interestingly, they assess the risk of a trajectory, and not only the risk of a final pay-off, and they derive “Bellman’s equations” for some notion of conditional risk. In contrast, [67, 72, 22] require the risk measure to have an inter-temporal property: dynamic consistency. A multi-period risk measure is dynamically consistent as information is revealed, in the sense that if a position is acceptable at time  $t + 1$  for any possible new information, then it must be acceptable at time  $t$ . We will also require this dynamic consistency condition to hold. The papers [67, 72] independently established a representation re-

sult for dynamically consistent coherent risk measures over a finite sample space and a finite time horizon. The former reference did not study any control setting, while the latter found a risk minimizing hedging strategy for a financial portfolio as an application. This optimal hedging problem corresponds to a limited control setting, where the decision maker does not influence the uncertain outcome, but the authors proved that “Bellman equations” hold for that example. Reference [22] extended these two papers from coherent to convex multi-period risk measures. More recently, [76] generalized further the representation theorem to general functional spaces – although they consider only finitely many time periods – and showed that minimizing dynamically consistent convex risk measure amounts to solving a dynamic zero-sum game between the controller and nature. This last paper has the most overlap with our work in Section 3.2 but it came to our knowledge only at the time of finalizing this dissertation: it provides a more general representation theorem than the one we derived independently; but it does not mention the dynamic consistency condition on risk measures because it only considers risk measures that are obtained by the composition of a finite number of coherent “conditional risk mappings,” which are dynamically consistent and *coherent* by construction. In contrast, we show that “Bellman’s equations” hold for dynamically consistent *convex* risk measures, even in the case of an infinite time horizon. Most important, our work differs from the literature by the introduction of the notion of Markovian risk measures. It allows us to exploit the state as a “sufficient statistics” in theoretical and computational results (notably, it allows us to tackle infinite horizon problems). Moreover, the notion of Markovian risk measures allows us to push further the correspondence between risk minimization and zero-sum games against nature since dynamically consistent Markovian convex risk measures can be minimized by solving a Markov game between the controller and nature.

**Worst-case control of MDPs.** In the control literature, it has long been known that the optimal policy and the optimal expected cost of an MDP is quite sensitive to parameter variations in practice. For recent illustrations of this observation by simulation examples, see e.g., [44, 86, 54, 38]. To mitigate this problem, the controller

can try to minimize the cost associated with the worst-case parameters within some given uncertainty set. In the case of “rectangular uncertainty set” and finite-horizon cost or infinite-horizon discounted cost, this problem has been addressed by Satia and Lave [78], Nilim and El Ghaoui [54] and Iyengar [38]. Under their assumptions, the robust optimal control problem is a zero-sum sequential Markov game [82] between the controller and an adversary (say, nature) that chooses the value of the uncertain parameters. As a result, the controller and nature can select a deterministic Markovian policy without loss of optimality. In addition, when the time horizon is infinite, the optimal policies can be chosen stationary. This approach not only yields policies with strong worst-case performance guarantees, but such policies can be computed with little extra effort relative to the nominal problem for many interesting uncertainty sets. However, the rectangularity assumption is potentially very conservative. Another criticism is that the decision-theoretic foundation of the worst-case optimal control formulation has not been investigated for MDPs, despite some analogies with the multiple recursive priors from economic theory [24, 99].

### 3.1.3 Chapter contributions

In order of exposition, this chapter makes the following contributions.

1. We introduce and motivate the notion of a *Markovian* multi-period risk measure in Subsection 3.3.3 and consider the problem of minimizing a Markovian dynamically consistent convex risk measure of the sample cost, over all Markovian randomized policies. We prove that under mild assumptions a risk-minimizing policy can be selected to be deterministic Markovian for finite horizon problems (cf. Theorem 3.4.3), or deterministic Markovian and stationary for infinite horizon problems (cf. Theorems 3.5.2 and 3.5.6). With our definition of Markovian risk, an optimal policy can be computed efficiently by classical dynamic programming techniques such as value iteration, even when the time horizon is infinite. Moreover, we show that a dynamically consistent Markovian convex risk measure of the sample cost can be minimized by solving a certain zero-sum

sequential Markov game between the decision maker and nature.

2. In Section 3.6, we point out that the robust control of uncertain MDPs proposed in [38, 54, 78] amounts to minimizing a multi-period coherent risk measure of the sample cost. This insight justifies from a decision-theoretic perspective this robust formulation of MDP control with rectangular uncertainty sets. When the uncertainty sets are not state-rectangular, we illustrate that the optimal robust controls can be dynamically inconsistent, in addition to being computationally intractable as we showed in Chapter 2.
3. The same connection allows us to motivate a new robust formulation that is also a sequential Markov game under some natural assumptions. Nature still picks the worse parameter but she has to pay a penalty for using “unlikely” parameters. Such a formulation has the potential to mitigate the conservativeness of the (classic) worst-case formulation of [78, 54, 38].
4. Finally, we show how to build a multi-period risk measure that is dynamically consistent, Markovian and convex starting from single-period convex risk measures. For example, we can construct a natural extension of the single-period conditional value at risk (CVaR) to a multi-period setting, whereas the naive application of CVaR to multi-period problems yields a dynamically inconsistent risk measure.

### 3.1.4 Chapter structure

We begin in Section 3.2 with a review of single-period coherent and convex risk measures. In Section 3.3, we describe a discrete-time model of controlled stochastic dynamical systems, and define dynamically consistent and Markovian risk measures. Section 3.4 deals with the minimization of a Markovian dynamically consistent convex risk measure over a finite horizon, whereas Section 3.5 deals with the infinite horizon case. Finally, Section 3.6 introduces a new formulation of the robust control of an MDP, which amounts to minimizing a convex risk measure, and show how to construct

Markovian dynamically consistent convex risk measures from single-period ones.

## 3.2 Convex and coherent risk measures

This section reviews in a unified manner existing work on single-period coherent and convex risk measures. This notion of risk captures more than the variability of positions with uncertain outcome: it also defines an order of preference on these positions. We will show that the axiomatically defined notion of a convex risk measure is equivalent to a worst-case penalized expectation. Therefore, minimizing a convex risk measure is equivalent to solving a zero-sum game between the decision maker and nature.

Let  $\Omega$  be a sample space endowed with a  $\sigma$ -algebra  $\mathcal{F}$ , and let  $\mu$  be a probability measure on  $(\Omega, \mathcal{F})$ . We let  $L^1(\Omega, \mathcal{F}, \mu)$  be the vector space of measurable integrable real-valued functions. We define  $l_1(f) = \int |f| d\mu$ , which is the  $l_1$ -norm of  $f \in L^1(\Omega, \mathcal{F}, \mu)$ , and  $l_p$  which is the  $p$ -norm on  $L^p(\Omega, \mathcal{F}, \mu)$ . The non-negative elements  $f$  of  $L^p(\Omega, \mathcal{F}, \mu)$ ,  $1 \leq p \leq \infty$ , such that  $\int f d\mu = 1$ , will sometimes be viewed as probability measures on  $(\Omega, \mathcal{F})$ , absolutely continuous with respect to  $\mu$ .

Let  $(\mathcal{H}, l)$  be a normed vector space whose elements are functions from  $\Omega$  into  $\mathbb{R}$  and containing the constants. For  $X, Y \in \mathcal{H}$ , we write  $X \leq Y$  if  $X(\omega) \leq Y(\omega)$  with  $\mu$ -probability 1. The norm  $l$  induces a topology generated by the open balls in  $\mathcal{H}$ .

To be consistent with the intuitive notion of risk,  $X \in \mathcal{H}$  is thought of as the uncertain *cost* of a position, which depends on the uncertain outcome  $\omega \in \Omega$ . This contrasts with the usual approach in the financial literature, which sees  $X$  as a payoff, but it suits better our line of exposition.

### 3.2.1 Definition of convex and coherent risk measures

Coherent and convex risk measures are functionals on a position space  $(\mathcal{H}, l)$  that satisfy a few basic properties, which have been motivated in the seminal paper [3]. These concepts have been subsequently refined and generalized; see e.g., [77]. A decision maker has a (subjective) notion of risk that maps uncertain positions to

their risk; the lower the risk of a position, the more attractive it is to the decision maker.

**Definition 3.2.1.** *A convex risk measure on  $\mathcal{H}$  is a functional  $\rho : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  that satisfies the following properties:*

1. *Normalization:*  $\rho(0) = 0$ .
2. *Monotonicity:* If  $X, Y \in \mathcal{H}$  and  $X \leq Y$ , then  $\rho(X) \leq \rho(Y)$ .
3. *Translation invariance:*  $\forall m \in \mathbb{R}, \forall X \in \mathcal{H}, \rho(X + m) = \rho(X) + m$ .
4. *Convexity:*  $\rho(\lambda X + (1 - \lambda)Y) \leq \lambda\rho(X) + (1 - \lambda)\rho(Y)$  for all  $X, Y \in \mathcal{H}$  and  $\lambda \in [0, 1]$ .
5. *Lower semicontinuity:*  $\{X \in \mathcal{H} \mid \rho(X) \leq 0\}$  is  $l$ -closed in  $\mathcal{H}$ .

*If, in addition,  $\rho$  is positively homogeneous, i.e., if for all  $X \in \mathcal{H}$  and all  $\lambda \geq 0$  we have  $\rho(\lambda X) = \lambda\rho(X)$ , then  $\rho$  is called a coherent risk measure.*

Sometimes the risk of a position increases nonlinearly with the position (for example, when the market is not liquid). In that case, the risk is convex but not coherent.

The properties of a convex risk measure are self-explanatory, except the last two. The convexity property essentially states that diversification does not increase risk, which is arguably reasonable [3, 25]. The last property of semicontinuity is rather technical, but it is also reasonable that the risk perception of a decision maker has some smoothness so that a little perturbation of a position does not affect radically its risk.

The following two classical results from convex analysis (e.g., Theorems 4.24 and 4.25 in [1]) show that, in order to satisfy condition (5), it is sufficient to check whether  $\rho$  is bounded on a  $l$ -neighborhood of 0.

**Lemma 3.2.2.**

- (a) *If a convex function  $\rho$  is defined on a  $l$ -neighborhood of 0 and bounded above in that neighborhood, then it is  $l$ -continuous at zero.*

(b) If  $\rho$  is continuous at a point in a convex open subset  $S$  of a topological vector space, then  $\rho$  is continuous on  $S$ .

Finally, the following lemma provides two ways to build new convex risk measures from existing ones.

**Lemma 3.2.3.** For  $i = 1, \dots, m$ , let  $\rho_i$  be convex risk measure on  $\mathcal{H}$  and  $\alpha_i \geq 0$  be weights such that  $\sum_{i=1}^m \alpha_i = 1$ .

(a) The functional  $\rho$  defined by  $\rho(X) = \sum_{i=1}^m \alpha_i \rho_i(X)$  is a convex risk measure on  $\mathcal{H}$ .

(b) The functional  $\rho$  defined by  $\rho(X) = \max_{i=1, \dots, m} \rho_i(X)$  is a convex risk measure on  $\mathcal{H}$ .

### 3.2.2 Examples of coherent and convex risk measures, and connection with expected utility

In this subsection, we provide examples of coherent and convex risk measures from the literature, e.g. [3, 25, 77] to illustrate the generality and the relevance of these notions. These notions are particularly popular in the financial literature.

- The expectation  $\rho(X) = E_\mu[X]$  with respect to  $\mu$  is the simplest coherent risk measure on  $L^p(\Omega, \mathcal{F}, \mu)$ ,  $p \geq 1$ .
- If  $P_1, \dots, P_m \in L^\infty(\Omega, \mathcal{F}, \mu)$  are probability measures and  $\alpha_i$  are non-negative weights such that  $\sum_{i=1}^m \alpha_i = 1$ , then  $\rho(X) = \sum_{i=1}^m \alpha_i E_{P_i}[X]$  is a coherent risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ .
- Conditional value at risk (CVaR) is a popular coherent risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$  [71]. Intuitively, the  $\alpha$ -CVaR of a position  $X \in L^1(\Omega, \mathcal{F}, \mu)$  is the conditional mean of the worst  $\alpha$ -tail of  $X$  ( $\alpha \in [0, 1]$ ). For example, the 0-CVaR( $X$ ) is simply the expectation  $E_\mu[X]$ , whereas 1-CVaR( $X$ ) corresponds to the worst-case value of  $X$ .

Formally, define the distribution function  $F_X$  of  $X$  by  $F_X(x) = \mu(\{X(\omega) \leq x\})$ , the value at risk  $\alpha$ -VaR( $X$ ) of position  $X$  by  $\alpha$ -VaR( $X$ ) =  $\inf_x\{x \mid F_X(x) \geq \alpha\}$  and the auxiliary distribution function  $\psi_{X,\alpha}(x)$  of the “ $\alpha$ -tail” of  $X$  by

$$\psi_{X,\alpha}(x) = \begin{cases} 0 & \text{for } x < \alpha\text{-VaR}(X) \\ (F_X(x) - \alpha)/(1 - \alpha) & \text{for } x \geq \alpha\text{-VaR}(X). \end{cases}$$

Then the conditional value at risk of position  $X$ ,  $\alpha$ -CVaR( $X$ ), is the expectation of the distribution  $\psi_{X,\alpha}$ .

**Remark 3.2.4.** *When the distribution of  $X$  is not continuous, the choice that  $\psi_{X,\alpha}(x) = 0$  when  $x < \alpha$ -VaR( $X$ ), instead of, say  $x \leq \alpha$ -VaR( $X$ ), has an impact on the value of the conditional value at risk (cf. [71] p. 8).*

It can be shown [71] that

$$\alpha\text{-CVaR}(X) = \inf_{z \in \mathbb{R}} \left( z + (1 - \alpha)^{-1} \int \max(X(\omega) - z, 0) d\mu \right).$$

- When exponential utility functions are considered, the logarithm of the expected utility  $\rho(X) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma X))$  with  $\gamma > 0$  is a convex risk measure on  $L^\infty(\Omega, \mathcal{F}, \mu)$  [25].

### 3.2.3 Representation of convex risk measures

Now, we will see that convex risk measures have the key property of being representable as worst-case penalized expectations. Although a given convex risk measure  $\rho$  need not rely on a probabilistic model of the the uncertain outcome  $\omega \in \Omega$ ,  $\rho$  implicitly defines a family of “test” probabilistic models  $\mathcal{P}$  for the uncertainty  $\omega$  and a penalty function on  $\mathcal{P}$  such that for all positions  $X$ ,

$$\rho(X) = \sup_{P \in \mathcal{P}} (E_P[X] - \phi(P)).$$

Consequently, when a decision maker chooses a position among  $(X_u)_{u \in U}$  in  $\mathcal{H}$  that minimizes its risk  $\rho$ , it solves the zero-sum game

$$\inf_{u \in U} \sup_{P \in \mathcal{P}} (E_P[X_u] - \phi(P)).$$

From now on, we will consider the set of positions  $(\mathcal{H}, l)$  to be space of  $\mu$ -integrable functions on  $\Omega$  endowed with its canonical norm, that is  $(\mathcal{H}, l) = (L^1(\Omega, \mathcal{F}, \mu), l_1)$ , unless specified otherwise. Nonetheless, the probability measure  $\mu$  need not be interpreted as a probabilistic model for the uncertainty  $\omega$ . It is introduced mostly for technical reasons (in order to invoke duality in a functional space) and it matters essentially through its support and the behavior of its tail. When  $\Omega$  is finite, we have  $L^1(\Omega, \mathcal{F}, \mu) = L^\infty(\Omega, \mathcal{F}, \mu)$ . Furthermore, we can even assume that it contains only the points that have positive  $\mu$ -probability by redefining the sample space  $\Omega$ . In this case, the probability measure  $\mu$  becomes unnecessary.

**Remark 3.2.5.** *When  $\Omega$  is finite, most of the technicalities of functional analysis in this section disappear. The reader is encouraged to keep this case in mind. However, in order to deal with general systems in the sequel, especially infinite horizon problems, we will need general sample spaces, and hence a reference probability measure.*

Consider the bilinear mapping on the dual pair  $(L^1(\Omega, \mathcal{F}, \mu), L^\infty(\Omega, \mathcal{F}, \mu))$  defined by  $(X, Y) = \int XY d\mu$ . When  $Y \geq 0$  and  $\int Y d\mu = 1$ ,  $(X, Y)$  can be interpreted as the expectation of  $X$  with respect to the probability measure on  $(\Omega, \mathcal{F})$  that has derivative  $Y$  with respect to  $\mu$ . With a slight abuse of notation, we will denote  $E_P[X] = (X, P)$  for all  $P \in L^\infty(\Omega, \mathcal{F}, \mu)$ , even when  $P$  cannot be interpreted as a probability density function.

Now, we can state the representation theorem for convex risk measures, which, intuitively, parallels the well-known statement that a closed convex function can be represented as a supremum of affine functions. In our case, the affine functions can be chosen to be only penalized expectations.

**Theorem 3.2.6.** *Let  $\mathcal{P}_0$  be a set of “test” probability measures in  $L^\infty(\Omega, \mathcal{F}, \mu)$ , and let  $\phi_0 : L^\infty(\Omega, \mathcal{F}, \mu) \rightarrow \mathbb{R} \cup \{+\infty\}$  be a “penalty” function such that  $\inf_{P \in \mathcal{P}_0} \phi_0(P) = 0$ .*

The functional  $\rho$  defined by

$$\rho(X) = \sup_{P \in \mathcal{P}_0} (E_P[X] - \phi_0(P)). \quad (3.2.1)$$

is a convex risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ .

If, in addition,  $\phi_0 = 0$ , then the functional  $\rho$  is a coherent risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ .

Conversely, let  $\rho$  be a convex risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ . Define a penalty function  $\phi$  from  $L^\infty(\Omega, \mathcal{F}, \mu)$  into  $\mathbb{R} \cup \{+\infty\}$  by

$$\phi(P) = \sup_{X \in \mathcal{H}} (E_P[X] - \rho(X))$$

and a set of probability measures

$$\mathcal{P} := \left\{ P \in L^\infty(\Omega, \mathcal{F}, \mu) \mid P \geq 0, \int P d\mu = 1, \phi(P) < +\infty \right\}.$$

Then,  $\phi$  is convex and  $l_\infty$ -lower-semicontinuous,  $\mathcal{P}$  is convex, and

$$\rho(X) = \sup_{P \in \mathcal{P}} (E_P[X] - \phi(P)). \quad (3.2.2)$$

In addition, if  $\rho$  is coherent, we have  $\phi(P) = 0, \forall P \in \mathcal{P}$ .

This theorem unifies several previous results. For coherent risk measures, it is the seminal representation theorem of [37] or Proposition 4.1 in [3] when  $\Omega$  is finite, or Theorem 2.3 in [21] when  $\mathcal{H} = L^\infty(\Omega, \mathcal{F}, \mu)$  and  $\Omega$  is general. The extension to convex risk measures and  $\mathcal{H} = L^\infty(\Omega, \mathcal{F}, \mu)$  was provided in [25]. In essence, we follow the proof in [25] to deal with a more general positions space, namely  $\mathcal{H} = L^1(\Omega, \mathcal{F}, \mu)$ , with arbitrary sample space  $\Omega$ . During the preparation of this chapter, we have become aware of a more general theorem by Ruszczynski and Shapiro [77] that unifies all aforementioned results. Essentially, they give sufficient conditions on convex risk measures on a given position space so that they can be written as worst-case penalized expectations with respect to a set of probability measures in a given space. Our

setting (with Condition (5) in Definition 3.2.1) happens to be a particular case of their analysis.

*Proof.* The first part of the theorem is obvious since the functional  $\rho$  is defined as the supremum of affine functions. The assumption that  $\inf_{P \in \mathcal{P}_0} \phi_0(P) = 0$  is necessary to guarantee the normalization of the functional  $\rho(X) = \sup_{P \in \mathcal{P}_0} E_P[X] - \phi_0(P)$ .

The second part of the theorem states that essentially all the convex risk measures take the form of the convex risk measures specified in the first part.

First, the function  $\phi(P) = \sup_{X \in \mathcal{H}} (E_P[X] - \rho(X))$  defined in the theorem statement indeed maps  $L^\infty(\Omega, \mathcal{F}, \mu)$  into  $[0, +\infty]$  since  $\phi(P) \geq E_P[0] - \rho(0) = 0$ . Furthermore, the function  $\phi$  is convex and  $l_\infty$ -lower-semicontinuous as the supremum of convex lower-semicontinuous affine functions.

Consider the  $l_1$ -closed convex set  $C = \{X \in L^1(\Omega, \mathcal{F}, \mu) \mid \rho(X) \leq 0\}$ . We show now that  $\phi(P) = \sup_{X \in C} E_P[X]$  and, as a result, that  $\phi$  is positively homogeneous on  $L^\infty(\Omega, \mathcal{F}, \mu)$ . First, observe that  $\phi(P) = \sup_{X \in \mathcal{H}} (E_P[X] - \rho(X)) \leq \sup_{X \in C} E_P[X]$ . On the other hand, note that for any position  $X$  such that  $\rho(X) < +\infty$ , we have  $\rho(X - \rho(X)) = 0$  by translation invariance, and therefore  $X - \rho(X) \in C$ . Thus,

$$\phi(P) = \sup_{\{X \in \mathcal{H} \mid \rho(X) < +\infty\}} E_P[X - \rho(X)] \geq \sup_{Y \in C} E_P[Y].$$

This proves that  $\phi(P) = \sup_{X \in C} E_P[X]$ .

The inequality  $\rho(X) \geq \sup_P (E_P[X] - \phi(P))$  (Fenchel weak duality) is immediate. The fact that equality holds (strong duality) will follow from the bipolar theorem applied to  $C$ . In order to use the bipolar theorem, we need to show that  $C$  is closed for the weak topology  $\sigma(L^1(\Omega, \mathcal{F}, \mu), L^\infty(\Omega, \mathcal{F}, \mu))$ . By the Riesz theorem (Theorem 10.28, p. 354 in [1]), the norm dual space of  $(L^1(\Omega, \mathcal{F}, \mu), l_1)$  is isometric to  $(L^\infty(\Omega, \mathcal{F}, \mu), l_\infty)$ . Since the  $l_1$ -topology (and of course the weak topology) on  $L^1(\Omega, \mathcal{F}, \mu)$  are consistent with the dual pair  $(L^1(\Omega, \mathcal{F}, \mu), L^\infty(\Omega, \mathcal{F}, \mu))$ , the closed convex sets for the norm and the weak topology are the same (Theorem 4.72, p. 154 in [1]), and  $C$  is closed for the weak topology.

The polar  $C^\circ$  of  $C$  is defined by

$$C^\circ = \{P \in L^\infty(\Omega, \mathcal{F}, \mu) \mid E_P[X] \leq 1, \forall X \in C\}$$

and its bipolar is

$$C^{\circ\circ} = \{X \in L^1(\Omega, \mathcal{F}, \mu) \mid E_P[X] \leq 1, \forall P \in C^\circ\}.$$

Since  $C$  is convex,  $\sigma(L^1(\Omega, \mathcal{F}, \mu), L^\infty(\Omega, \mathcal{F}, \mu))$ -closed and contains zero, the bipolar theorem (Theorem 4.77, p. 157 in [1]) states that  $C = C^{\circ\circ}$ .

Now, we show that for all  $Y \in \mathcal{H}$  such that  $\rho(Y) > 0$ , there exists  $Q \in L^\infty(\Omega, \mathcal{F}, \mu)$  such that  $E_Q[Y] - \phi(Q) > 0$ . For the sake of contradiction, assume there exists a position  $Y \in \mathcal{H}$  such that  $\rho(Y) > 0$  and  $E_P[Y] - \sup_{X \in C} E_P[X] \leq 0$  for all  $P$ . Since  $Y$  is not in  $C = C^{\circ\circ}$ , by the bipolar theorem, there is  $Q \in C^\circ$  such that  $E_Q[Y] > 1$ , and by definition of the  $C^\circ$ ,  $\sup_{X \in C} E_Q[X] = \phi(Q) \leq 1$ . The last two inequalities yield the contradiction that  $E_Q[Y] - \phi(Q) > 0$ .

For the sake of contradiction, assume that strong duality does not hold, i.e., there exists  $X \in \mathcal{H}$  such that  $\rho(X) > \sup_P E_P[X] - \phi(P)$ . In particular,  $\sup_P E_P[X] - \phi(P) < +\infty$ . Hence, the position  $Y := X - \sup_P (E_P[X] - \phi(P))$  is in  $\mathcal{H}$  and  $\rho(Y) > 0$ . By the result of the previous paragraph, there exists a  $Q$  such that  $E_Q[X] - \sup_P (E_P[X] - \phi(P)) - \phi(Q) > 0$ . This contradiction concludes the proof that for all  $X \in \mathcal{H}$ ,  $\rho(X) = \sup_{P \in \mathcal{P}} (E_P[X] - \phi(P))$ .

Now, we show that the supremum can be restricted to the  $l_\infty$ -closed convex set of probability measures  $\mathcal{P}$  in  $L^\infty(\Omega, \mathcal{F}, \mu)$ . Recall that  $\phi(P) = \sup_{X \in C} E_P[X]$ . Since the non-positive functions in  $L^1(\Omega, \mathcal{F}, \mu)$  are in  $C$ ,  $\phi(P) = +\infty$  if  $P$  is not non-negative  $\mu$ -almost surely. In addition, since  $\phi$  is positively homogeneous,  $\mathcal{Q} := \{P \mid \phi(P) < +\infty\}$  is a convex cone included into the positive orthant. In the representation theorem, it is enough to pick one representative for each ray of  $\mathcal{Q}$ , namely define  $\mathcal{P} = \{P \in \mathcal{Q} \mid \int P d\mu = 1\}$ .

Finally, when the risk measure  $\rho$  is coherent, for all  $\lambda \geq 0$   $\rho(\lambda X) = \lambda \rho(X)$  and  $X \in C$  implies that  $\lambda X \in C$ . Since  $\phi(P) = \sup_{X \in C} E_P[X]$ , it follows that  $\phi(P) = +\infty$

as soon as  $\phi(P) > 0$ . □

**Remark 3.2.7.** *Starting from a penalty function  $\phi_0$  and a set of test probability measures  $\mathcal{P}_0$ , the first part of the previous theorem shows that the functional  $\rho(X) = \sup_{P \in \mathcal{P}_0} (E_P[X] - \phi_0(P))$  is a convex risk measure to which the second part of the theorem applies. Without getting into the technical details, it can be shown that the penalty function  $\phi$  induced by  $\rho$  is the closed convex envelope of  $\phi_0 \mathbf{1}_{\mathcal{P}_0}$  in  $L^\infty(\Omega, \mathcal{F}, \mu)$ .*

It will be useful in Section 3.5 to know when the supremum in Equation (3.2.2) is achieved. This question is answered by the following proposition from the theory of Fenchel duality (e.g., Proposition 3, p. 203 in [6]).

**Proposition 3.2.8.** *Let  $X \in \mathcal{H}$  be a position in the domain of the convex risk measure  $\rho$ . Then the following are equivalent:*

- (a)  $\rho(X) = E_{P^*}[X] - \phi(P^*)$  for  $P^* \in \mathcal{P}$ .
- (b)  $P^*$  belongs to the subdifferential of  $\rho$  at  $X \in \mathcal{H}$ ,  $\partial\rho(X)$ .

*Proof.* For  $P^* \in \mathcal{P}$ ,

$$\begin{aligned} E_{P^*}[X] - \phi(P^*) &= \sup_{P \in \mathcal{P}} (E_P[X] - \phi(P)) \\ \Leftrightarrow E_{P^*}[X] - \phi(P^*) &\geq E_P[X] - \phi(P), \quad \forall P \in \mathcal{P} \\ \Leftrightarrow P^* &\in \partial\rho(X). \end{aligned}$$

□

When  $\mathcal{H} = L^\infty(\Omega, \mathcal{F}, \mu)$ , there is an analog of the previous theorem, which yields a representation involving expectations with respect to the continuous dual space of  $L^\infty(\Omega, \mathcal{F}, \mu)$ , which is  $\mathbf{ba}(\Omega, \mathcal{F}, \mu)$ , the Banach space of signed bounded finitely additive measures on  $(\Omega, \mathcal{F})$  that are absolutely continuous with respect to  $\mu$  [1, 21, 25]. If  $\Omega$  is countable and  $\mathcal{F} = 2^\Omega$ , finitely additive measures are  $\sigma$ -additive. Therefore, convex risk measures on  $L^\infty$  can be represented as worst-case penalized expectations with

respect to probability measures in  $L^1$ . This result is an easy extension of Theorem 2.3 in [21].

There is also a representation theorem on the position space  $\mathcal{H} = L^\infty(\Omega, \mathcal{F}, \mu)$  that involves test measures in  $L^1(\Omega, \mathcal{F}, \mu)$ . However, for this stronger result, we need to replace Condition (5) in the definition of a convex risk measure with the stronger Fatou property. The Fatou property can be defined by three equivalent conditions (as shown in Theorem 6 in [25]).

**Definition 3.2.9.** *A translation invariant mapping  $\rho : L^\infty(\Omega, \mathcal{F}, \mu) \rightarrow \mathbb{R} \cup \{+\infty\}$  is said to satisfy the Fatou property if any of the following three conditions is satisfied:*

- (a)  $\{X \in L^\infty(\Omega, \mathcal{F}, \mu) \mid \rho(X) \leq 0\}$  is  $\sigma(L^\infty(\Omega, \mathcal{F}, \mu), L^1(\Omega, \mathcal{F}, \mu))$ -closed.
- (b)  $\rho(X) \leq \liminf \rho(X_n)$  for any sequence of functions  $(X_n)$  on  $L^\infty(\Omega, \mathcal{F}, \mu)$ , uniformly bounded by 1 and converging to  $X$  in  $\mu$ -probability.
- (c)  $\rho(X_n) \rightarrow \rho(X)$  for any uniformly bounded sequence  $X_n$  that decreases to  $X$   $\mu$ -almost surely

**Remark 3.2.10.** *The Fatou property is a refinement of condition (5) in the definition of a convex risk measure on  $L^\infty(\Omega, \mathcal{F}, \mu)$ . If  $\rho$  has the Fatou property, then  $\{X \in L^\infty(\Omega, \mathcal{F}, \mu) \mid \rho(X) \leq 0\}$  is weakly closed and a fortiori  $l_\infty$ -closed [1].*

Now, we can state an analog of Theorem 3.2.6 on  $L^\infty(\Omega, \mathcal{F}, \mu)$ , namely Theorem 6 in [25].

**Theorem 3.2.11.** *Let  $\rho$  be a convex risk measure on  $L^\infty(\Omega, \mathcal{F}, \mu)$ . Assume that  $\rho$  verifies the Fatou property. Then there exists a  $l_1$ -closed, convex set of probability measures  $\mathcal{P} \subset L^1(\Omega, \mathcal{F}, \mu)$  and a penalty function  $\phi$  from  $\mathcal{P}$  to  $\mathbb{R}$  such that for all  $X \in L^\infty(\Omega, \mathcal{F}, \mu)$ ,*

$$\rho(X) = \sup_{P \in \mathcal{P}} (E_P[X] - \phi(P)). \quad (3.2.3)$$

Moreover, the penalty function can be chosen as the Fenchel conjugate of the risk measure:

$$\phi(P) = \sup_{X \in \mathcal{H}} (E_P[X] - \rho(X)).$$

In addition, if  $\rho$  is coherent, we can have  $\phi = 0$ .

Let us illustrate the theorem with the convex risk measure  $\rho(X) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma X))$  on  $L^\infty(\Omega, \mathcal{F}, \mu)$ . By the monotone convergence theorem,  $\rho$  has the Fatou property. Hence,  $\rho$  satisfies all the conditions of the theorem, and thus, can be represented as a worst-case penalized expectation.

Indeed, it is well-known ([25], p. 441) that

$$\log (E_\mu[\exp(X)]) = \sup_{P \in \mathcal{P}} (E_P[X] - D(P; \mu)),$$

where the set of test probability measures is  $\mathcal{P} = L^1(\Omega, \mathcal{F}, \mu)$  and the penalty function  $\phi$  is the divergence  $D(P; \mu) = \int P \log P d\mu \geq 0$ . Consequently, we have

$$\rho(X) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma X)) = \sup_{P \in \mathcal{P}} \left( E_P[X] - \frac{1}{\gamma} D(P; \mu) \right),$$

as predicted by Theorem 3.2.11.

### 3.3 Risk-averse control of dynamical systems

In this section, we describe a discrete-time model for a controlled stochastic dynamical system, which includes MDPs as a special case and we comment on our model choice. Then, we define dynamically consistent multi-period risk measures, and introduce the new notion of Markovian risk. We conclude with the definition of a sequential Markov game associated with a risk measure.

### 3.3.1 Model description and notation

#### Uncertainty description

Let  $\mathcal{X}$  and  $\mathcal{A}$  be a finite state and action space, respectively. The state space  $\mathcal{X}$  not only describes the states of the dynamical system, but it can also capture information about the risk perception of the controller. Defining a state space incorporating the latter will be critical when we will consider Markovian risk. Consider the sample space  $\Omega = \mathcal{X} \times \left[ \mathbb{R}^{\mathcal{X}} \times (\mathcal{X} \times \mathbb{R})^{\mathcal{X}\mathcal{A}} \right]^T$ , where  $T$  is the time horizon length ( $1 \leq T \leq +\infty$ ). A sample point  $\omega \in \Omega$  will be thought of as a sequence realized in successive steps  $\omega = (s_1, (r_{1,i})_i, (\nu_{1,i,a}, q_{1,i,a})_{(i,a)}, (r_{2,i})_i, (\nu_{2,i,a}, q_{2,i,a})_{(i,a)}, \dots)$  with the following interpretation:  $s_1$  is the initial state of the system at  $t = 1$ ; for  $t \geq 1$ ,  $r_{t,i} \in \mathbb{R}$  is used to randomize the action choice at the state  $i \in \mathcal{X}$  at time  $t$ ,  $\nu_{t,i,a} \in \mathcal{X}$  and  $q_{t,i,a} \in \mathbb{R}$  are, respectively, the new state at time  $t + 1$  and the associated transition cost out of the state-action pair  $(i, a) \in \mathcal{X}\mathcal{A}$ .

For  $\omega \in \Omega$  and  $t \geq 1$ , we denote the partial histories of length  $t$  by

$$\omega_t = (s_1, (r_{1,i})_i, (\nu_{1,i,a}, q_{1,i,a})_{(i,a)}, \dots, (r_{t,i})_i)$$

and

$$\omega_t^+ = ((r_{1,i})_i, (\nu_{1,i,a}, q_{1,i,a})_{(i,a)}, \dots, (r_{t,i})_i, (\nu_{t,i,a}, q_{t,i,a})_{(i,a)}).$$

We also let  $\omega_0^+ = s_1$ . Let  $\Omega' = \{\omega_t, \omega_{t-1}^+ | \omega \in \Omega, t \geq 1\}$  be the set of partial histories and  $\Omega_t = \{\omega_{t-1}^+ | \omega \in \Omega\}$ . If  $\omega' \in \Omega'$  is of the form  $\omega_t$  or  $\omega_t^+$  for  $\omega \in \Omega$ , we will write  $\omega' \preceq \omega$ . For  $\omega' \in \Omega'$ , let  $F(\omega') = \{\omega \in \Omega | \omega' \preceq \omega\}$  and  $F(\emptyset) = \Omega$ .

Define  $\mathcal{F} = \mathcal{F}(\emptyset)$  the  $\sigma$ -field on  $\Omega$  generated by the sets  $F(\omega'')$ ,  $\omega'' \in \Omega'$ . Similarly let  $\mathcal{F}(\omega')$  be the  $\sigma$ -field on  $F(\omega')$  generated by  $F(\omega'')$ ,  $\omega' \preceq \omega''$ ,  $\omega'' \in \Omega'$ . The sub- $\sigma$ -field  $\mathcal{F}^s(\omega') \subset \mathcal{F}(\omega')$  contains the events that are decidable as soon as the next step after  $\omega'$  occurs. More precisely,  $\mathcal{F}^s(\omega_t)$  is the sub- $\sigma$ -field of  $\mathcal{F}(\omega_t)$  generated by the sets  $F(\omega_t^+)$ . Similarly,  $\mathcal{F}^s(\omega_{t-1}^+)$  is the sub- $\sigma$ -field of  $\mathcal{F}(\omega_{t-1}^+)$  generated by the set  $F(\omega_t)$ .

Let  $\mu = \mu_{|\emptyset}$  be a reference probability measure on the measurable space  $(\Omega, \mathcal{F})$ ,

and let  $\mu_{|\omega'}$  be a regular conditional probability law of  $\mu$  on  $(F(\omega'), \mathcal{F}(\omega'))$  given  $\omega' \in \Omega'$  (see e.g. [16], p.430). Random variables will be denoted by upper case letters (e.g.,  $S_1, R_{t,i}, N_{t,i,a}, Q_{t,i,a}$ ), whereas realizations of a random variable (for an implicit  $\omega$ ) will be denoted with lower case letters (e.g.,  $s_1, r_{t,i}, \nu_{t,i,a}, q_{t,i,a}$ ).

The random numbers  $R_{t,i}$  allow control randomization so that the decision maker has more power, but we will see in Sections 3.4 and 3.5 that the optimal decisions can be chosen deterministically under the appropriate conditions. We will assume that under  $\mu$  the random numbers  $r_{t,i}$  are generated independently for each  $t \geq 1$  and  $i \in \mathcal{X}$ , and independently from the state-cost process, and such that for all  $\delta \in \Delta = \{x \in \mathbb{R}^{\mathcal{A}} \mid x_a \geq 0, \sum_{a \in \mathcal{A}} x_a = 1\}$ , there exists a measurable function  $B : \mathbb{R} \rightarrow \mathcal{A}$  verifying for all  $a \in \mathcal{A}$ ,  $\mu(\{B(r) = a\}) = \delta_a$ . Intuitively, the last condition guarantees that one can sample any probability distribution of action over  $\mathcal{A}$  using the random number generator.

The role of the probability measure  $\mu$  is to describe which events are known to occur with zero probability a priori, but it need not refer to the “true probabilistic model” of the system of interest. For example, we can have the following prior knowledge encoded in the probability measure  $\mu$ .

- If  $\mu(\{\nu_{t,i,a} = j\}) = 0$ , then the transition at time  $t$  from the state-action pair  $(i, a)$  to the state  $j$  is not expected to happen.
- If  $\mu(\{q_{t,i,a} = g(t, i, a)\}) = 1$  for some function  $g$ , then the immediate cost  $q_{t,i,a}$  is a deterministic function of  $(t, i, a)$ .

Observe that we do not assume that  $\mu$  is stationary or that there is independence across steps, as it would be the case if  $\mu$  was a Markovian model. The probability measure  $\mu$  could even encode history-dependent restrictions. For example, if we let  $\mu(\{\nu_{t_1,i_1,a} = \nu_{t_2,i_2,a}, \forall t_1, t_2, i, a\}) = 1$ , then all the one-step transitions are the same with probability one. However, some choices of  $\mu$  might make certain properties of risk measures on  $L^1(\Omega, \mathcal{F}, \mu)$ , such as dynamic consistency or Markovianity, impossible to hold.

## System trajectories

A policy  $\pi$  is a sequence of measurable mappings  $(\pi_1, \pi_2, \dots)$  such that  $\pi_t : \Omega_t \times \mathbb{R} \rightarrow \mathcal{A}$ . A policy associates with  $\omega \in \Omega$  a state-action-cost trajectory defined recursively by

$$\begin{aligned} S_1^\pi(\omega) &= s_1, \\ A_t^\pi(\omega) &= \pi_t(\omega_t, R_{t, S_t^\pi(\omega)}), \\ Q_t^\pi(\omega) &= q_{t, S_t^\pi(\omega), A_t^\pi(\omega)}, \\ S_{t+1}^\pi(\omega) &= \nu_{t, S_t^\pi(\omega), A_t^\pi(\omega)}. \end{aligned} \tag{3.3.1}$$

It will be useful later to have, for all time  $t$ , states  $j$ , actions  $a$ , the notation  $T_{t,j,a}^\pi(\omega)$  for the system's trajectory defined by the above recursion when the system is initialized at time  $t$  in the state action pair  $(j, a)$  and the sequence of controls  $\pi = (\pi_{t+1}, \pi_{t+2}, \dots)$  is applied.

In each state  $i \in \mathcal{X}$ , there is a non-empty set  $\mathcal{A}_i$  of available actions. A policy  $\pi$  is *admissible* if for all  $\omega \in \Omega$ ,  $A_t^\pi(\omega) \in \mathcal{A}_{S_t^\pi(\omega)}$ . We use  $\Pi_{h,r}$  to denote the set of admissible “history-dependent” policies selecting only actions available at the current state and depending only on the observed state-action-cost trajectory and the last randomization variable, i.e., if  $\omega^1, \omega^2 \in \Omega$  are such that  $S_t^\pi(\omega^1) = S_t^\pi(\omega^2)$ ,  $A_t^\pi(\omega^1) = A_t^\pi(\omega^2)$ ,  $Q_t^\pi(\omega^1) = Q_t^\pi(\omega^2)$ , for  $t \leq \tau - 1$ , and  $S_\tau^\pi(\omega^1) = S_\tau^\pi(\omega^2)$ , and  $r_{\tau, S_\tau^\pi(\omega^1)}^1 = r_{\tau, S_\tau^\pi(\omega^2)}^2$ , then  $A_\tau^\pi(\omega^1) = A_\tau^\pi(\omega^2)$ . The policy  $\pi \in \Pi_{h,r}$  can be restricted to depend on the history only through the current state, time index, and randomization variable (then the policy  $\pi$  belongs to the Markovian policy space  $\Pi_{m,r}$ ), or depend only on the current state and randomization variable (then the policy  $\pi$  is in the stationary Markovian policy space  $\Pi_{s,r}$ ). The sets of deterministic policies are denoted with the index  $d$  instead of  $r$  ( $\Pi_{h,d}$ ,  $\Pi_{m,d}$ , and  $\Pi_{s,d}$ ).

For a policy  $\pi \in \Pi_{h,r}$  and a discount factor  $\beta \in [0, 1]$ , we let

$$C_\tau^\pi = \liminf_{h \rightarrow T} \sum_{t=\tau}^h \beta^t Q_t^\pi,$$

which is the tail sample cost incurred by policy  $\pi$ . When  $\tau > T$ , we let  $C_\tau^\pi = 0$ . When the context is clear, we will write  $C^\pi$  instead of  $C_1^\pi$ . Also, we define the tail sample cost  $C_{t,j,a}^\pi$  (resp.  $C_{t,j}^\pi$ ) of a policy  $\pi$  initialized at time  $t$  in state-action pair  $(j, a)$  (resp. in state  $j$ ).

The sample costs are measurable random variables taking values in  $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ . If  $T < +\infty$ , then  $C_t^\pi$  is finite for all  $t$  and  $\omega \in \Omega$ . However, without additional assumptions, there might be  $\omega \in \Omega$  such that  $C_t^\pi$  is not finite when  $T = +\infty$ . The following assumption makes sure that the sample cost is integrable.

**Assumption 3.3.1.** *For all  $\pi \in \Pi_{m,r}$ ,  $t \geq 1$ ,  $j \in \mathcal{X}$ , and  $a \in \mathcal{A}$ , the sample costs  $C_t^\pi$ ,  $C_{t,j}^\pi$ , and  $C_{t,j,a}^\pi$  are in  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  for all  $\omega' \in \Omega' \cup \{\emptyset\}$ .*

Although there is one version of the tail sample cost in  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  per  $\omega' \in \Omega' \cup \{\emptyset\}$ , we denote all of them with the same notation, namely  $C_t^\pi$ .

**Example 3.3.2.**

In order to illustrate our model definition, let us consider a toy example of a dynamical system on a time horizon  $T = 2$  with two states  $\mathcal{X} = \{b, c\}$ , and  $\mathcal{A} = \{l, m\}$ .

A sample point  $\omega$  is any element of  $\Omega = \mathcal{X} \times \left[ \mathbb{R}^{\mathcal{X}} \times (\mathcal{X} \times \mathbb{R})^{\mathcal{X}\mathcal{A}} \right]^T$ . Let us pick one sample point  $\omega$  in  $\Omega$  and arrange its components in a matrix form, where the first two rows correspond to state  $b$  and the last two to state  $c$  and where the first row in these pairs of rows corresponds to action  $l$  and the other to action  $m$ ,

$$\bar{\omega} = \left( \begin{array}{c|c|c|c|c} s_1 & r_{1,b} & \nu_{1,b,l}, q_{1,b,l} & r_{2,b} & \nu_{2,b,l}, q_{2,b,l} \\ & & \nu_{1,b,m}, q_{1,b,m} & & \nu_{2,b,m}, q_{2,b,m} \\ \hline & r_{1,c} & \nu_{1,c,l}, q_{1,c,l} & r_{2,c} & \nu_{2,c,l}, q_{2,c,l} \\ & & \nu_{1,c,m}, q_{1,c,m} & & \nu_{2,c,m}, q_{2,c,m} \end{array} \right) = \left( \begin{array}{c|c|c|c|c} b & 2.4 & b, 1 & 9.34 & c, -5 \\ & & c, 0 & & b, 6 \\ \hline & 6 & c, -2 & 18.4 & b, 25 \\ & & c, 7 & & c, 7 \end{array} \right).$$

This specific sample point specifies that the initial state  $s_1 = b$ , that given that the system in state  $b$  and that action  $m$  is selected at time 1, then the new state would be  $\nu_{1,b,m} = c$  and the associated transition cost would be  $q_{1,b,m} = 0$ . On the other

hand, if the system is in state  $c$  and action  $l$  is selected at time 1, then the next state would be  $\nu_{1,c,l} = c$  and the associate cost would be  $q_{1,c,l} = -2$ .

If the controller follows the policy  $\pi$  that always chooses action  $m$ , then the uncertainty realization  $\bar{\omega}$  induces the state trajectory  $b, c, c$  and the total cost is  $C^\pi = 0 + 7 = 7$ . The information encoded in the sample point  $\bar{\omega}$  that is relevant to the system trajectory in this example is underlined

$$\bar{\omega} = \left( \begin{array}{c|c|c|c|c} \underline{b} & \underline{2.4} & b, 1 & 9.34 & c, -5 \\ & & \underline{c, 0} & & b, 6 \\ \hline & 6 & c, -2 & \underline{18.4} & b, 25 \\ & & c, 7 & & \underline{c, 7} \end{array} \right).$$

Recall that the support of the probability measure  $\mu$  on  $(\Omega, \mathcal{F})$  encodes the constraints of the system that the controller knows a priori. For example, if the controller knows a priori that

- the system always goes immediately from state  $c$  to state  $b$  under all actions,
- that the system stays in state  $b$  when it is in state  $b$  and control  $l$  (for loop) is chosen, and
- the immediate costs are either zero or one,

then we can let

$$\begin{aligned} \mu(\{S_1 = s_1, N_{t,i,a} = \nu_{t,i,a}, Q_{t,i,a} = q_{t,i,a}, t \geq 1, i \in \mathcal{X}, a \in \mathcal{A}\}) \\ = \mathbb{P}_{S_1}(s_1) \prod_{t,i,a} \mathbb{P}_{N_{t,i,a}}(\nu_{t,i,a}) \mathbb{P}_{Q_{t,i,a}}(q_{t,i,a}), \end{aligned}$$

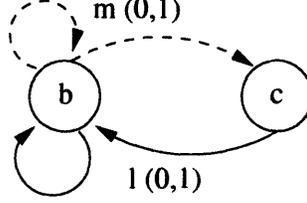


Figure 3-1: Graphical representation of a single-step transition for a simple example of dynamical system.

where for all  $t, i, a$

$$\begin{aligned} \mathbb{P}_{S_1}(S_1 = b) &= \mathbb{P}_{S_1}(S_1 = c) = 1/2, \\ \mathbb{P}_{N_{t,b,m}}(N_{t,b,m} = b) &= \mathbb{P}_{N_{t,b,m}}(N_{t,b,m} = c) = 1/2, \\ \mathbb{P}_{N_{t,b,l}}(N_{t,b,l} = b) &= 1, \\ \mathbb{P}_{N_{t,c,m}}(N_{t,c,m} = b) &= \mathbb{P}_{N_{t,c,l}}(N_{t,c,l} = b) = 1, \\ \mathbb{P}_{Q_{t,i,b}}(Q_{t,i,b} = 0) &= \mathbb{P}_{Q_{t,i,b}}(Q_{t,i,b} = 1) = 1/2. \end{aligned}$$

This information is best summarized and represented by Figure 3-1.

### 3.3.2 Comments on our model choice

In this subsection, we illustrate the modeling richness of our framework. It might seem unnecessary to define such a general and high-dimensional uncertainty  $\omega \in \Omega$ . There is one separate component of  $\omega$  for each possible outcome of interest, in particular the state  $\nu_{t,i,a} \in \mathcal{X}$  following each state-action pair  $(i, a) \in \mathcal{XA}$ , and the associated immediate cost  $q_{t,i,a} \in \mathbb{R}$ .

In our class of models, we have the models in which the same low-dimensional uncertain outcome  $w_t$  influences all states:

$$\begin{aligned} A_t^\pi &= \pi_t(S_t), \\ S_{t+1}^\pi &= f(S_t, A_t, w_t), \\ Q_t^\pi &= g(S_t, A_t, w_t). \end{aligned}$$

This model is simpler and, therefore, seems more appealing than our general framework. But, we will see in Sections 3.4 and 3.5 that our more general model yields tractable risk-sensitive decision problems. Furthermore, when the uncertain outcome  $w_t$  has too low dimension, it cannot capture some important preferences in face of uncertainty and ambiguity, which our general model allows.

For example, consider the following situation inspired from Ellsberg's paradox [23]. There are two coins, respectively called coin  $A$  and  $B$ , and a player can bet on the outcome of a flip of one of the two coins. If his guess is right, he gets \$100, and \$0 otherwise. It might seem appropriate to have  $w_1$  be 0 if the flipped coin (either coin  $A$  or  $B$ ) turns up head and 1 otherwise, but we will see next that this model cannot capture a different uncertainty aversion for each coin, in contrast with the more general uncertainty where  $w_A$  and  $w_B$  is the uncertain results of coin  $A$  and  $B$ , respectively.

Assume that the player believes that coin  $A$  is fair and will in heads or tails with equal probability, whereas coin  $B$  is thought to give always the same result but the player does not know whether it is head or tail. If the player had to choose a subjective probability for the outcome of coin  $B$ , it needs to be the probability giving half and half for head and tail by symmetry. If the player does not like "ambiguity" as most surveyed people in [23], he will prefer to bet on the outcome of coin  $A$ . This preference cannot be captured by decision theory based on expected utility as Ellsberg demonstrated. Neither can it be captured by risk aversion as defined in Section 3.2 if the uncertain outcome  $w_1 \in \{0, 1\}$  represents the uncertain result of flipping the selected coin. Indeed, since both coins  $A$  and  $B$  are described by the same uncertainty, they cannot be discriminated. However, if the uncertain outcome is  $(w_A, w_B)$ , where  $w_A \in \{0, 1\}$  and  $w_B \in \{0, 1\}$  correspond respectively to the realization of coin  $A$  and  $B$ , then this model for uncertainty allows to differentiate both coins and thus capture ambiguity aversion. For example, consider the probability measure  $P_A^0$  for the uncertain outcome  $w_A$  defined by  $P_A(0) = P_A(1) = 1/2$ , and the two probability measures  $P_B^1, P_B^2$  for the uncertain outcome  $w_B$  of coin  $B$  defined by  $P_B^1(0) = P_B^2(1) = 0.6$  and  $P_B^1(1) = P_B^2(0) = 0.4$ . Then, the coherent risk measure associated with test

probability measures  $P = P_A \times P_B$  on  $(w_A, w_B)$  in the set  $\mathcal{P} = \{P_A^0 \times P_B^1, P_A^1 \times P_B^0\}$  gives a risk of  $-\$50$  for betting on coin  $A$  and  $-\$40$  for coin  $B$ .

The advantage of using simple uncertainty model is that some technical difficulties can be avoided. For example, if the controller does not use randomization, if the immediate costs  $Q_{t,i,a}$  take values in a finite set, and if the time horizon is finite, then the set of uncertainty  $\Omega$  can be chosen finite. In this case, the functional analytic issues disappear. (However, a finite  $\Omega$  cannot model infinite time horizon problems.)

### 3.3.3 Properties of multi-period risk

In this subsection, we define dynamically consistent multi-period risk measures in the control setting described previously. We introduce the new notion of Markovian multi-period risk measures and establish some useful properties of dynamically consistent and Markovian risk measures.

#### Multi-period risk measures

**Definition 3.3.3.** *A multi-period risk measure on the space  $L^1(\Omega, \mathcal{F}, \mu)$  is a mapping that assigns to each partial history  $\omega' \in \Omega' \cup \{\emptyset\}$  a risk measure on  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$ , denoted  $\rho(\cdot|\omega')$ . We let  $\rho(\cdot) = \rho(\cdot|\emptyset)$ .*

**Remark 3.3.4.** *The risk measure is defined on a space of “positions” containing many more random variables than the sample costs  $C_t^\pi$ , in order to invoke the results on convex risk measures of Section 3.2, (whereas the set of sample costs is not even convex). Nonetheless, this requirement is not restrictive since a decision maker should have a well-defined risk measure, which really represents his preferences in the face of uncertainty, not only on the space of possible sample costs  $C_t^\pi$ , but for all conceivable positions.*

**Definition 3.3.5.** *A multi-period risk measure  $\rho$  is coherent (resp. convex) if for all  $\omega' \in \Omega' \cup \{\emptyset\}$ ,  $\rho(\cdot|\omega')$  is coherent (resp. convex) on  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$ .*

### Dynamically consistent multi-period risk measure

Fix a partial history  $\omega' \in \Omega'$ . Let  $\mathcal{E}(\omega')$  be the set of possible “single-step outcomes” following  $\omega'$ . Specifically,  $\mathcal{E}(\omega') = \mathbb{R}^{\mathcal{X}}$  if  $\omega' = \omega_{t-1}^+$  and  $\mathcal{E}(\omega') = (\mathcal{X} \times \mathbb{R})^{\mathcal{X}\mathcal{A}}$  if  $\omega' = \omega_t$ . For any  $\omega \succeq \omega'$ , let  $V_{\omega'}(\omega)$  be the unique “immediate continuation”  $v \in \mathcal{E}(\omega')$  such that  $\omega \succeq \omega'v$ . If  $\omega'$  is of the form  $\omega' = \omega_{t-1}^+$ , then the immediate continuation is  $V_{\omega'}(\omega) = (R_{t,i}(\omega))_i$ , and if  $\omega' = \omega_t$ , then the immediate continuation is  $V_{\omega'}(\omega) = (N_{t,i,a}(\omega), Q_{t,i,a}(\omega))_{(i,a)}$ . An element of  $L^1(F(\omega'), \mathcal{F}^s(\omega'), \mu_{|\omega'})$  will be called a *single-step position*, since intuitively it is a position whose payoff is determined by the immediate continuation of  $\omega'$ .

For a given  $X \in L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  and a multi-period risk measure  $\rho$ , consider the extended real-valued function on  $F(\omega')$ ,  $\rho(X|\omega' \cdot)$ , that maps  $\omega$  to  $\rho(X|\omega'v)$ , where  $v = V_{\omega'}(\omega)$ . It will be convenient to denote this random variable by  $\rho(X|\omega'V)$ , where the randomness enters through  $V$ .

**Definition 3.3.6.** *A multi-period risk measure is dynamically consistent if*

(a) *For all  $\omega' \in \Omega'$  and all*

$$X, Y \in \left[ L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'}) \cap \left( \bigcap_{v \in \mathcal{E}(\omega')} L^1(F(\omega'v), \mathcal{F}(\omega'v), \mu_{|\omega'v}) \right) \right]$$

*the following implication holds*

$$[\forall v \in \mathcal{E}(\omega'), \rho(X|\omega'v) \geq \rho(Y|\omega'v)] \Rightarrow \rho(X|\omega') \geq \rho(Y|\omega').$$

(b) *For all  $\omega' \in \Omega'$ , all  $\pi \in \Pi_{m,r}$  and  $t \geq 1$ ,  $\rho(C_t^\pi|\omega' \cdot)$  is a single-step position, i.e., it belongs to  $L^1(F(\omega'), \mathcal{F}^s(\omega'), \mu_{|\omega'})$ .*

This definition is the same as the usual definition of a dynamically consistent risk measure when the sample space is finite (e.g., [72]) but with the necessary technical assumptions to cope with general sample spaces.

The restriction of the risk measure  $\rho(\cdot|\omega')$ , which is defined on all of  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$ ,

to the set of single-step positions  $L^1(\Omega, \mathcal{F}^s(\omega'), \mu_{|\omega'})$  will be denoted by  $\rho_{\omega'}$ . If  $\rho(\cdot|\omega')$  is convex (resp. coherent), it follows easily that  $\rho_{\omega'}$  is convex (resp. coherent).

**Lemma 3.3.7.** *If  $\rho$  is a dynamically consistent convex multi-period risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ , then for all policies  $\pi \in \Pi_{m,r}$ , all  $t \geq 1$ , and all partial histories  $\omega' \in \Omega'$ , we have  $\rho(C_t^\pi|\omega') = \rho_{\omega'}[\rho(C_t^\pi|\omega' \cdot)]$ .*

*Proof.* Fix  $\pi \in \Pi_{m,r}$ ,  $t \geq 1$  and  $\omega' \in \Omega'$ . Consider the single-step position  $\tilde{X}$  defined by  $\tilde{X} = \rho(C_t^\pi|\omega' \cdot)$ . (Note that  $\tilde{X}$  is indeed in  $L^1(F(\omega'), \mathcal{F}^s(\omega'), \mu_{|\omega'})$  thanks to property (b) in Definition 3.3.6.) Since the payoff of  $\tilde{X}$  is determined after the next single-step outcome  $v$ , we have  $\rho(\tilde{X}|\omega'v) = \tilde{X}(\omega'v) = \rho(C_t^\pi|\omega'v)$  for all  $v \in \mathcal{E}(\omega')$ , where the first equality follows from the translation invariance of  $\rho(\cdot|\omega'v)$ . Consequently, property (a) in Definition 3.3.6 yields that  $\rho(\tilde{X}|\omega') = \rho(C_t^\pi|\omega')$ . Using the definition of  $\tilde{X}$ ,  $\rho(C_t^\pi|\omega') = \rho[\rho(C_t^\pi|\omega' \cdot)|\omega'] = \rho_{\omega'}[\rho(C_t^\pi|\omega' \cdot)]$ .  $\square$

Now we make a distinction between the risk associated with the controller's randomization via  $(R_{t,i})_i$ , and the system uncertainty captured by  $(N_{t,i,a}, Q_{t,i,a})_{(i,a)}$ . We will essentially assume that the controller is "risk neutral" (or has no uncertainty about the randomness of the random number generator) with respect to the  $(R_{t,i})_i$ .

**Assumption 3.3.8.** *For all  $\omega \in \Omega$ ,  $t \geq 1$ , and for all single-step positions  $X^s \in L^1(F(\omega_{t-1}^+, \mathcal{F}^s(\omega_{t-1}^+), \mu_{|\omega_{t-1}^+})$ , we have  $\rho(X^s|\omega_{t-1}^+) = E_{\mu_{|\omega_{t-1}^+}}[X^s]$ .*

Under this assumption, the following notations will be handy. For  $\pi \in \Pi_{m,r}$ , denote  $\pi_t(a|\omega_{t-1}^+) = \mu_{|\omega_{t-1}^+}(\{A_t^\pi = a\})$ . Let  $\Delta_j = \{x \in \mathbb{R}^A \mid x_a \geq 0; \sum_{a \in A_j} x_a = 1\}$ .

### Markovian risk measures

As time goes and new information is revealed, two drivers of decision-making change:

1. the unrealized or unobserved uncertainty in the system,
2. the risk-sensitivity of the decision maker.

Intuitively, a multi-period risk measure is Markovian if its conditional risk measures depend on the system history only through the current state-action pair.

Let  $\mathcal{G}_{t,j,a}$  be the subset of positions that are functions of the tail trajectory of some Markovian policy initialized in state-action pair  $(j, a)$  at time  $t$ , i.e.,

$$\mathcal{G}_{t,j,a} = \{f(T_{t,j,a}^\pi), \pi \in \Pi_{m,r}, f \in L^1\}.$$

Similarly, let

$$\mathcal{G}_{t,j} = \{f(T_{t,j}^\pi), \pi \in \Pi_{m,r}, f \in L^1\}.$$

**Definition 3.3.9.** *A multi-period risk measure  $\rho$  is Markovian if for all times  $t \geq 1$ , states  $j \in \mathcal{X}$ , and actions  $a \in \mathcal{A}$ , the following hold:*

- (a) *For all  $g \in \mathcal{G}_{t,j}$ ,  $\rho(g|\omega_{t-1}^+)$  does not depend on  $\omega_{t-1}^+$ . For any  $\omega_{t-1}^+$ , we denote this risk mapping on  $\mathcal{G}_{t,j}$  by  $\rho(\cdot|t, j) = \rho(\cdot|\omega_{t-1}^+)$ .*
- (b) *For all  $g \in \mathcal{G}_{t,j,a}$ ,  $\rho(g|\omega_t)$  does not depend on  $\omega_t$ . For any  $\omega_t$ , we denote this risk mapping on  $\mathcal{G}_{t,j,a}$  by  $\rho(\cdot|t, j, a) = \rho(\cdot|\omega_t)$ .*

For example,  $C_{t,j}^\pi \in \mathcal{G}_{t,j}$  and  $C_{t,j,a}^\pi \in \mathcal{G}_{t,j,a}$  for all Markovian policies  $\pi \in \Pi_{m,r}$ . Hence, the above definition justifies writing  $\rho(C_{t,j}^\pi|t, j)$  and  $\rho(C_{t,j,a}^\pi|t, j, a)$ .

If  $\rho$  is Markovian, the respective notations  $\rho_{(t,j)}$  and  $\rho_{(t,j,a)}$  for the risk measures  $\rho(\cdot|t, j)$  and  $\rho(\cdot|t, j, a)$  restricted to the single-step positions in  $\mathcal{G}_{t,j}$  and  $\mathcal{G}_{t,j,a}$  are well-defined. Observe also that the risk measures  $\rho(\cdot|t, j)$  and  $\rho(\cdot|t, j, a)$  are fixed for all positions in  $\mathcal{G}_{t,j}$  and  $\mathcal{G}_{t,j,a}$ .

The assumption that a risk measure is Markovian is not very restrictive if we are willing to have a large state space. Indeed, we can add to the state space  $\mathcal{X}$  the information that is necessary so that the decision maker's risk is Markovian with respect to  $\mathcal{X}$ . The extreme case is to have one state per partial history. In this extreme case, some previous works (e.g., [72, 76, 4]) derived equations relating the  $\rho(C_t^\pi|\omega')$  for different  $\omega'$  and  $t$ , for dynamically consistent convex risk measures over a finite time horizon  $T$ . These equations can be solved recursively to evaluate the risk of a policy or to find a Markovian policy with minimum risk. However, there are as many equations as partial histories  $\omega' \in \Omega'$ . Accordingly, the computational cost grows exponentially with the time horizon  $T$ . Hence, the problem becomes quickly

intractable from a computational standpoint and nothing was said about the infinite horizon case.

The main motivation for defining Markovian risk measures is to be able to minimize risk efficiently for problems with large state-action space and long, even infinite time horizon. This will be accomplished in Sections 3.4 and 3.5. As a preview of the results of these sections and to compare with the aforementioned works, when the time horizon  $T$  is finite, a Markovian policy that minimizes a Markovian dynamically consistent convex risk measure can be computed by carrying out only  $O(|\mathcal{X}|T)$ , instead of  $O(|\mathcal{X}|^T)$ .

**Lemma 3.3.10.** *Let  $\rho$  be a convex dynamically consistent Markovian multi-period risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ . Assume that Assumption 3.3.8 is satisfied. For any policy  $\pi \in \Pi_{m,r}$ , we have the identities*

$$\rho(C_{t,j}^\pi | t, j) = \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho(C_{t,j,a}^\pi | t, j, a), \quad (3.3.2)$$

$$\rho(C_{t,j,a}^\pi | t, j, a) = \rho_{(t,j,a)} \left[ Q_{t,j,a} + \rho(C_{t+1,N_{t,j,a}}^\pi | t+1, N_{t,j,a}) \right]. \quad (3.3.3)$$

**Remark 3.3.11.** *It is important to notice that the risk mappings that appear in these identities do not depend on  $\pi$ , but only on  $t, j, a$ .*

*Proof.* Both identities stem from the combination of the definition of Markovian risk and Lemma 3.3.7 on dynamically consistent risk. In addition, the first identity relies on Assumption 3.3.8, which assumes that the controller is risk-neutral with respect to the random numbers  $(R_{t,i})_i$ .

(a) Since  $\rho$  is Markovian and  $C_{t,j}^\pi \in \mathcal{G}_{t,j}$ , we have  $\rho(C_{t,j}^\pi | t, j) = \rho(C_{t,j}^\pi | \omega_{t-1}^+)$  for an arbitrary  $\omega_{t-1}^+$ . By Lemma 3.3.7, there holds

$$\rho(C_{t,j}^\pi | \omega_{t-1}^+) = \rho_{\omega_{t-1}^+} [\rho(C_{t,j}^\pi | \omega_{t-1}^+(R_{t,i})_i)].$$

Assumption 3.3.8 applied to the single-step position  $\rho(C_{t,j}^\pi | \omega_{t-1}^+(R_{t,i})_i)$  yields

$$\rho(C_{t,j}^\pi | \omega_{t-1}^+) = E_{\mu_{\omega_{t-1}^+}} [\rho(C_{t,j}^\pi | \omega_{t-1}^+(R_{t,i})_i)].$$

When  $R_{t,j}$  is such that action  $a$  is selected under policy  $\pi$ , i.e.,  $\pi_t(\omega_{t-1}^+, (R_{t,i})_i) = a$ , then we have, for all  $\omega \geq \omega_{t-1}^+, (R_{t,i})_i$ ,  $C_{t,j}^\pi(\omega) = C_{t,j,a}^\pi(\omega)$ . Now, using the Markovian property of  $\rho$  with  $C_{t,j,a}^\pi$ , we can write

$$\rho(C_{t,j}^\pi | \omega_{t-1}^+, (R_{t,i})_i) = \rho(C_{t,j,a}^\pi | \omega_{t-1}^+, (R_{t,i})_i) = \rho(C_{t,j,a}^\pi | t, j, a).$$

This concludes the proof that

$$\rho(C_{t,j}^\pi | t, j) = \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho(C_{t,j,a}^\pi | t, j, a).$$

(b) Since the risk  $\rho$  is Markovian, we have  $\rho(C_{t,j,a}^\pi | t, j, a) = \rho(C_{t,j,a}^\pi | \omega_t)$  for an arbitrary  $\omega_t$ . By dynamic consistency of  $\rho$ , the following decomposition holds

$$\rho(C_{t,j,a}^\pi | \omega_t) = \rho_{\omega_t} \left[ \rho \left( C_{t,j,a}^\pi | \omega_t, (N_{t,i,b}, Q_{t,i,b})_{(b,i)} \right) \right].$$

By definition of the tail sample cost, and then using the Markov property of  $\rho$  with  $C_{t+1, N_{t,j,a}}^\pi \in \mathcal{G}_{t+1, N_{t,j,a}}$ , we have

$$\begin{aligned} \rho \left( C_{t,j,a}^\pi | \omega_t, (N_{t,i,b}, Q_{t,i,b})_{(b,i)} \right) &= Q_{t,j,a} + \rho \left( C_{t+1, N_{t,j,a}}^\pi | \omega_t, (N_{t,i,b}, Q_{t,i,b})_{(b,i)} \right) \\ &= Q_{t,j,a} + \rho \left( C_{t+1, N_{t,j,a}}^\pi | t+1, N_{t,j,a} \right). \end{aligned}$$

Combining this with the above decomposition, we obtain

$$\rho(C_{t,j,a}^\pi | \omega_t) = \rho_{\omega_t} \left[ Q_{t,j,a} + \rho \left( C_{t+1, N_{t,j,a}}^\pi | t+1, N_{t,j,a} \right) \right].$$

Invoking again the Markov property of  $\rho$  yields the second identity.  $\square$

### 3.3.4 A Markov game induced by a risk measure

In this subsection, we define a zero-sum sequential Markov game between nature and the controller, induced by a dynamically consistent Markovian convex risk measure  $\rho$  verifying Assumption 3.3.8. In the next two sections, we will see that minimizing  $\rho$

amounts to solving this game.

Given time  $t \geq 1$ , state  $j \in \mathcal{X}$ , and action  $a \in \mathcal{A}$ , consider the single-period risk mapping  $\rho_{(t,j,a)}$  defined on the single-step positions in  $\mathcal{G}_{t,j,a}$  let  $\mathcal{S}_{t,j,a}^\pi := \{\omega \in \Omega | S_t^\pi(\omega) = j, A_t^\pi(\omega) = a\}$ . Pick an arbitrary representative  $\bar{\omega}_t$ . Since  $\rho$  is Markovian,  $\rho(\cdot | t, j, a) = \rho(\cdot | \bar{\omega}_t)$  on  $\mathcal{G}_{t,j,a}$ . Observe that  $\rho_{\bar{\omega}_t}$  is a convex single-period risk measure on  $L^1(F(\bar{\omega}_t), \mathcal{F}^s(\bar{\omega}_t), \mu_{|\bar{\omega}_t})$ . By the representation theorem 3.2.6 applied to  $\rho_{\bar{\omega}_t}$ , there exists a  $l_\infty$ -closed convex set of probability measures  $\mathcal{P} \subset L^\infty(F(\bar{\omega}_t), \mathcal{F}^s(\bar{\omega}_t), \mu_{|\bar{\omega}_t})$  and a  $l_\infty$ -lower semicontinuous convex penalty function  $\phi$  such that

$$\rho_{\bar{\omega}_t}(X^s) = \sup_{P \in \mathcal{P}} (E_P[X^s] - \phi(P))$$

for all  $X^s \in L^1(F(\bar{\omega}_t), \mathcal{F}^s(\bar{\omega}_t), \mu_{|\bar{\omega}_t})$ .

For  $P \in \mathcal{P}$ , denote  $P^{t,j,a}$  the marginal of  $P$  on the space of random variables depending only on  $(Q_{t,j,a}, N_{t,j,a})$  and let  $\mathcal{P}^{t,j,a} = \{P_{(t,j,a)} | P \in \mathcal{P}\}$  be the collection of marginal distributions of  $\mathcal{P}$ .

For a probability measure  $P_{t,j,a} \in L^\infty(F(\bar{\omega}), \mathcal{F}^s(\bar{\omega}), \mu_{|\bar{\omega}})$  on  $(\mathbb{R}, \mathcal{X})$ , define the “penalty functions”

$$\bar{\phi}_{t,j,a}(P_{t,j,a}) = \inf_{\{P \in \mathcal{P} | P^{t,j,a} = P_{t,j,a}\}} \phi(P) \geq 0,$$

with the usual convention that  $\bar{\phi}_{t,j,a}(P_{t,j,a}) = +\infty$  if the set  $\{P \in \mathcal{P} | P^{t,j,a} = P_{t,j,a}\}$  is empty. Also let  $P(\cdot | t, j, a)$  be the marginal probability distribution on  $\mathcal{X}$  of  $P_{t,j,a}$  and let  $\bar{q}_{t,j,a}(P_{t,j,a}) := E_{P_{t,j,a}}[Q_{t,j,a}] \in \bar{\mathbb{R}}$ . When  $P_{t,j,a}$  is the marginal distribution of  $P \in L^\infty(F(\bar{\omega}), \mathcal{F}^s(\bar{\omega}), \mu_{|\bar{\omega}})$ , i.e.,  $P_{t,j,a} = P^{t,j,a}$ , the expectation  $\bar{q}_{t,j,a}(P_{t,j,a})$  is finite since for an appropriate Markovian policy  $\pi$  we have  $Q_{t,j,a} = C_t^\pi - C_{t+1}^\pi$ , which belongs to  $L^1(F(\bar{\omega}), \mathcal{F}^s(\bar{\omega}), \mu_{|\bar{\omega}})$  by Assumption 3.3.1.

Define  $\mathcal{P}^{t,j,a} = \{P^{t,j,a} | \bar{\phi}_{t,j,a}(P^{t,j,a}) < +\infty, P \in \mathcal{P}\}$ .  $\mathcal{P}^{t,j,a}$  is a non-empty set of “test probability measures”.

Now, we associate with  $\rho(\cdot | s_1)$  a zero-sum sequential Markov game on  $\mathcal{X} \cup \mathcal{XA}$  initialized at  $s_1 \in \mathcal{X}$  at time  $t = 1$ . The controller is the minimizer and nature the

maximizer. In state  $j \in \mathcal{X}$  at time  $t \geq 1$  the controller chooses  $\pi_t(\cdot|j) \in \Delta_j = \{x \in \mathbb{R}^{\mathcal{A}} \mid x_a \geq 0; \sum_{a \in \mathcal{A}_j} x_a = 1\}$ . The game state becomes  $(j, a)$ , for  $a \in \mathcal{A}_j$ , with probability  $\pi_t(a|j)$ . Then, nature chooses  $P_{t,j,a} \in \mathcal{P}^{t,j,a}$  and the controller pays to nature  $(\bar{q}_{t,j,a}(P_{t,j,a}) - \bar{\phi}_{t,j,a}(P_{t,j,a}))$ . The game continues for  $t = 1, \dots, T$ . At time  $T$ , there is no final cost.

For a fixed controller's policy  $\pi \in \Pi_{m,r}$ , this Markov game becomes an MDP on  $\mathcal{X}$ , controlled by nature, with the following characteristics. From the state  $j \in \mathcal{X}$  at time  $t$ , nature chooses  $(P_{t,j,a})_{a \in \mathcal{A}} \in \prod_{a \in \mathcal{A}} \mathcal{P}^{t,j,a}$ , the system moves to state  $k$  with probability  $\sum_{a \in \mathcal{A}} \pi_t(a|j) P(k|t, j, a)$  and nature receives the immediate reward  $\sum_{a \in \mathcal{A}} \pi_t(a|j) (\bar{q}_{t,j,a}(P_{t,j,a}) - \bar{\phi}_{t,j,a}(P_{t,j,a}))$ .

For  $\pi_1 = (\pi_1(\cdot|j))_{j \in \mathcal{X}} \in \prod_{j \in \mathcal{X}} \Delta_j$ , we can define the classical dynamic programming operators. For  $\beta \in (0, 1)$ , define the operator for  $\beta$ -discounted problem  $\mathbf{T}_{\pi_1}^\beta : \mathbb{R}^{\mathcal{X}} \rightarrow \mathbb{R}^{\mathcal{X}}$  by

$$\begin{aligned} (\mathbf{T}_{\pi_1}^\beta V)(j) &= \sum_{a \in \mathcal{A}} \pi_1(a|j) \rho_{(1,j,a)} (Q_{1,j,a} + \beta V(N_{1,j,a})) \\ &= \sum_{a \in \mathcal{A}} \pi_1(a|j) \sup_{P_{j,a} \in \mathcal{P}^{1,j,a}} \left[ \bar{q}_{1,j,a}(P_{j,a}) - \bar{\phi}_{1,j,a}(P_{j,a}) + \beta \sum_{k \in \mathcal{X}} P(k|1, j, a) V(k) \right]. \end{aligned}$$

In undiscounted cost problems, we assume that there is a special "termination" state  $\sigma \in \mathcal{X}$  in which the system is eventually absorbed at no cost so that the total cost is finite. For undiscounted cost problems, we will need the subspace  $\mathcal{V} := \{V \in \mathbb{R}^{\mathcal{X}} \mid V(\sigma) = 0\}$ , and the operator  $\mathbf{T}_{\pi_1} : \mathcal{V} \rightarrow \mathcal{V}$  defined by  $(\mathbf{T}_{\pi_1} V)(\sigma) = V(\sigma)$ , and for  $j \neq \sigma$ ,

$$\begin{aligned} (\mathbf{T}_{\pi_1} V)(j) &= \sum_{a \in \mathcal{A}} \pi_1(a|j) \rho_{(1,j,1)} (Q_{1,j,a} + V(N_{1,j,a})) \\ &= \sum_{a \in \mathcal{A}} \pi_1(a|j) \sup_{P_{j,a} \in \mathcal{P}^{1,j,a}} \left[ (\bar{q}_{1,j,a}(P_{j,a}) - \bar{\phi}_{1,j,a}(P_{j,a})) + \sum_{k \in \mathcal{X}} P(k|1, j, a) V(k) \right]. \end{aligned}$$

Define also the operators  $\mathbf{T}$  and  $\mathbf{T}^\beta$  respectively by  $\mathbf{T}V = \inf_{\pi_1 \in \prod_j \Delta_j} (\mathbf{T}_{\pi_1} V)$  and  $\mathbf{T}^\beta V = \inf_{\pi_1 \in \prod_j \Delta_j} (\mathbf{T}_{\pi_1}^\beta V)$ .

The following properties of these operators have classical proofs.

**Lemma 3.3.12.**

- (a) *The operators  $\mathbf{T}_\pi$ ,  $\mathbf{T}_\pi^\beta$ ,  $\mathbf{T}$ , and  $\mathbf{T}^\beta$  are monotonic, i.e., if  $V \leq V'$  component-wise, then  $\mathbf{T}_\pi V \leq \mathbf{T}_\pi V'$ , etc.*
- (b) *Let  $\mathbf{e} \in \mathbb{R}^\mathcal{X}$  be the vector with all components equal to one, except that  $\mathbf{e}(\sigma) = 0$ , and  $k \geq 0$ . For all  $V \in \mathcal{V}$ ,  $\mathbf{T}_\pi(V + k\mathbf{e}) \leq \mathbf{T}_\pi(V) + k\mathbf{e}$  and  $\mathbf{T}(V + k\mathbf{e}) \leq \mathbf{T}(V) + k\mathbf{e}$ .*
- (c) *The operators  $\mathbf{T}_\pi^\beta$  and  $\mathbf{T}^\beta$  are  $\beta$ -contractions under the sup-norm on  $\mathbb{R}^\mathcal{X}$ , i.e., for all  $V, V' \in \mathbb{R}^\mathcal{X}$ ,*

$$\|(\mathbf{T}^\beta V) - (\mathbf{T}^\beta V')\|_\infty \leq \beta \|V - V'\|_\infty.$$

- (d) *The operators  $\mathbf{T}_\pi$  and  $\mathbf{T}$  are 1-Lipschitz for the sup-norm on  $\mathbb{R}^\mathcal{X}$ , and thus continuous, i.e.,*

$$\|(\mathbf{T}V) - (\mathbf{T}V')\|_\infty \leq \|V - V'\|_\infty.$$

Now, we have the necessary framework and definitions to tackle, in Sections 3.4 and 3.5, the problem of finding a Markovian policy that minimizes a dynamically consistent Markovian convex multi-period risk measure of the sample cost.

### 3.4 Risk minimization over a finite horizon

In this section, we assume that the time horizon length  $T$  is finite. Without loss of generality, we let the discount factor  $\beta$  be equal to one. We will show that we can efficiently compute the minimum of a convex dynamically consistent Markovian risk of the sample cost  $C^\pi$  over the policy space  $\Pi_{m,r}$  and find a deterministic Markovian policy  $\pi \in \Pi_{m,d}$  that achieves the minimum risk. We will also establish that minimizing the risk amounts to solving the zero-sum sequential Markov game described in Subsection 3.3.4.

The next proposition shows how to evaluate efficiently the risk associated with a Markovian policy.

**Proposition 3.4.1.** *Let  $\rho$  be a convex dynamically consistent Markovian multi-period risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$  and fix  $\pi = (\pi_1, \dots, \pi_T) \in \Pi_{m,r}$ . If Assumptions 3.3.1 and 3.3.8 hold, then the backward recursion in  $\mathbb{R} \cup \{+\infty\}$*

$$V^\pi(T, j) = \sum_{a \in \mathcal{A}} \pi_T(a|j) \rho_{(T,j,a)}(Q_{T,j,a}), \quad j \in \mathcal{X},$$

$$V^\pi(t, j) = \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho_{(t,j,a)}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})], \quad j \in \mathcal{X}, \quad 1 \leq t \leq T-1,$$

yields  $\rho(C_1^\pi | s_1) = V^\pi(1, s_1)$  for all  $s_1 \in \mathcal{X}$ .

*Proof.* We will show by induction that  $V^\pi(t)$  is well-defined and

$$V^\pi(t, j) = \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho(C_{t,j,a}^\pi | t, j, a),$$

for all time  $t$ ,  $1 \leq t \leq T$ , and all states  $j \in \mathcal{X}$ .

At time  $T$ ,  $Q_{T,j,a} \in \mathcal{G}_{T,j,a}$  so that  $V^\pi(T, j)$  is well-defined for all  $j \in \mathcal{X}$ . Also, in light of the first identity of Lemma 3.3.10, it is clear that

$$V^\pi(T, j) = \sum_{a \in \mathcal{A}} \pi_T(a|j) \rho(C_{T,j,a}^\pi | T, j, a), \quad \forall j \in \mathcal{X}.$$

Assume  $V^\pi(\tau, j)$  verifies the claimed recursion for all  $\tau \geq t+1$  and  $j \in \mathcal{X}$ . Since  $Q_{t,j,a} + V^\pi(t+1, N_{t,j,a}) \in \mathcal{G}_{t,j,a}$ , the expression for  $V^\pi(t)$  is well-defined.

Combining the two identities of Lemma 3.3.10 yields for all  $j \in \mathcal{X}$

$$V^\pi(t, j) = \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho_{(t,j,a)}[Q_{t,j,a} + \rho(C_{t+1, N_{t,j,a}}^\pi | t+1, N_{t,j,a})].$$

Replacing  $\rho(C_{t+1, N_{t,j,a}}^\pi | t+1, N_{t,j,a})$  by  $V^\pi(t+1, N_{t,j,a})$  concludes the proof.  $\square$

When the multi-period risk measure  $\rho$  is dynamically consistent, Markovian, and convex, Proposition 3.4.1 shows that  $\rho$  is completely specified for the positions of the

form  $C_t^\pi$  through the single-period risk measures  $\rho_{(t,j,a)}$  for all  $t, j, a$ , instead of  $\rho(\cdot|\omega')$  for all  $\omega' \in \Omega'$ .

The recursion in the previous proposition is actually the Bellman recursion for the MDP induced by  $\rho$  (cf. Subsection 3.3.4) when the controller uses a fixed policy  $\pi \in \Pi_{m,r}$ , as we show now.

**Proposition 3.4.2.** *Let  $\hat{V}^\pi(t, j)$  the maximal reward secured by nature on the MDP induced by  $\rho$  and initialized at state  $j$  at time  $t$ . Then,  $V^\pi(t, j)$ , the risk of the tail sample cost under policy  $\pi \in \Pi_{m,r}$ , satisfies  $V^\pi(t, j) = \hat{V}^\pi(t, j)$  for all  $j \in \mathcal{X}$  and  $t = 1, \dots, T$ .*

*Proof.* From Proposition 3.4.1, we have

$$\begin{aligned} V^\pi(T, j) &= \sum_{a \in \mathcal{A}} \pi_T(a|j) \rho_{(T,j,a)}(Q_{T,j,a}), \quad j \in \mathcal{X}, \\ V^\pi(t, j) &= \sum_{a \in \mathcal{A}} \pi_t(a|j) \rho_{(t,j,a)}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})], \quad j \in \mathcal{X}, \quad 1 \leq t \leq T-1. \end{aligned}$$

In subsection 3.3.4, we associated with the risk mapping  $\rho_{(t,j,a)}$  an uncertainty set  $\mathcal{P}_{t,j,a}$  and penalty functions  $\bar{\phi}_{t,j,a}$ , for all  $t, j, a$ , such that

$$\rho_{(t,j,a)}(X) = \sup_{P \in \mathcal{P}^{t,j,a}} (E[X] - \bar{\phi}_{t,j,a}(P)),$$

for all single-step positions  $X \in \mathcal{G}_{t,j,a}$ . Consequently,

$$\rho_{(t,j,a)}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})] = \sup_{P_{t,j,a} \in \mathcal{P}^{t,j,a}} (E_{P_{t,j,a}}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})] - \bar{\phi}_{t,j,a}(P_{t,j,a})).$$

Hence, the recursion of Proposition 3.4.1 becomes

$$\begin{aligned} V^\pi(T, j) &= \sum_{a \in \mathcal{A}} \pi_T(a|j) \sup_{P_{T,j,a} \in \mathcal{P}^{T,j,a}} (E_{P_{T,j,a}}[Q_{T,j,a}] - \bar{\phi}_{t,j,a}(P_{T,j,a})), \\ V^\pi(t, j) &= \sum_{a \in \mathcal{A}} \pi_t(a|j) \sup_{P_{t,j,a} \in \mathcal{P}^{t,j,a}} (E_{P_{t,j,a}}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})] - \bar{\phi}_{t,j,a}(P_{t,j,a})), \end{aligned}$$

or equivalently,

$$V^\pi(T, j) = \sup_{(P_{T,j,a})_a \in \Pi_a \mathcal{P}^{T,j,a}} \sum_{a \in \mathcal{A}} \pi_T(a|j) [E_{P_{T,j,a}}[Q_{T,j,a}] - \bar{\phi}_{t,j,a}(P_{T,j,a})],$$

$$V^\pi(t, j) = \sup_{(P_{t,j,a})_a \in \Pi_a \mathcal{P}^{t,j,a}} \sum_{a \in \mathcal{A}} \pi_t(a|j) [E_{P_{t,j,a}}[Q_{t,j,a} + V^\pi(t+1, N_{t,j,a})] - \bar{\phi}_{t,j,a}(P_{t,j,a})].$$

These are the Bellman equations for the nature's MDP induced by  $\rho$ , when the controller's policy is  $\pi$ . Hence, the solution to this recursion is the maximal reward for nature in the MDP, and  $V^\pi(t, j) = \hat{V}^\pi(t, j)$ , for  $t = 1, \dots, T$ .  $\square$

**Theorem 3.4.3.** *Let  $\rho$  be a convex dynamically consistent Markovian multi-period risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ , and let Assumptions 3.3.1 and 3.3.8 hold. Consider the backward recursion:*

$$V(T, j) = \min_{a \in \mathcal{A}_j} \sup_{P_{T,j,a} \in \mathcal{P}^{T,j,a}} [\bar{q}_{T,j,a}(P_{T,j,a}) - \bar{\phi}_{T,j,a}(P_{T,j,a})], \quad j \in \mathcal{X}$$

$$V(t, j) = \min_{a \in \mathcal{A}_j} \sup_{P_{t,j,a} \in \mathcal{P}^{t,j,a}} \left[ \bar{q}_{t,j,a}(P_{t,j,a}) - \bar{\phi}_{t,j,a}(P_{t,j,a}) + \sum_{k \in \mathcal{X}} P(k|t, j, a) V(t+1, k) \right], \quad j \in \mathcal{X},$$

(3.4.1)

for  $t = 1, \dots, T-1$ .

Then  $\inf_{\pi \in \Pi_{m,r}} \rho(C_1^\pi | s_1) = V(1, s_1)$  for all  $s_1 \in \mathcal{X}$ .

Let  $a^*(t, j) \in \mathcal{A}_j$  be an action achieving the minimum in Equation (3.4.1). The deterministic policy  $\pi^*$  selecting action  $a^*(t, j)$  when the system is in state  $j$  at time  $t$ , achieves the minimum sample cost risk.

**Remark 3.4.4.** *This theorem rely heavily on the fact that the risk mapping  $\rho_{(t,j,a)}$  is fixed over  $\mathcal{G}_{t,j,a}$ , and thus does not depend on the policy  $\pi$ .*

*Proof.* First, an immediate induction using the recursion of Proposition 3.4.1 shows that  $V(t, j) \leq V^\pi(t, j)$  for all  $t, j$  and all  $\pi \in \Pi_{m,r}$ . Equivalently, we have  $V(t, j) \leq \rho(C_{t,j}^\pi | t, j)$  for all  $\pi \in \Pi_{m,r}$ .

Now, we can apply Proposition 3.4.1 to the policy  $\pi^* \in \Pi_{m,d}$  defined in the theorem. The recursion defining  $V$  and  $V^{\pi^*}$  are identical. Thus,  $V = V^{\pi^*}$ .

Finally, we conclude by observing

$$\rho(C^{\pi^*} | s_1) = V^{\pi^*}(1, s_1) \leq V^\pi(1, s_1) = \rho(C_1^\pi | s_1), \quad \forall \pi \in \Pi_{m,r},$$

where the first equality follows from Proposition 3.4.1 and the inequality is proved at the beginning of the proof.  $\square$

Observe that we made almost no assumption on the dynamical system itself; in particular, we did not have a probabilistic model for the system (except for the reference probability space  $(\Omega, \mathcal{F}, \mu)$ , which is needed essentially for technical reason). All the results presented so far, including the ones indicating that the system behaves like a stochastic Markov game from the controller's perspective, are obtained as a consequence of properties of the controller's risk aversion.

Similar to dynamic programming for MDP control, the controller's risk objective can be minimized over a "very large" policy space (exponential in the time horizon  $T$ ) by an algorithm with a complexity linear in  $T$ .

To the best of our knowledge, the only dynamic decision problem investigated in the risk literature is the pricing of a derivative product in an incomplete market [72]. In that paper, some assets are traded on a finite horizon by a decision maker who minimizes a coherent risk measure of his portfolio position. The asset prices take values in a finite set of price levels, with a distribution independent of the investor's decisions, and are observed at each trading period. The authors establish a recursion similar to the one in Theorem 3.4.3. However, without the Markovianity of the risk measure, they need to keep track of the conditional risk measure for all possible market histories, so that the complexity of their recursion grows exponentially with the horizon length. In contrast, if the risk measure has the Markov property with respect to the state space  $\mathcal{X}$ , Theorem 3.4.3 establishes that the solution can be found with only  $O(T|\mathcal{X}|)$  updates of the form  $V(i, t) = \min_{a \in \mathcal{A}_i} \rho_{(t,i,a)}[Q_{(i,a)}(t) + V(t+1, N_{t,i,a})]$ .

Before we deal with the more technical infinite-horizon problems in the next section, let us illustrate our system model and our results on an inventory management

problem.

### An example in inventory management

Let us illustrate our framework with an example. We consider a multi-period newsboy problem [66] with a spot market and no back-orders. Fix a finite time horizon  $T$  and the discount factor  $\beta = 1$ . Consider an inventory of a non-perishable commodity traded on a spot market. We assume that the inventory can hold a maximum of  $N$  units of the commodity. Hence, the inventory level is represented by the number  $i$  of units in the inventory, where  $i \in \mathcal{I} = \{0, 1, \dots, N\}$ , and  $N$  is a positive integer. Let  $h$  be the inventory holding cost per unit of stock at each period.

At each time period, there is an uncertain demand in  $\{0, \dots, D\}$  from customers, which drives the revenue of the firm. Customers pay a fixed price  $\bar{p}$  per unit. If a customer's order is not fulfilled immediately, the order is lost and a penalty of  $c$  is paid per missing unit.

The firm has access to a spot market from which it can replenish its inventory. We assume that the market price of the commodity takes values in a finite set  $\mathcal{X}_p = \{p_1, \dots, p_m\}$ . At the beginning of each period, the firm buys  $a$  units of the commodity ( $a \in \mathcal{A} = \{0, \dots, N\}$ ) at the current spot price in order to build inventory for the demand in the current and future periods. Then, the current demand is observed.

Let  $I(t)$  be the inventory at the beginning of the period,  $A(t)$  be the quantity purchased on the market at price  $P(t)$ , and  $D(t)$  be the demand at time  $t$  for each period indexed by  $t$ . In this problem, it is always better for the firm to meet the current demand as much as possible, rather than leaving demand unsatisfied and saving inventory for subsequent periods. Hence, we can write the dynamics of the inventory level as

$$I(t+1) = \max\{I(t) - D(t) + A(t), 0\},$$

and the associated immediate cost is

$$A(t)P(t) - \bar{p} \min(D(t), I(t)) + hI(t) + c \max(0, D(t) - I(t) - A(t)).$$

This inventory management problem not only entails the supply of an uncertain demand, but also the management of price uncertainty. It fits our general model for dynamical systems presented above, where we let  $\mathcal{X} = \mathcal{I} \times \mathcal{X}_p$  be the state space, and  $\mathcal{A}_{(n,p)} = \{0, \dots, N - n\}$  be the set of available actions in state  $(n, p) \in \mathcal{X}$ , where there is already  $n$  units in inventory.

In contrast with standard formulations of newsboy problems, we do not specify a probabilistic model for the demand or the market price. If the controller only uses deterministic policies, we can model this problem with a finite sample space  $\Omega$  since the only uncertainty is in the demand and market prices (and there are only finitely many of them to describe when  $T$ ,  $m$ ,  $D$ , and  $N$  are finite). In this case, the system model  $(\Omega, 2^\Omega, \mu)$  just specifies which events  $\omega \in \Omega$  are possible (i.e.,  $\mu(\omega) > 0$ ).

Let us assume that the decision maker has a dynamically consistent Markovian convex risk measure. Intuitively, this assumes that past demands and market prices bear no influence on the current risk assessment given the current market price and inventory level. For example, if the controller believes that the demand is independent period by period, and that the market price evolution can be described by a probability law that is Markovian with respect to  $\mathcal{X}$ , then its risk preference satisfies this assumption. In this case, we saw earlier in the present section that the risk preference of the decision maker can be represented as a certain Markov game, described in Subsection 3.3.4. In other words, the risk preference of the decision maker induces a worst-case probabilistic model of the system.

When the firm seeks to minimize its risk over all possible Markovian policies, the main challenge is to deal with the exponentially large policy space. Theorem 3.4.3 shows how to efficiently compute the minimal risk over the Markovian policies, and find an optimal deterministic policy by solving Shapley's equations for the associated zero-sum Markov game.

## 3.5 Risk minimization over an infinite horizon

In this section, we minimize a convex dynamically consistent Markovian risk measure of the sample cost of Markovian policies over an infinite horizon  $T = +\infty$ . We show that a risk minimizing policy can be chosen deterministic and stationary and that it can be computed efficiently.

**Definition 3.5.1.** *For a policy  $\pi \in \Pi_{m,r}$  and a time delay  $t \geq 0$ , define the delayed policy  $\tilde{\pi}^t \in \Pi_{m,r}$  by  $\tilde{\pi}_\tau^t = \pi_{\tau+t}$ . A multi-period Markovian risk measure is stationary if for all policies  $\pi \in \Pi_{m,r}$ , and all  $i, t$ ,*

$$\rho(C_t^\pi | t, S_t^\pi = i) = \rho(\beta^{t-1} C_1^{\tilde{\pi}^{t-1}} | 1, S_1 = i),$$

where  $\beta$  is the discount factor.

### 3.5.1 Coherent risk measure of discounted sample cost

We assume in this subsection that the discount factor is less than one and that the risk measure is coherent.

**Theorem 3.5.2.** *Let  $\rho$  be a coherent dynamically consistent stationary Markovian multi-period risk measure, and denote  $V^\pi(s_1) = \rho(C_1^\pi | s_1)$  for  $\pi \in \Pi_{m,r}$ , and  $V^*(i) = \inf_{\pi \in \Pi_{m,r}} V^\pi(i)$ . Assume that the discount factor  $\beta$  is in  $[0, 1)$  and that Assumptions 3.3.1 and 3.3.8 hold.*

(a) *If a stationary policy  $\pi \in \Pi_{s,r}$  satisfies  $V^\pi < +\infty$ , the vector  $V^\pi$  is the unique fixed point in  $\mathbb{R}^{\mathcal{X}}$  of  $\mathbf{T}_\pi^\beta$ .*

(b) *If there is a Markovian policy  $\pi \in \Pi_{m,r}$  such that  $V^\pi < +\infty$  and  $V^* > -\infty$ , then  $V^*$  is the unique fixed point in  $\mathbb{R}^{\mathcal{X}}$  of  $\mathbf{T}^\beta$ :  $V^* = \mathbf{T}^\beta V^*$  (Shapley-Bellman equations).*

(c) *Furthermore, for all  $V \in \mathbb{R}^{\mathcal{X}}$ ,  $\|(\mathbf{T}^\beta)^k V - V^*\|_\infty \leq \beta^k \|V - V^*\|_\infty$  so that  $(\mathbf{T}^\beta)^k V \rightarrow V^*$  for all  $V \in \mathbb{R}^{\mathcal{X}}$  (value iteration).*

(d) *There is a deterministic stationary Markovian policy  $\pi^* \in \Pi_{s,d}$  such that  $\mathbf{T}_{\pi^*}^\beta V^* = \mathbf{T}^\beta V^*$ . Moreover, the policy  $\pi^*$  achieves the minimum risk  $V^{\pi^*} = V^*$  and*

does not depend on the initial state.

**Remark 3.5.3.** *If all the immediate costs are bounded in absolute value by  $M$  with  $\mu$ -probability one, then there holds  $V^\pi < +\infty$  for all  $\pi \in \Pi_{m,r}$ .*

*On the other hand, the condition  $V^\pi < +\infty$  for  $\pi \in \Pi_{m,r}$  is not a consequence of Assumption 3.3.1, which states that the sample cost of any Markovian policy is integrable. The assumption that  $E_\mu[|C_1^\pi|] < +\infty$  implies that  $E_P[|C_1^\pi|] < +\infty$  for all  $P \in L^\infty(\Omega, \mathcal{F}, \mu)$ , but it does not necessarily imply that  $\sup_{P \in \mathcal{P}} E_P[|C_1^\pi|] < +\infty$  when  $\mathcal{P} \subset L^\infty(\Omega, \mathcal{F}, \mu)$ .*

*Proof.* (a) For  $\pi \in \Pi_{m,r}$ , the decomposition lemma 3.3.10 specialized to  $t = 1$  yields

$$V^\pi(s_1) = \rho(C_1^\pi | 1, s_1) = \sum_{a \in A} \pi_1(a | s_1) \rho_{(1, s_1, a)} [Q_1^\pi + \rho(C_2^\pi | 2, S_2^\pi)].$$

Since  $\rho$  is stationary and coherent,  $\rho(C_2^\pi | 2, S_2^\pi = j) = \rho(\beta C_1^{\tilde{\pi}^1} | 1, S_1^\pi = j) = \beta V^{\tilde{\pi}^1}(j)$ , and the previous equation becomes  $V^\pi = \mathbf{T}_{\pi_1}^\beta V^{\tilde{\pi}^1}$ . If the policy  $\pi$  is stationary and satisfies  $V^\pi < +\infty$ , then  $\pi = \tilde{\pi}^1$ . Hence,  $V^\pi$  is finite and is a fixed point of  $\mathbf{T}_{\pi_1}^\beta$  in  $\mathbb{R}^\mathcal{X}$ . Since  $\mathbf{T}_{\pi_1}^\beta$  is a  $\beta$ -contraction for the sup norm on  $\mathbb{R}^\mathcal{X}$  (Lemma 3.3.12), for all  $V \in \mathbb{R}^\mathcal{X}$ ,  $(\mathbf{T}_{\pi_1}^\beta)^k V \rightarrow V^\pi$  as  $k \rightarrow +\infty$ , and  $V^\pi$  is the unique fixed point of  $\mathbf{T}_{\pi_1}^\beta$ .

(b) By monotonicity of  $\mathbf{T}_{\pi_1}^\beta$ ,  $V^\pi = \mathbf{T}_{\pi_1}^\beta V^{\tilde{\pi}^1} \geq \mathbf{T}_{\pi_1}^\beta V^* \geq \mathbf{T}^\beta V^*$ . Taking the infimum with respect to  $\pi \in \Pi_{m,r}$  yields  $V^* \geq \mathbf{T}^\beta V^*$ .

If there exists  $\pi \in \Pi_{s,r}$  such that  $V^\pi(j) < +\infty$  for all  $j \in \mathcal{X}$ , then  $V^* \leq V^\pi < +\infty$ . Since  $V^* > -\infty$ ,  $V^* \in \mathbb{R}^\mathcal{X}$  is finite.

Consider a deterministic stationary Markovian policy  $\pi^* \in \Pi_{s,d}$  such that  $\mathbf{T}^\beta V^* = \mathbf{T}_{\pi_1^*}^\beta V^*$ . Since  $\mathbf{T}^\beta V^* = V^*$ , we have  $V^* = \mathbf{T}_{\pi_1^*}^\beta V^*$ . By monotonicity of  $\mathbf{T}_{\pi_1^*}^\beta$ ,  $V^* \geq (\mathbf{T}_{\pi_1^*}^\beta)^k V^*$  for all positive integers  $k$ . Since  $(\mathbf{T}_{\pi_1^*}^\beta)^k V^* \rightarrow V^{\pi^*} \geq V^*$ , we have  $V^{\pi^*} = V^* = \mathbf{T}^\beta V^*$ .

(c) Since  $\mathbf{T}^\beta$  is a  $\beta$ -contraction on  $\mathbb{R}^\mathcal{X}$ , this part is straightforward.

(d) This part was proved along the way in (a). □

### 3.5.2 Convex risk measure of undiscounted sample cost

To guarantee that an infinite horizon undiscounted sample cost is well-defined, we will assume that there is a special “termination” state  $\sigma \in \mathcal{X}$ . Hence, infinite horizon undiscounted cost problems can be thought as finite horizon problems with uncertain time horizon length, where the time horizon is the time until absorption by state  $\sigma$ . This set up generalizes the analysis of stochastic shortest path problem by dynamic programming when the discount factor  $\beta$  is one. Also zero-sum Markov games have been studied in this setting [60].

For infinite-horizon undiscounted cost problems, we can analyze the more general case of convex risk measure. We will establish results analogous to the finite-horizon case under the following additional assumption.

**Assumption 3.5.4.** (a) *There is a special state  $\sigma \in \mathcal{X}$ , which has zero-cost and is absorbing with  $\mu$ -probability 1, i.e.,*

$$\mu(N_{t,\sigma,a} = \sigma, Q_{t,\sigma,a} = 0, \forall a \in \mathcal{A}, t \geq 1) = 1.$$

(b) *Recall that for any fixed policy  $\pi \in \Pi_{s,r}$ , the Markov game defined in Subsection 3.3.4 becomes an MDP controlled by nature. We require that for every stationary policy  $\theta$  of nature either there exists an initial state  $i \in \mathcal{X}$  from which the expected reward of nature goes to  $-\infty$  or  $\theta$  is a proper policy. Recall that a stationary policy  $\theta$  is proper if for all initial states  $s_1$ , the limit as  $t \rightarrow +\infty$  of the probability under  $\theta$  that the state at time  $t$  is  $\sigma$  is equal to one.*

**Remark 3.5.5.** *Assumption 3.5.4 generalizes two usual assumptions in deterministic shortest path problems: 1) that every node is connected to the destination node  $\sigma$ , and 2) all cycles have positive length.*

**Theorem 3.5.6.** *Let  $\rho$  be a dynamically consistent stationary Markovian convex risk measure verifying Assumptions 3.3.1, 3.3.8 and 3.5.4. Moreover, assume that for  $\omega' \in \Omega'$   $\rho(\cdot|\omega')$  has a subdifferential everywhere on its domain, i.e., for all  $X \in \mathcal{H}$*

such that  $\rho(X|\omega') < +\infty$ , we have  $\partial\rho(X|\omega') \neq \emptyset$ . Let  $V^\pi(j) = \rho(C_1^\pi|j)$  be the risk of the sample cost under policy  $\pi \in \Pi_{m,r}$ .

(a) If a stationary policy  $\pi \in \Pi_{s,r}$  is such that  $V^\pi(j) < +\infty$  for all  $j \in \mathcal{X}$ , then  $V^\pi$  is the unique solution in  $\mathcal{V} = \{V \in \mathbb{R}^\mathcal{X} \mid V(\sigma) = 0\}$  of  $\mathbf{T}_{\pi_1} V^\pi = V^\pi$  (Bellman equation). Moreover,  $V^\pi = \lim_{k \rightarrow +\infty} \mathbf{T}_{\pi_1}^k V$  for any  $V \in \mathcal{V}$  (value iteration).

(b) Assume that for all deterministic stationary Markovian policies  $\pi \in \Pi_{s,d}$ ,  $V^\pi < +\infty$  and that  $V^*(j) = \inf_{\pi \in \Pi_{m,r}} V^\pi(j) > -\infty$ . If  $V^*(j) > -\infty$  for all  $j \in \mathcal{X}$ , then  $V^*$  is the unique fixed point in  $\mathcal{V}$  of  $\mathbf{T}$  (Shapley-Bellman equation). There is a deterministic stationary Markovian policy  $\pi^* \in \Pi_{s,d}$ ,  $\pi^* = (\pi_1^*, \pi_1^*, \dots)$ , such that  $\mathbf{T}_{\pi_1^*} V^* = \mathbf{T}V^*$ . Furthermore, the policy  $\pi^*$  is optimal,  $V^{\pi^*} = V^*$ , and does not depend on the initial state.

(c) Finally,  $\lim_{k \rightarrow +\infty} \mathbf{T}^k V = V^*$  for all  $V \in \mathcal{V}$  (value iteration).

**Remark 3.5.7.** In contrast to the discounted case,  $\mathbf{T}_\pi$  need not be a contraction with respect to any norm, as in [13]. However, in contrast to the results in [13], this theorem does not provide a tool to evaluate the risk associate with non-stationary Markovian policies.

**Remark 3.5.8.** The assumption that the convex risk measure  $\rho$  is subdifferentiable on its domain is mild since  $\rho$  is already a proper lower-semicontinuous convex functional. Hence, it is subdifferentiable on the relative interior of its domain. Essentially, the assumption that  $\rho$  is subdifferentiable on its domain rules out the possibility of  $\rho$  having “vertical derivatives,” that is positions around which the risk perception of the decision maker changes “infinitely fast.”

It is not clear whether the assumptions that  $V^\pi < +\infty$  for all  $\pi \in \Pi_{s,d}$  and  $V^* > -\infty$  are not redundant, because before proving that Bellman’s equations hold we do not know whether the test probability measures associated with  $\rho$  have the Markov property. If we knew this, we could use stronger properties of MDPs to by-pass these assumptions.

*Proof.* (a) Fix a stationary policy  $\pi \in \Pi_{s,r}$  such that  $V^\pi < +\infty$ . By mimicking the proof of Theorem 3.5.2, we have that  $V^\pi \in \mathcal{V} = \{V \in \mathbb{R}^\mathcal{X} \mid V(\sigma) = 0\}$  is a fixed point

of  $\mathbf{T}_\pi$ .

Now, we will prove that  $(\mathbf{T}_{\pi_1})^k V \rightarrow V^\pi$  as  $k \rightarrow +\infty$ . This property will follow from the analysis of Bertsekas and Tsitsiklis in [13]. Indeed, the operator  $\mathbf{T}_{\pi_1}$  can be interpreted as the classical dynamic programming operator for the MDP controlled by nature that is obtained when the decision maker uses policy  $\pi$  in the Markov game defined in Subsection 3.3.4. We will apply the results of [13] to this MDP.

For the reader's convenience, the main result of [13] is re-stated here with the set up and notations adapted to our situation, where nature maximizes her reward.

**Assumption 1**

State  $\sigma$  is absorbing and cost-free. Furthermore, there exists at least one proper stationary policy, and each improper stationary policy yields a reward to nature of  $-\infty$  for at least one initial state.

Since  $V^\pi = \mathbf{T}_{\pi_1} V^\pi$ , Lemma 1(b) in [13] guarantees the existence of a proper policy for nature when the stationary controller's policy  $\pi$  is fixed. This together with Assumption 3.5.4 imply that Assumption 1 is satisfied.

**Assumption 2**

For all states  $j$ , nature's action set  $\mathcal{P}^{1,j,a}$  is compact, the immediate cost functions  $(\bar{q}_{1,j,a}(P_{j,a}) - \bar{\phi}_{1,j,a}(P_{j,a}))$  are upper-semicontinuous in  $P_{j,a} \in \mathcal{P}^{1,i,a}$ , and the marginal distributions  $P(k|j, a)$  of  $P_{j,a}$  on  $\mathcal{X}$  are continuous in  $P_{j,a} \in \mathcal{P}^{1,i,a}$ .

Observe that Assumption 2 is not satisfied in general in our set up because the sets  $\mathcal{P}^{1,j,a}$  need not be compact. This assumption is needed for two reasons: 1) it guarantees that the supremum in Bellman equation is achieved and 2) it is used in [13] to show the existence of a fixed point to  $\mathbf{T}_{\pi_1}$ . Since in our case, we know that  $V^\pi$  is a fixed point, we could replace Assumption 2 by the weaker assumption that the supremum in the definition of  $\mathbf{T}_{\pi_1} V$ , i.e.,

$$(\mathbf{T}_{\pi_1} V)(j) = \sum_{a \in \mathcal{A}} \pi_1(a|j) \sup_{P_{j,a} \in \mathcal{P}^{1,j,a}} \left[ (\bar{q}_{1,j,a}(P_{j,a}) - \bar{\phi}_{1,j,a}(P_{j,a})) + \sum_{k \in \mathcal{X}} P(k|1, j, a) V(k) \right],$$

is achieved for all stationary policies  $\pi \in \Pi_{s,r}$  and all  $V \in \mathbb{R}^{\mathcal{X}}$ . Since we are assuming (in Theorem 3.5.6) that the conditional risk measures of  $\rho$  have a non-empty

subdifferential on their domain, Proposition 3.2.8 implies that the supremum in the representation of  $\rho(\cdot|1, j, a)$  is achieved for all  $j, a$ ; thus, the supremum in the definition of  $\mathbf{T}_{\pi_1} V$  is also achieved.

**Proposition 3.5.9** (Proposition 2 in [13]). *Let Assumption 1 and 2 hold. Then:*

(i) *The optimal cost vector  $V^\pi$  is the unique fixed point of  $\mathbf{T}_{\pi_1}$  in  $\mathcal{V}$ .*

(ii) *For every  $V \in \mathcal{V}$ , there holds  $\lim_{k \rightarrow +\infty} \mathbf{T}_{\pi_1}^k V = V^\pi$ .*

This proposition implies that  $(\mathbf{T}_{\pi_1})^k V \rightarrow V^\pi$  and that  $V^\pi$  is the unique fixed point of  $\mathbf{T}_{\pi_1}$  in  $\mathcal{V}$ , which concludes the proof for point (a).

(b) By following again the proof of Theorem 3.5.2, we have that  $V^* \geq \mathbf{T}V^*$ . Let  $\pi^* = (\pi_1^*, \pi_1^*, \dots) \in \Pi_{s,d}$  be a stationary deterministic policy such that  $\mathbf{T}V^* = \mathbf{T}_{\pi_1^*}V^*$ .

By applying  $\mathbf{T}_{\pi_1^*}^k$  on both sides, we have  $V^* \geq (\mathbf{T}_{\pi_1^*})^k V^* \rightarrow V^{\pi^*}$  as  $k \rightarrow +\infty$ , by (a). Since  $V^* \leq V^{\pi^*}$ , there holds  $V^* = \mathbf{T}V^* = V^{\pi^*}$ , and the deterministic stationary policy  $\pi^*$  is optimal and does not depend on the initial state.

In fact, there is a unique fixed point of  $\mathbf{T}$  in  $\mathcal{V}$ . Assume  $V^1, V^2 \in \mathcal{V}$  are fixed points of  $\mathbf{T}$ . Let  $\pi^1, \pi^2$  be two policies such that  $\mathbf{T}_{\pi_i^i} V^i = \mathbf{T}V^i$ ,  $i = 1, 2$ . We have  $V^1 = \mathbf{T}V^1 \leq \mathbf{T}_{\pi_1^2} V^1$ . Iterating this inequality, we obtain  $V^1 \leq \mathbf{T}_{\pi_1^2}^k V^1 \rightarrow V^2$ . The symmetric argument yields  $V_1 \geq V_2$ . Hence,  $V^1 = V^2$ , and  $V^*$  is the unique fixed point of  $\mathbf{T}$  in  $\mathcal{V}$ .

(c) Now, we prove that  $\lim_{k \rightarrow +\infty} \mathbf{T}^k V = V^*$  for all  $V \in \mathcal{V}$ . Recall the definition of  $\mathbf{e} \in \mathbb{R}^{\mathcal{X}}$  as the vector with all components equal to one, except for  $\mathbf{e}(\sigma) = 0$ . For a fixed  $\delta \geq 0$ , define  $\tilde{\mathbf{T}}_\delta$  the operator on  $\mathcal{V}$  by  $\tilde{\mathbf{T}}_\delta V = \mathbf{T}_{\pi^*} V + \delta \mathbf{e}$ . Note that  $\tilde{\mathbf{T}}$  is the dynamic programming operator for a modified MDP where all the immediate transition costs, at states other than  $\sigma$ , have been increased by  $\delta > 0$ . First, we show that there exists  $V_\delta \in \mathcal{V}$  such that  $\tilde{\mathbf{T}}_\delta V_\delta = V_\delta$ . This MDP verifies the assumptions of Proposition 2 in [13], which states that  $\tilde{\mathbf{T}}_\delta$  has a unique fixed point in  $\mathcal{V}$ , namely  $V_\delta$ . Moreover, since all the costs have been increased,  $V_\delta \geq V^{\pi^*}$ .

Now, we can write  $V^{\pi^*} = \mathbf{T}V^{\pi^*} \leq \mathbf{T}V_\delta \leq \mathbf{T}_{\pi^*} V_\delta = V_\delta - \delta \mathbf{e} \leq V_\delta$ . Applying the operators  $\mathbf{T} \leq \mathbf{T}_{\pi^*}$  to these inequalities yield  $V^{\pi^*} \leq \mathbf{T}^k V_\delta \leq \mathbf{T}^{k-1} V_\delta \leq V_\delta$ . Consequently,  $(\mathbf{T}^k V_\delta)_k$  is a decreasing sequence lower-bounded by  $V^{\pi^*}$ . Since  $\mathbf{T}$  is

continuous on  $(\mathbb{R}^{\mathcal{X}}, \|\cdot\|_{\infty})$  (Lemma 3.3.12, (d)), the sequence converges to the unique fixed point of  $\mathbf{T}$ ,  $V^*$ .

On the other hand, we can write  $V^{\pi^*} - \delta \mathbf{e} = \mathbf{T}V^{\pi^*} - \delta \mathbf{e} \leq \mathbf{T}(V^{\pi^*} - \delta \mathbf{e}) \leq \mathbf{T}V^{\pi^*} = V^{\pi^*}$ . Applying  $\mathbf{T}$  to these inequalities imply that  $\mathbf{T}^k(V^{\pi^*} - \delta \mathbf{e})$  is an increasing sequence, which is upper bounded by  $V^{\pi^*}$ . Hence, it converges and its limit is the unique fixed point of  $\mathbf{T}$ .

Since the state space  $\mathcal{X}$  is finite, for all  $V \in \mathcal{V}$ , we can find  $\delta \geq 0$  such that  $V^{\pi^*} - \delta \leq V \leq V_{\delta}$ . Applying  $\mathbf{T}$  on each side yields  $\mathbf{T}^k(V^{\pi^*} - \delta) \leq \mathbf{T}^k V \leq \mathbf{T}^k V_{\delta}$ . Taking the limit when  $k$  goes to  $+\infty$  shows that  $\mathbf{T}^k V \rightarrow V^*$ .  $\square$

### 3.5.3 An illustration: minimization of exponential utility function

We have seen in Section 3.2 that the functional  $\rho(X) = \frac{1}{\gamma} \log E_{\mu}[\exp(\gamma X)]$  with  $\gamma > 0$  is a convex risk measure on  $L^{\infty}(\Omega, \mathcal{F}, \mu)$  for any probability space  $(\Omega, \mathcal{F}, \mu)$ . Under an additional assumption on  $\mu$ , we will show that the risk measure  $\rho$  is Markovian and dynamically consistent, and (an adaptation of) Theorem 3.5.6 will apply. This subsection is meant to be an illustration of Theorem 3.5.6 since we will restrict ourselves to systems where the infinite horizon undiscounted cost belongs to  $L^{\infty}(\Omega, \mathcal{F}, \mu)$ . Moreover, the results obtained in this subsection are already known in the literature (e.g., [61]) but they were derived differently.

As detailed in Subsection 3.2.2, the risk measure  $\rho$  is convex only on  $L^{\infty}(\Omega, \mathcal{F}, \mu)$ , and not  $L^1(\Omega, \mathcal{F}, \mu)$ . Since the results of this chapter have been derived for risk measures on  $L^1(\Omega, \mathcal{F}, \mu)$ , we would need to adapt all the definitions and re-derive all the results to deal with  $L^{\infty}(\Omega, \mathcal{F}, \mu)$ . For the sake of brevity, we will not do it; recall from Theorem 3.2.11 that if a convex risk measure on  $L^{\infty}(\Omega, \mathcal{F}, \mu)$  satisfies the Fatou property, then the representation theorem holds with the role of  $L^1$  and  $L^{\infty}$  switched. As a result, all the definitions and results exposed so far can be adapted by:

- (i) replacing “convex risk measure” by “convex risk measure with the Fatou property,”

(ii) switching the role of  $L^1$  and  $L^\infty$ .

Until the end of this section, we will assume that the probability law  $\mu$  has a product form, i.e.,

$$\mu = \mathbb{P}_{S_1} \prod_{t=1}^T \left( \prod_{j \in \mathcal{X}} \mathbb{P}_{R_{t,j}} \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathbb{P}_{N_{t,j,a}} \mathbb{P}_{Q_{t,j,a}} \right). \quad (3.5.1)$$

In other words, all the components of a sample point  $\omega$  are generated independently of each other, including the cost  $Q_{t,j,a}$  and the state  $N_{t,j,a}$  that follows state-action pair  $(j, a)$  at time  $t$ . In particular, the trajectories distributed according to  $\mu$  have the Markov property. This assumption is more restrictive than before because in the present context  $\mu$  is supposed to represent the subjective probabilistic model for the system with respect to which the decision maker is basing its expectations (as opposed to the earlier situation where the support of  $\mu$  mostly mattered).

We note that the risk measure  $\rho(\cdot) = \frac{1}{\gamma} \log E_\mu[\exp(\gamma \cdot)]$  is dynamically consistent. Indeed, let  $\omega'$  be a partial history, and  $X, Y$  two positions. If for all  $\nu \in \mathcal{E}(\omega')$   $E_\mu[\exp(\gamma X)|\omega' \nu] \leq E_\mu[\exp(\gamma Y)|\omega' \nu]$ , then  $E_\mu[\exp(\gamma X)|\omega'] \leq E_\mu[\exp(\gamma Y)|\omega']$ , since  $E_\mu[\exp(\gamma X)|\omega'] = E_\nu E_\mu[\exp(\gamma X)|\omega' \nu]$ .

For a product form  $\mu$ , the conditional risk measure  $\rho(\cdot) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma \cdot))$  is Markovian. Indeed, let  $\omega \in \Omega$  such that  $S_t^\pi(\omega) = j$  and  $A_t^\pi(\omega) = a$ , then, by conditioning on the subsequent state,

$$\begin{aligned} \log E_\mu[\exp(\gamma C_t^\pi)|\omega_t] &= \log \sum_{k \in \mathcal{X}} \mathbb{P}_{N_{t,j,a}}(k) E_\mu[\exp(\gamma Q_{t,j,a}) \cdot \exp(\gamma C_{t+1}^\pi) | \omega_t, S_{t+1}^\pi = k] \\ &= \log \sum_{k \in \mathcal{X}} \mathbb{P}_{N_{t,j,a}}(k) E_{\mathbb{P}_{Q_{t,j,a}}}[\exp(\gamma Q_{t,j,a})] \cdot E_\mu[\exp(\gamma C_{t+1}^\pi) | S_{t+1}^\pi = k], \\ &= \log E_{\mathbb{P}_{Q_{t,j,a}}}[\exp(\gamma Q_{t,j,a})] + \log \sum_{k \in \mathcal{X}} \mathbb{P}_{N_{t,j,a}}(k) \cdot E_\mu[\exp(\gamma C_{t+1}^\pi) | S_{t+1}^\pi = k], \end{aligned}$$

where the second equality uses the independence between the immediate transition cost and the next occupied state as well as the Markov property of the state process. Since the right-hand side depends on  $\omega$  only via  $S_t^\pi(\omega)$  and  $A_t^\pi(\omega)$ , the risk measure

$\rho(\cdot) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma \cdot))$  satisfies condition (a) of the definition of a Markovian risk measure. A similar argument shows that condition (b) is also satisfied.

Now, we can rewrite Theorem 3.5.6 for the risk measure  $\rho(X) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma X))$ ,  $\gamma > 0$ , associated with the exponential utility function.

**Proposition 3.5.10.** *We assume that the probability law  $\mu$  has the product form of Equation (3.5.1), is such that Assumption 3.5.4 holds, and that for all  $\pi \in \Pi_{m,r}$  and  $t \geq 1$ , the sample cost  $C_t^\pi$  is in  $L^\infty(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  for all  $\omega' \in \Omega' \cup \{\emptyset\}$  (adaptation of Assumption 3.3.1).*

Let  $V^\pi(j) = \frac{1}{\gamma} \log E_\mu[\exp(\gamma C_1^\pi) | S_1 = j]$  and  $V^*(j) = \inf_{\pi \in \Pi_{m,r}} V^\pi(j)$  with  $\gamma > 0$ . Then, the following statements hold:

(a)  $V^\pi$  is the unique solution in  $\mathcal{V}$  of

$$V^\pi(j) = \sum_{a \in \mathcal{A}_j} \pi(a|j) \left( \frac{1}{\gamma} \log E_{\mathbb{P}_{Q_{1,j,a}}}[\exp(\gamma Q_{1,j,a})] + \frac{1}{\gamma} \log \sum_{k \in \mathcal{X}} \mathbb{P}_{N_{1,j,a}}(k) \exp(\gamma V^\pi(k)) \right).$$

(b) The limit  $\lim_{k \rightarrow +\infty} \mathbf{T}_\pi^k V$  is  $V^\pi$  for all  $V \in \mathcal{V}$  (value iteration).

(c) If  $V^*(j) > -\infty$  for all  $j \in \mathcal{X}$ ,  $V^*$  is the unique solution in  $\mathcal{V}$  of the equations

$$V^*(j) = \min_{a \in \mathcal{A}_j} \left( \frac{1}{\gamma} \log E_{\mathbb{P}_{Q_{1,j,a}}}[\exp(\gamma Q_{1,j,a})] + \frac{1}{\gamma} \log \sum_{k \in \mathcal{X}} \mathbb{P}_{N_{1,j,a}}(k) \exp(\gamma V^*(k)) \right).$$

(d) If  $V^*(j) > -\infty$  for all  $j \in \mathcal{X}$ , there exists an optimal deterministic stationary Markovian policy  $\pi^*$  that satisfied  $\mathbf{T}_{\pi^*} V^* = \mathbf{T} V^*$  and that does not depend on the initial state.

(e) If  $V^*(j) > -\infty$  for all  $j \in \mathcal{X}$ , the limit  $\lim_{k \rightarrow +\infty} \mathbf{T}^k V$  is  $V^*$  for all  $V \in \mathcal{V}$  (value iteration).

**Remark 3.5.11.** *This proposition is not entirely contained in [61], which assumes that the immediate costs are positive (Assumption 1 (i)).*

*Proof.* Since the functionals  $\rho(\cdot|\omega') = \frac{1}{\gamma} \log E_\mu[\exp(\gamma\cdot)|\omega']$ ,  $\omega' \in \Omega'$ , form a dynamically consistent Markovian convex risk measure, this proposition is an application of (an adaptation of) Theorem 3.5.6, so we simply need to verify that its assumptions are satisfied.

Without any assumption on the probability space  $(\Omega, \mathcal{F}, \mu)$ , let  $(X_n)$  be a uniformly bounded sequence in  $L^\infty(\Omega, \mathcal{F}, \mu)$  such that  $X_n \downarrow X$   $\mu$ -almost surely. By the monotone convergence theorem,  $E_\mu(\exp(\gamma X_n)) \downarrow E_\mu(\exp(\gamma X))$ , and  $\rho(X_n) \rightarrow \rho(X)$ . Hence,  $\rho(X) = \frac{1}{\gamma} \log E_\mu(\exp(\gamma X))$  with  $\gamma > 0$  has the Fatou property (Definition 3.2.9 (c)).

Assumption 3.3.8 is easily verified in our present case.

Since  $C_1^\pi \in L^\infty(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  for all  $\omega' \in \Omega'$ , there exists  $M < +\infty$  such that  $l_\infty(C_1^\pi) < M$   $\mu$ -almost surely, and thus  $\rho(C_1^\pi|j) < M < +\infty$ .

□

The functional  $\rho$  defined by  $\rho(X) = \frac{1}{\gamma} \log E_\mu[\exp(\gamma X)]$  is a risk measure on  $L^\infty(\Omega, \mathcal{F}, \mu)$ , but not on  $L^p(\Omega, \mathcal{F}, \mu)$  for  $p \in [1, +\infty)$ . Intuitively, condition (5) in the definition of convex risk measure is not satisfied when the tails of the functional space  $\mathcal{H}$  do not decrease at least exponentially. In order to have a result of the form of Proposition 3.5.10 that is more practical, one could consider a functional space whose elements have exponentially decreasing tails and use the representation theorem for convex risk measures on general functional spaces in [77]. We did not investigate this avenue because it does not serve our main research objectives.

In summary of the last two sections, we established new representation theorems for dynamically consistent Markovian convex risk measures of the sample cost, and showed that their minimization is equivalent to solving a zero-sum sequential game with nature.

In the rest of this chapter, we will take the converse approach of starting from uncertainty sets and penalty functions to define well-behaved risk measures.

### 3.6 From robust control to risk minimization

In this section we still deal with systems of the form described in Section 3.3, but we take a converse approach to that of Sections 3.4 and 3.5. Now, instead of representing risk measures as worst-case penalized expectations, we will construct dynamically consistent Markovian convex risk measures starting from uncertainty sets and penalty functions. The minimization of the coherent risk measures of this class of risk corresponds to the worst-case robust control of uncertain MDPs of [54, 38]. Furthermore, the convex risk measures of this class of risk allows us to motivate another —potentially less conservative— robust formulation in which nature is penalized for choosing “unlikely” parameters. Together with Theorems 3.4.3, 3.5.2 and 3.5.6, this establishes the equivalence of minimizing dynamically consistent Markovian convex risk measures of sample costs (which is well-motivated from the perspective of risk-averse decision theory) and of solving zero-sum games between the decision maker and nature (which are computationally “tractable”). This gives us two ways to think about robust control of MDPs. Finally, we generalize the aforementioned construction of multi-period risk measures to define multi-period risk measures with desirable properties (dynamical consistency, Markovianity) starting from single-period convex risk measures.

In the sequel, we deal with technicalities from measure theory and functional analysis to cope with systems as general as the ones defined in Section 3.3, but the reader is encouraged to consider a finite sample space  $\Omega$  to focus on the heart of the matter.

Our model for dynamical systems is still the one described at the beginning of Section 3.3, with the reference probability space  $(\Omega, \mathcal{F}, \mu)$ . Denote  $\mu_{|\omega_{t-1}^+}^t$  the single-step marginal probability measure of  $\mu_{|\omega_{t-1}^+}$  on  $(R_{t,i})_{i \in \mathcal{X}}$ . Recall that we have assumed that each random number  $R_{t,i}$  is generated independently of everything else so that  $\mu_{|\omega_{t-1}^+}^t$  does not depend explicitly on  $\omega_{t-1}^+$ . For a given partial history  $\omega_t$ , we denote by  $P^t$  the one-step marginal distribution on  $\mathcal{E}(\omega_t)$  of any probability measure  $P$  on the measurable space  $(F(\omega_t), \mathcal{F}(\omega_t))$ , and by  $P^{t,j,a}$  the marginal of  $P$  over  $(N_{t,j,a}, \mathcal{Q}_{t,j,a})$ .

Also denote  $P_{|v}$  a version (they are all equal with  $P$ -probability one) of the conditional probability distribution of  $P$  given  $\omega_t$  and  $v \in \mathcal{E}(\omega_t)$ .

### 3.6.1 A multi-period risk measure induced by uncertainty sets and penalty functions

In this subsection we construct dynamically consistent Markovian convex multi-period risk measures starting from uncertainty sets and penalty functions. This construction provides a converse to the representation results of Proposition 3.4.1 and of Theorems 3.5.2 and 3.5.6.

Let  $\mathcal{P}_{t,j,a}$  be an (uncertainty) set of probability measures on  $(\mathcal{X} \times \mathbb{R})$ , absolutely continuous with respect to  $\mu_{|\omega_t}^{t,j,a}$  for all  $\omega \in \Omega$ , and such that all the derivatives are in  $L^\infty(F(\omega_t), \mathcal{F}^s(\omega_t), \mu_{|\omega_t}^t)$ . (A given probability measure can be continuous with respect to several different measures and have different derivatives with respect to each of them. Therefore, in this subsection, we will deal directly with probability measures, not with their derivatives. For example, the uncertainty sets  $\mathcal{P}_{t,j,a}$  are now sets of probability measures, not of derivatives with respect to some reference probability measure.) Observe that the uncertainty set  $\mathcal{P}_{t,j,a}$  contain test probability measures for both the next state occupied at time  $t + 1$  after the state-action pair  $(j, a)$  and the associated transition cost. Given the uncertainty sets  $\mathcal{P}_{t,j,a}$ , we can define the sets of probability measures  $\mathcal{P}(\omega')$  on  $(F(\omega'), \mathcal{F}(\omega'))$ ,  $\omega' \in \Omega'$  by

$$\mathcal{P}(\omega_t) = \left\{ \left( \prod_{\tau=t}^{T-1} Z_{\omega_\tau} \mu_{|\omega_\tau}^t \right) Z_{\omega_T} \text{ with } Z_{\omega_\tau} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{\tau,j,a}, \tau \geq t \right\}$$

and

$$\mathcal{P}(\omega_{t-1}^+) = \left\{ \mu_{|\omega_{t-1}^+}^t \left( \prod_{\tau=t}^{T-1} Z_{\omega_\tau} \mu_{|\omega_\tau}^t \right) Z_{\omega_T} \text{ with } Z_{\omega_\tau} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{\tau,j,a}, \tau \geq t \right\}.$$

Observe that these sets are “rectangular” across time and state-action pairs. Intuitively, the knowledge of some components of an element of these sets does not provide

information on the values of other components.

Let  $\phi_{t,j,a}$  be non-negative measurable (not necessarily convex) penalty functions on the uncertainty sets  $\mathcal{P}_{t,j,a}$ . These penalty functions do not appear in the worst-case control of uncertain MDPs, but they will be important for our extension. To guarantee normalization of the risk measure, defined later by Equation (3.6.1), (i.e.,  $\rho(0|\omega') = 0$  for all  $\omega' \in \Omega'$ ), we assume that for all  $t, j, a$ , there is some  $P_{t,j,a}^0 \in \mathcal{P}_{t,j,a}$  such that  $\phi_{t,j,a}(P_{t,j,a}^0) = 0$ . For  $r \geq 1$ , define penalty functions from  $\mathcal{P}(\omega_r)$  into  $\mathbb{R} \cup \{+\infty\}$  by

$$\phi_{\omega_r}(P) = E_P \left[ \sum_{t=r}^T \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{|\omega_t}^{t,j,a}) \right].$$

Similarly, define the penalty function  $\phi_{\omega_{r-1}^+}(P) = E_P \left[ \sum_{t=r}^T \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{|\omega_t}^{t,j,a}) \right]$  for  $P \in \mathcal{P}(\omega_{r-1}^+)$ . The following lemma guarantees that these objects are well-defined.

**Lemma 3.6.1.** *For  $T$  finite or infinite, the penalty functions  $\phi_{\omega_r}$  and  $\phi_{\omega_{r-1}^+}$  are measurable functions respectively from  $\mathcal{P}(\omega_r)$  and from  $\mathcal{P}(\omega_{r-1}^+)$  to  $\mathbb{R} \cup \{+\infty\}$ . In particular, their value do not depend on the choice of the conditional probabilities of  $P$ .*

Furthermore,  $\phi_{\omega_{r-1}^+}(P) = E_{P^r} \left[ \phi_{\omega_{r-1}^+ V}(\tilde{P}_V) \right]$  and  $\phi_{\omega_r}(P) = \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{r,j,a}(P^{r,j,a}) + E_{P^r} \left[ \phi_{\omega_r V}(\tilde{P}_V) \right]$ .

*Proof.* First, observe that the expressions  $\phi_{\omega_r}$  and  $\phi_{\omega_{r-1}^+}$  are sums of measurable and non-negative terms.

Now we show that  $\phi_{\omega_r}(P)$  does not depend on the choice of the conditional distributions of  $P \in \mathcal{P}(\omega_r)$ . For  $t \geq r+1$ , let  $Z_{\omega_t}, Z'_{\omega_t}$  be versions of the conditional probability measure of  $P_{|\omega_t}$ . They agree with  $P$ -probability one and  $E_P \left[ \phi_{t,j,a}(Z_{\omega_t}^{t,j,a}) \right] = E_P \left[ \phi_{t,j,a}(Z'_{\omega_t}{}^{t,j,a}) \right]$  for all  $t \geq r$ ,  $(j, a) \in \mathcal{X}\mathcal{A}$ . This implies that

$$\phi_{\omega_r}(P) = E \left[ \sum_{t=r}^T \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{|\omega_t}^{t,j,a}) \right]$$

does not depend on the versions of the conditional distributions of  $P$ .

A similar argument shows that the penalty functions  $\phi_{\omega_{r-1}^+}$  do not depend on the version of the conditional probability measures of  $P$ .

By the law of iterated expectation, we have for  $P \in \mathcal{P}(\omega_{r-1}^+)$  that

$$\phi_{\omega_{r-1}^+}(P) = E_P \left[ E_{P|V} \left[ \sum_{t=r}^T \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{|\omega_t}^{t,j,a}) \right] \right],$$

where  $P|V$  is a regular conditional probability measure of  $P$  given  $V \in \mathcal{E}(\omega_{r-1}^+)$  [16]. More precisely, we can pick a version  $P|_v$  of the conditional probability measure of  $P$  such that  $\tilde{P}_v$ , the restriction of  $P|_v$  on the measurable subspace  $(F(\omega_t v), \mathcal{F}(\omega_t v))$ , belongs to  $\mathcal{P}(\omega_{r-1}^+ v)$  for all  $v \in \mathcal{E}(\omega_{r-1}^+)$ , and write  $\phi_{\omega_{r-1}^+}(P) = E_P \left[ \phi_{\omega_{r-1}^+ v}(\tilde{P}_v) \right]$ .

The decomposition of  $\phi_{\omega_r}$  follows a similar argument.  $\square$

Now, we can define a convex risk measure  $\rho(\cdot|\omega')$  on  $L^1(F(\omega'), \mathcal{F}(\omega'), \mu_{|\omega'})$  associated with the uncertainty sets  $\mathcal{P}_{t,j,a}$  and the penalty functions  $\phi_{t,j,a}$ :

$$\rho(X|\omega') = \sup_{P \in \mathcal{P}(\omega')} (E_P[X|\omega'] - \phi_{\omega'}(P)). \quad (3.6.1)$$

When the penalty functions  $\phi_{t,j,a}$  are all zero, the conditional risk measure  $\rho(\cdot|\omega')$  is coherent.

**Remark 3.6.2.** Equation (3.6.1) suggests that nature pays some penalty even for parameters “outside of the trajectory.” But since nature picks the probability measure  $P \in \mathcal{P}(\omega')$  after she observes the partial history  $\omega_t$ , she knows the current state-action pair, and can pick the parameters  $P_{|\omega_t}^{t,j,a} = P_{t,j,a}^0$  for all the state-action pairs  $(j,a)$  that are not the current state-action pair. Hence, our proposed form for the penalty function works similarly as in the MDP introduced in Subsection 3.3.4.

**Proposition 3.6.3.** The mapping that associates to  $\omega' \in \Omega'$  the convex risk measure defined by (3.6.1) is a dynamically consistent and Markovian multi-period risk measure on  $L^1(\Omega, \mathcal{F}, \mu)$ , which satisfies Assumption 3.3.8 (i.e., controller risk-neutral with respect to the uncertainty of the random number generator).

*Proof.* First, let us show that it is dynamically consistent. To verify condition (a) of Definition 3.3.6, it is enough to show that  $\rho(X|\omega') = \rho_{\omega'}[\rho(X|\omega' \cdot)]$  for all  $\omega' \in \Omega' \cup \emptyset$ . Indeed, if  $\rho(X|\omega'v) \leq \rho(Y|\omega'v)$  for all immediate continuations  $v \in \mathcal{E}(\omega')$ , we have by monotonicity of  $\rho_{\omega'}$  that  $\rho_{\omega'}[\rho(X|\omega'v)] = \rho(X|\omega'v) \leq \rho(Y|\omega'v) = \rho_{\omega'}[\rho(Y|\omega'v)]$ , and thus  $\rho(X|\omega') \leq \rho(Y|\omega')$ .

For  $P \in \mathcal{P}(\omega_t)$  and  $X \in L^1(F(\omega_t), \mathcal{F}(\omega_t), \mu_{|\omega_t})$ ,

$$\begin{aligned} E_P[X|\omega_t] - \phi_{\omega_t}(P) &= E_{P^t} [E_P[X|\omega_t V]] - \phi_{\omega_t}(P) \\ &= E_{P^t} [E_P[X|\omega_t V] - \phi_{\omega_t V}(P)] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P^{t,j,a}) \\ &\leq E_{P^t} [\rho(X|\omega_t V)] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P^{t,j,a}) \\ &\leq \sup_{\bar{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}} \left[ E_{\bar{P}} \rho(X|\omega_t V) - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(\bar{P}^{j,a}) \right]. \end{aligned}$$

Hence, taking the supremum of the left-hand side with respect to  $P \in \mathcal{P}(\omega_t)$ ,

$$\rho(X|\omega_t) \leq \sup_{\bar{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}} \left[ E_{\bar{P}} \rho(X|\omega_t V) - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(\bar{P}^{j,a}) \right].$$

A similar computation yields for  $X \in L^1(F(\omega_{t-1}^+), \mathcal{F}(\omega_{t-1}^+), \mu_{|\omega_{t-1}^+})$

$$\rho(X|\omega_{t-1}^+) \leq \sup_{\bar{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}} E_{\bar{P}} [\rho(X|\omega_t)].$$

To show that in fact equality holds, we will show that the left-hand side of the above expression is arbitrarily close to  $E_{\bar{P}}[\rho(X|\omega_t)]$  for any given  $\bar{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}$ . Let  $\bar{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}$  and  $\epsilon > 0$ . For all  $v \in \mathcal{E}(\omega_t)$ , there exists  $\bar{P}_v \in \mathcal{P}(\omega_t v)$  such that  $\rho(X|\omega_t v) \geq E_{\bar{P}_v}[X|v] - \phi_{\omega_t v}(\bar{P}_v) - \epsilon$ . Since  $\mathcal{P}(\omega_t)$  is “time-rectangular”, we can choose  $P \in \mathcal{P}(\omega_t)$  such that  $P^t = \bar{P}$  and  $\tilde{P}_v = \bar{P}_v$  for which we can write

$$\begin{aligned}
\rho(X|\omega_t) &\geq E_P[X|\omega_t] - \phi_{\omega_t}(P) = E_{P^t} \left[ E_P[X|\omega_t V] - \phi_{\omega_t V}(\tilde{P}_V) \right] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P^{t,j,a}) \\
&\geq E_{P^t} [\rho(X|\omega_t V) - \epsilon] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P^{t,j,a})
\end{aligned}$$

Since this inequality holds for all  $\tilde{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}$  and all  $\epsilon > 0$ ,

$$\rho(X|\omega_t) \geq \sup_{\tilde{P} \in \prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}} \left[ E_{\tilde{P}} [\rho(X|\omega_t V)] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(\tilde{P}^{j,a}) \right].$$

Consequently,  $\rho(X|\omega_t) = \rho_{\omega_t}[\rho(X|\omega_t \cdot)]$ .

A similar argument holds to show that  $\rho(X|\omega_{t-1}^+) = \rho_{\omega_{t-1}^+}[\rho(X|\omega_{t-1}^+ \cdot)]$ . Condition (b) of Definition 3.3.6 is satisfied. Consequently, the multi-period risk measure  $\rho$  defined by Equation (3.6.1) is dynamically consistent.

Now, we show that the risk measure  $\rho$  is Markovian. Fix  $\pi \in \Pi_{m,r}$  and consider  $\omega, \bar{\omega} \in \Omega$  such that  $S_t^\pi(\omega) = S_t^\pi(\bar{\omega})$  and  $A_t^\pi(\omega) = A_t^\pi(\bar{\omega})$ .

By definition,  $\rho(C_t^\pi|\omega_t) = \sup_{P \in \mathcal{P}(\omega_t)} (E_P[C_t^\pi] - \rho_{\omega_t}(P))$ . Here,  $\mathcal{P}(\omega_t)$  (resp.  $\mathcal{P}(\bar{\omega}_t)$ ) is a set of probability measures on the measurable space  $(F(\omega_t), \mathcal{F}(\omega_t))$  (resp.  $(F(\bar{\omega}_t), \mathcal{F}(\bar{\omega}_t))$ ). Although  $\mathcal{P}(\omega_t)$  and  $\mathcal{P}(\bar{\omega}_t)$  live on distinct probability spaces, we will see that they differ only by the prefix up to time  $t-1$  of the sample point and that we can identify them with each other by ignoring the prefix. Indeed, write  $P \in \mathcal{P}(\omega_t)$  as  $P = \left( \prod_{\tau=t}^{T-1} Z_{\omega_\tau} \mu_{|\omega_\tau^+}^s \right) Z_{\omega_T}$  and  $\tilde{P} \in \mathcal{P}(\bar{\omega}_t)$  as  $\tilde{P} = \left( \prod_{\tau=t}^{T-1} \tilde{Z}_{\bar{\omega}_\tau} \mu_{|\bar{\omega}_\tau^+}^s \right) \tilde{Z}_{\omega_T}$ , where  $Z_{\omega_\tau}, \tilde{Z}_{\bar{\omega}_\tau}$  are in  $\prod_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{\tau,j,a}$  for  $\tau \geq t$ . When  $\omega_\tau$  and  $\bar{\omega}_\tau$  differ only by their prefix, we can let  $Z_{\omega_\tau} = \tilde{Z}_{\bar{\omega}_\tau}$  for  $\tau \geq t$ . Also recall that the distribution of the random numbers  $(R_{t,i})_i, \mu_{|\omega_\tau^+}^s$ , does not depend explicitly on  $\omega_\tau^+$  since each random number is generated independently of the rest. Given this warning, we can write with a slight abuse of notation,  $\mathcal{P}(\omega_t) = \mathcal{P}(\bar{\omega}_t)$  and  $\phi_{\omega_t}(P) = \phi_{\bar{\omega}_t}(P)$ . Therefore,  $\rho(C_t^\pi|\omega_t) = \rho(C_t^\pi|\bar{\omega}_t)$ .

The same argument holds for the case of  $\omega_{t-1}^+$ . This concludes the proof that  $\rho$  is Markovian.

Finally, from the definition of the uncertainty sets  $\mathcal{P}(\omega')$  and penalty functions  $\phi_{\omega'}$ , Assumption 3.3.8 is satisfied.  $\square$

The single-period risk measure associated with  $\rho(\cdot|\omega_t)$  applied to a single-period position  $X^s \in L^1(F(\omega_t), \mathcal{F}^s(\omega_t), \mu_{|\omega_t})$  takes the form

$$\begin{aligned} \rho_{\omega_t}(X^s) &= \sup_{P \in \mathcal{P}} (E_P[X^s] - \phi_{\omega_t}(P)) & (3.6.2) \\ &= \sup_{(P_{t,j,a})_{(j,a) \in \Pi_{(j,a) \in \mathcal{X}\mathcal{A}} \mathcal{P}_{t,j,a}}} \left( E_{(P_{t,j,a})}[X^s] - \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{t,j,a}) \right). \end{aligned}$$

Indeed, we do not lose anything by letting  $P_{|\omega_\tau}^t = \prod_{(j,a) \in \mathcal{X}\mathcal{A}} P_{\tau,j,a}^0$  for all time  $\tau \geq t+1$  in the supremum of Equation (3.6.1). This observation closes the loop in the following way: from the uncertainty sets  $\mathcal{P}_{t,j,a}$  and penalty functions  $\phi_{t,j,a}$  we defined a dynamically consistent Markovian convex risk measure  $\rho$  to which we can apply the analysis of Sections 3.4 and 3.5. The theorems therein lead to test probability measures and penalty functions associated with  $\rho$ , but these are essentially the uncertainty sets and the penalty functions we started with.

For the sake of illustration, consider a finite time horizon  $T$ . According to Theorem 3.4.3, the minimal risk  $\inf_{\pi \in \Pi_{m,r}} \rho(C^\pi | s_1) = V(1, s_1)$  is given by Bellman recursion, which has the form

$$\begin{aligned} V(T, j) &= \min_{a \in \mathcal{A}_j} \sup_{P_{T,j,a}} E_{P_{T,j,a}}[Q_{T,j,a}] - \phi_{T,j,a}(P_{T,j,a}), \quad j \in \mathcal{X} \\ V(t, j) &= \min_{a \in \mathcal{A}_j} \sup_{P_{t,j,a}} E_{P_{t,j,a}} [Q_{t,j,a} - \phi_{t,j,a}(P_{t,j,a}) + \beta V(t+1, N_{t,j,a})], \quad j \in \mathcal{X}, \quad t = 1, \dots, T-1. \end{aligned}$$

Observe how the penalties  $\phi_{t,j,a}$  appear in a natural way in Bellman's equations, even though the formal penalty function  $\phi_{\omega_r}(P) = E_P \left[ \sum_{t=r}^T \sum_{(j,a) \in \mathcal{X}\mathcal{A}} \phi_{t,j,a}(P_{|\omega_t}^{t,j,a}) \right]$  seems intricate.

### 3.6.2 Connection with robust control formulation of MDPs

In this subsection, we point out a connection between the worst-case robust control of uncertain MDPs and risk-sensitive control of MDPs. Then, we propose an improved robust formulation for the control of uncertain MDPs based on the minimization of a certain convex risk measure of the sample cost.

When the multi-period risk measure  $\rho$  is defined by Equation (3.6.1) with  $\phi_{\omega'} = 0$  for all  $\omega' \in \Omega'$ , the problem  $\min_{\pi} \rho(C^{\pi}|s_1)$  is equivalent to the robust control problem formulation of [78, 54, 38] with uncertainty sets  $\mathcal{P}_{t,j,a}$ . Actually, their robust control formulation does not explicitly allow for uncertain transition costs (while ours does), but this can be easily accommodated. Since their robust control formulation amounts to minimizing a dynamically consistent Markovian coherent risk measure, Theorems 3.4.3 and 3.5.2 apply and recover most of their results, while Theorem 3.5.6 analyzes the case of undiscounted sample costs over an infinite horizon.

The papers [78, 54, 38] assume that the uncertainty sets are rectangular, but do not study, neither from a computational, nor from a decision theoretic perspective, what happens if this assumption fails to hold. In these papers, the uncertainty sets are not updated as new information comes in during the control phase. They propose to design uncertainty sets based on probabilistic models, e.g., relative entropy or likelihood level sets. In these cases, we may consider updating the uncertainty sets on the basis of new observations, for example using Bayes rule.

This suggests that the controller could decide at some time point to refine the uncertainty sets and solve again the corresponding robust control problem. However, the subsequent example will show that this procedure can lead to severe time inconsistency in the decision maker preferences, when the uncertainty sets are not rectangular. This pitfall adds to negative computational complexity results established in Chapter 2, where we showed that the worst-case optimal control of uncertain MDPs becomes at least NP-hard when the uncertainty sets are not rectangular.

**Example** The following example is inspired from [72] and illustrated in Figure 3-2. Consider a two-period economy with two outcomes per period, where the market

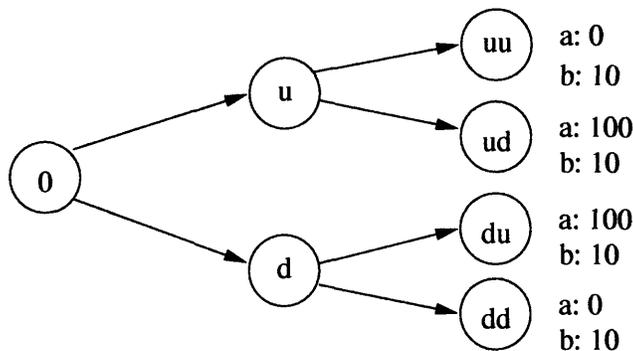


Figure 3-2: Simple model of a two-period market

goes either up ( $u$ ) or down ( $d$ ), but has uncertain dynamics. To model this situation, let us consider two possible MDPs, parameterized by an (uncertain) probability  $p_u \in \{0.01, 0.99\}$ , on the state space  $\mathcal{X} = \{\emptyset, u, d, uu, dd\}$  and action space  $\{a, b\}$ . In both MDPs, the initial state is  $\emptyset$  and there are two time steps. In the first period, there is no control and no cost, and the system moves from state  $\emptyset$  to state  $u$  with probability  $p_u$ , and to state  $d$  with probability  $1 - p_u$ . In the second period, the system state transition is still exogenous: state  $u$  is followed by state  $uu$  with probability  $p_u$ , and by state  $ud$  with probability  $1 - p_u$ ; and state  $d$  is followed by state  $du$  with probability  $p_u$ , and by state  $dd$  with probability  $1 - p_u$ . However, in the second period, the controller chooses between actions  $a$  and  $b$  as a function of the current state. Action  $a$  in state  $u$  has cost 100 if the system moves to state  $ud$  and zero otherwise; action  $a$  has cost 100 in state  $d$  if the next state is  $du$  and zero otherwise. Action  $b$  always costs 10. Observe that the uncertainty set is not “rectangular” since  $p_u = 0.99$  in the first period cannot be followed by  $p_u = 0.01$  in the second period.

For  $\omega' \in \{\emptyset, u, d, uu, ud, du, dd\}$ , define

$$\rho(C^\pi|\omega') = \max_{i=1,2} E_{P_i}[C^\pi|\omega'], \quad (3.6.3)$$

where  $P_1$  (resp.  $P_2$ ) refers to the MDP with  $p_u = 0.99$  (resp.  $p_u = 0.01$ ). Then,  $\rho$  is a coherent multi-period risk measure, but we show now that it is not dynamically consistent. In this simple model, there are only four deterministic policies  $\pi$  (since

there are two possible actions in the two states  $u$  and  $d$ ), and their respective worst-case costs  $\rho(\pi)$  are listed in the following table, where the rows (resp. columns) are indexed by the action chosen in state  $u$  (resp.  $v$ ).

	$a$	$b$
$a$	$2 \times 0.01 \times 0.99 \times 100 = 1.98$	$0.99 \times 10 + 0.01 \times 0.99 \times 100 = 10.89$
$b$	$0.99 \times 10 + 0.01 \times 0.99 \times 100 = 10.89$	10

Therefore, the policy that chooses action  $a$  in both states  $u$  and  $d$  minimizes  $\rho(C_1^\pi|0)$  over all Markovian deterministic policies  $\pi$ .

To compare the actions  $a$  and  $b$  given a first period outcome, we need to specify how the uncertainty about  $p_u$  is updated as new information is available to the decision maker. The definition of  $\rho$  by Equation (3.6.3) assumes that the uncertainty is unchanged, i.e.  $p_u$  can be either 0.99 or 0.01. In that case, the worst-case cost of action  $a$  in the second period given that the system is in state  $u$  increases to 99, which is worse than the cost of  $b$ . Similarly, if the system goes to state  $d$ ,  $X$  also becomes worse than  $Y$ . Hence, the decision rule based on  $\rho$  is not dynamically consistent.

In addition to the dynamic inconsistency highlighted by the above example, another criticism of the robust control formulation that solves

$$\inf_{\pi \in \Pi_{m,r}} \sup_{P \in \mathcal{P}} E_P[C^\pi],$$

without penalty functions, (as in [54, 38]) is as follows. When  $\phi_{t,j,a} = 0$ , Theorems 3.4.3, 3.5.2, and 3.5.6 show that nature's optimal move lies on the boundary of the uncertainty sets  $\mathcal{P}_{t,j,a}$ . Hence, nature's policy is strongly dependent on the design of these sets. Furthermore, the uncertainty sets are often obtained from a probabilistic model of the uncertain parameters (e.g., derived statistically from historical data) that allows for all "likely" values of the parameters (see, e.g., [92, 54, 38]). Typically, the more unlikely values lie at the boundary of the uncertainty set. As a result, the decision maker ends up assuming pessimistically that nature will choose very unlikely parameter values, leading to potentially conservative policies.

The construction in the previous subsection can mitigate both types of problems

(namely, dynamic inconsistency and conservatism) and serve as the basis for a better and more general robust control formulation than the one of [78, 54, 38]. When we consider non-zero penalty functions  $\phi_{t,j,a}$ , Equation (3.6.1) defines a dynamically consistent Markovian convex risk measure, which we can optimize efficiently over the set of randomized Markovian policies, and for which we have appealing structural results, when either the time horizon is finite, or when the discount factor  $\beta$  is one (cf. Sections 3.4 and 3.5). The problem

$$\inf_{\pi \in \Pi_{m,r}} \rho(C^\pi | s_1) = \inf_{\pi \in \Pi_{m,r}} \sup_{P \in \mathcal{P}(s_1)} (E_P[C^\pi | s_1] - \phi_{s_1}(P))$$

can be interpreted as a robust control formulation where the uncertain parameters are penalized, for example according to their “unlikeliness.” In contrast to the case where  $\phi_{t,j,a} = 0$ , nature’s optimal parameter choice need not be at the boundary of the uncertainty sets  $\mathcal{P}_{t,j,a}$ . As a result, the controller’s policy could be less conservative.

In the practical setting where only a prior distribution  $\mathbb{P}_{t,j,a}^0$  on the probability distribution  $P_{t,j,a}$  over  $(N_{t,j,a}, Q_{t,j,a})$ , is available, natural candidates for the penalty functions  $\phi_{t,j,a}$  could be  $\phi_{t,j,a}(P_{t,j,a}) = -\mathbb{P}_{t,j,a}^0(P_{t,j,a})$ , or  $\phi_{t,j,a}(P_{t,j,a}) = -\log \mathbb{P}_{t,j,a}^0(P_{t,j,a})$ .

### 3.6.3 From single-period risk measures to dynamically consistent Markovian multi-period risk measures

At the beginning of this section, we started from uncertainty sets  $\mathcal{P}_{t,j,a}$  and penalty functions  $\phi_{t,j,a}$  to construct dynamically consistent Markovian convex risk measures. The same procedure allows us to define dynamically consistent Markovian convex risk measures starting from single-period convex measures  $\bar{\rho}_{t,j,a}$  since these risk measures can be associated with sets of test measures  $\mathcal{P}_{t,j,a}$  and penalty functions  $\phi_{t,j,a}$ , thanks to the representation theorem 3.2.6. We will analyze such multi-period risk measures and argue that they are desirable counterparts to the single-period risk.

**Proposition 3.6.4.** *Let  $\bar{\rho}_{t,j,a}$  be convex risk measures on  $L^1(\mathcal{X} \times \mathbb{R}, \mathcal{F}_{j,a}^s(\omega_t), \mu_{|\omega_t}^{t,j,a})$  for all  $\omega \in \Omega$ , where  $\mathcal{F}_{j,a}^s(\omega_t)$  is the  $\sigma$ -field obtained as the restriction of the  $\sigma$ -field*

$\mathcal{F}^s(\omega_t)$  to the component  $(\nu_{t,j,a}, q_{t,j,a})$  of  $\omega$ . Let  $\mathcal{P}_{t,j,a}$  and  $\phi_{t,j,a}$ , respectively, be test probability measures and penalty functions associated with  $\bar{\rho}_{t,j,a}$  by Theorem 3.2.6, i.e.,

$$\rho_{t,j,a}(X) = \sup_{P \in \mathcal{P}_{t,j,a}} (E_P[X] - \phi_{t,j,a}(P)).$$

Then Equation (3.6.1) defines a dynamically consistent Markovian convex risk measure  $\rho$  on  $L^1(\Omega, \mathcal{F}, \mu)$ . If the risk measures  $\bar{\rho}_{t,j,a}$  are coherent, so is  $\rho$ ; and if they do not depend on  $t$ . i.e., for all  $t \geq 1$ ,  $\bar{\rho}_{t,j,a} = \bar{\rho}_{j,a}$ ,  $\rho$  is a stationary multi-period risk measure. Furthermore, if  $V(j) = \rho(C^\pi|j)$ , then the following results hold:

(a) When the time horizon  $T$  is finite, and the assumptions of Theorem 3.4.3 satisfied, the recursion therein becomes

$$V(T, j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{T,j,a}(Q_{T,j,a}), \quad j \in \mathcal{X}$$

$$V(t, j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{t,j,a} [Q_{t,j,a} + \beta V(t+1, N_{t,j,a})], \quad j \in \mathcal{X}, \quad t = 1, \dots, T-1.$$

(b) When the time horizon is infinite, the discount factor  $\beta$  is one, the single-period risk measures  $\bar{\rho}_{t,j,a}$  are constant equal to  $\bar{\rho}_{j,a}$  for all  $t$ , and the assumptions of Theorem 3.5.6 are satisfied, Bellman-Shapley equations therein take the form

$$V(j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{j,a}(Q_{t,j,a} + V(N_{1,j,a})).$$

(c) When the single-step risk measures  $\bar{\rho}_{t,j,a}$  are coherent and equal to  $\bar{\rho}_{j,a}$  for all  $t$ , and the assumptions of Theorem 3.5.6 are satisfied, Bellman-Shapley equations therein take the form

$$V(j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{j,a}(Q_{t,j,a} + \beta V(N_{1,j,a})).$$

*Proof.* The first part of the proposition is essentially Proposition 3.6.3.

If the  $\bar{\rho}_{t,j,a}$  are coherent ( $\phi_{t,j,a} = 0$ ) for all  $t, j, a$ , then the global penalty function  $\phi = 0$  and the induced risk  $\rho$  is coherent.

It is easy to see that  $\rho$  is stationary if the risk measures  $\bar{\rho}_{t,j,a}$  are constant over time.

The final part of the proposition amounts to checking that essentially  $\rho_{(t,j,a)} = \bar{\rho}_{t,j,a}$  for all  $t, j, a$ . Formally, these functions are defined respectively on  $L^1(F(\omega_t), \mathcal{F}^s(\omega_t), \mu|_{\omega_t})$  and  $L^1(\mathcal{X} \times \mathbb{R}, \mathcal{F}_{j,a}^s(\omega_t), \mu_{|\omega_t}^{t,j,a})$ , which are different spaces. But they coincide on the single-step position that depends only on  $(N_{t,j,a}, Q_{t,j,a})$ .  $\square$

This proposition suggests a “risk averse” or “robust” formulation based on a modification of Bellman’s equations. Fix a dynamical system with finite state and action space, and let  $\mu$  be a probability measure describing the likelihood of the trajectories of an MDP. To simplify the exposition, we will assume that the time horizon  $T$  is finite in this paragraph, but the infinite horizon case can be handled as well. The MDP whose probability law at time  $t$  from state-action pair  $(j, a)$  is given by the marginal  $\mu^{t,j,a}$  has a value function  $V$ , which satisfied Bellman’s equations [12]

$$V(t, j) = \min_{a \in \mathcal{A}_j} E_{\mu^{t,j,a}}[Q_{T,j,a}], \quad j \in \mathcal{X},$$

$$V(t, j) = \min_{a \in \mathcal{A}_j} E_{\mu^{t,j,a}}[Q_{t,j,a} + V(t+1, N_{t,j,a})], \quad j \in \mathcal{X}, \quad t = 1, \dots, T-1.$$

The value function gives the minimal expected cost of Markovian policies over the MDP described by  $\mu$ . If the expectation operators  $E_{\mu^{t,j,a}}$  in these equations are replaced by convex risk measures  $\bar{\rho}_{t,j,a}$  on  $\mathcal{X} \times \mathbb{R}$ , i.e.,

$$V(T, j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{T,j,a}[Q_{T,j,a}], \quad j \in \mathcal{X},$$

$$V(t, j) = \min_{a \in \mathcal{A}_j} \bar{\rho}_{t,j,a}(Q_{t,j,a} + V(t+1, N_{t,j,a})), \quad j \in \mathcal{X}, \quad t = 1, \dots, T-1,$$

then  $V(1, j)$  is the minimal risk  $\rho(C^\pi|j)$  over the Markovian policies, where  $\rho$  is a dynamically consistent Markovian convex induced by the single period risks  $\bar{\rho}_{t,j,a}$ . A risk minimizing policy is obtained (as usual) by selecting deterministically the minimizing actions in the above recursion.

### Illustration with the conditional value at risk

As an illustration of the result of this subsection on an infinite horizon discounted cost problem, let us consider a popular coherent risk measure conditional, namely the conditional value at risk (CVaR). It is well-known that CVaR, when considered as a multi-period risk measure, is not dynamically consistent.

Given (known) probability distributions  $p_{t,j,a}$  of the next state and associated transition cost  $(N_{t,j,a}, Q_{t,j,a})$  given the state-action pair  $(j, a)$  at time  $t$ , together with parameters  $\alpha_{j,a} \in [0, 1]$ , we can define the  $\alpha_{j,a}$ -CVaR of functions of  $(N_{t,j,a}, Q_{t,j,a})$ , denoted here by  $CVaR_{j,a}$  for the sake of notation. Let them play the role of the single-period risk measures, that is,  $\bar{\rho}_{t,j,a} = CVaR_{j,a}$ . As explained in Proposition 3.6.4, these single-period risk measures can be combined to define a dynamically consistent Markovian stationary coherent risk measure  $\rho$  on  $L^1(\Omega, \mathcal{F}, \mu)$ . The multi-period risk measure  $\rho$  is the natural multi-period counterpart of the single-period CVaRs.

Under the assumptions of Theorem 3.5.2,  $V^\pi(s_1) = \rho(C^\pi|s_1)$  is the unique solution in  $\mathbb{R}^{\mathcal{X}}$  of Bellman's equations,

$$V^\pi(j) = \sum_{a \in \mathcal{A}} \pi(a|j) CVaR_{j,a}(Q_{t,j,a} + \beta V^\pi(N_{t,j,a})).$$

Furthermore,  $V^*(j) = \inf_{\pi \in \Pi_{m,r}} V^\pi(j)$  is the unique solution in  $\mathbb{R}^{\mathcal{X}}$  of Shapley-Bellman's equations, that is,

$$V^*(j) = \min_{a \in \mathcal{A}_j} \alpha_{j,a} CVaR [Q_{t,j,a} + \beta V^*(N_{t,j,a})].$$

### 3.7 Conclusion

In this chapter, we motivate and define the notion of Markovian multi-period risk measure. This concept is key to finding a risk-minimizing policy when the time horizon is large, or even infinite. Moreover, it allows us to minimize a Markovian dynamically consistent convex risk measure of the sample cost by solving a zero-sum Markov game between the controller and nature. This correspondence can be further exploited to transfer results from game theory, in particular large-scale games, to the

problem of risk minimization.

We also show how to build multi-period risk measures that are dynamically consistent, Markovian, and convex, from single-period convex risk measures. This construction has better properties than the straightforward application of single-period risk measure to multi-period sample space, as illustrated with the conditional value at risk.

This chapter points out that the worst-case control of uncertain MDPs with rectangular uncertainty sets amounts to minimizing a Markovian dynamically consistent coherent risk measure of the sample cost, and thereby guarantees that the robust policies are sound from a decision-theoretic perspective. It also proposes an extension of the worst-case robust control of uncertain MDPs by adding a penalty to “unlikely” parameters in a principled fashion. This formulation has analogous structural and computational results, and it has the potential of generating less conservative policies.

An interesting research direction would be to come up with specific penalty functions that are well-motivated statistically and appealing to decision makers, and to study numerically the benefits of using penalized worst-case formulation in practical problems of sequential decision under uncertainty.

## 3.8 Glossary

### 3.8.1 Definitions

Convex risk measure	Definition 3.2.1, p. 68
Fatou property	Definition 3.2.9, p. 76
Multi-period risk measure	Definition 3.3.3, p. 85
Dynamically consistent multi-period risk measure	Definition 3.3.6, p. 86
Markovian multi-period risk measure	Definition 3.3.9, p. 88
Stationary Markovian risk measure	Definition 3.5.1, p. 100

### 3.8.2 Notations

Notations related to the model description in Subsection 3.3.1	
$\omega \in \Omega$	sample point in sample space
$\omega' \in \Omega'$	partial history in set of partial histories
$\omega_t \preceq \omega$	partial history of $\omega$ , of length $t$ , up to $r_{t,i}$
$\omega_t^+ \preceq \omega$	partial history of $\omega$ , of length $t$ , up to $(\nu_{t,i,a}, q_{t,i,a})$
$\mathcal{E}(\omega')$	set of possible single-step continuations after the partial history $\omega'$
$F(\omega') \subset \Omega$	set of sample points starting with partial history $\omega'$
$\mathcal{F}(\omega')$	$\sigma$ -field on $F(\omega')$
$\mathcal{F}^s(\omega')$	$\sigma$ -field, on $F(\omega')$ , of events realized with the immediate continuation of $\omega'$
$\mu$	reference probability measure on $(\Omega, \mathcal{F})$
$\nu_{t,i,a} \in \mathcal{X}$	state following the state-action pair $(i, a)$ at time $t$
$q_{t,i,a} \in \mathbb{R}$	cost associated with state-action pair $(i, a)$ at time $t$
$r_{t,i} \in \mathbb{R}$	number to randomize action choice in state $i$ at time $t$
$\pi$	policy
$\mathcal{A}_i$	set of available action in state $i$
$\Delta_i$	probability simplex over $\mathcal{A}_i$
$S_t^\pi \in \mathcal{X}$	random state occupied at time $t$ under policy $\pi$
$A_t^\pi \in \mathcal{A}$	random action chosen at time $t$ by policy $\pi$
$Q_t^\pi \in \mathbb{R}$	random cost incurred by policy $\pi$ at time $t$
$C_t^\pi \in \mathbb{R}$	tail sample cost starting at time $t$ of policy $\pi$
$C_{t,j,a}^\pi \in \mathbb{R}$	tail sample cost starting at time $t$ of policy $\pi$ from the state-action pair $(j, a)$
$T_{t,j,a}^\pi$	Trajectory of the system following policy $\pi$ and initialized in $(j, a)$ at time $t$ .
$\mathcal{G}_{t,j,a}$	Set of positions of the form $f(T_{t,j,a}^\pi)$ for policy $\pi \in \Pi_{m,r}$ .
Notations for risk measures	
$\rho(X)$	risk measure of position $X$
$\phi(P)$	penalty associated with the parameter $P$
$\rho(\cdot \omega')$	risk measure conditional on the partial history $\omega'$
$\rho_{\omega'}(\cdot)$	risk measure conditional on the partial history $\omega'$ on single-step positions
$\rho(X \omega')$	single-step position taking the value of the conditional risk $\rho(X \omega'V)$



# Chapter 4

## Data-driven approach to Markov Decision Processes

### 4.1 Introduction

When the dynamic aspect of a decision problem plays an important role, MDPs offer an appealing modeling framework. If an underlying MDP model was given or if there were enough data available to calibrate accurately such a model, dynamic programming methods could compute the optimal expected performance and an optimal decision rule. However, in many practical applications, there is no accurate MDP model available to the decision maker. This situation is very common in the social and medical sciences where, oftentimes, experts have little mechanistic insight about the phenomenon of interest. (We will see specific examples in the next subsection). On the other hand, building an MDP model from data, and a fortiori an uncertain MDP model, requires a fair amount of observations. In order to avoid the curse of uncertainty of uncertain MDPs (cf. Chapter 2) and in order to deal with the case where there is insufficient data to even attempt a model estimation, quantitative methods have to exploit directly system observations in order to gain insights into the problem. In the present chapter, we will tackle the problem of estimating the expected performance of a given policy (and its gradient) from a training set comprising observed trajectories sampled under a known policy. We will seek good estimators,

in the sense that they should be unbiased and have the lowest possible training set to training set variance <sup>1</sup>.

### 4.1.1 Motivating examples

The following two examples, one from marketing and one from medical decision making, will serve as motivation for our work in this chapter. We introduce them briefly now, but we will come back to them in this chapter’s conclusion to see how our findings apply to them. These examples have three features that are fundamental for our approach:

1. the sampling policy under which the observations were made is known to the estimator,
2. the sampling policy explores different actions by randomizing its action choice,
3. the number of observed trajectories is sufficient for a “good” estimator to have low training set to training set variance.

#### Catalog mailing problem

In his dissertation on the catalog mailing problem [86] and subsequent work [83], Sun observed that: “catalog firms mailed almost 17 billion catalogs in 2000. Printing and mailing these catalogs is the second largest expense in the industry (behind the cost of the goods), representing approximately 20% of net sales. As a result, catalog managers view improving their policies for deciding who should receive mail catalogs as one of their highest priorities.”

The catalog mailing policies of firms are mostly myopic: they mail catalogs to customers who they believe will be profitable in the short-term, neglecting the long-term effect on customer relationship of advertising. In the references [86, 83], an optimal dynamic catalog mailing policy that factors in the long-term dynamic effects

---

<sup>1</sup>An estimator maps an observed training set to an estimate. The training set to training set variance of an estimator is the variance of the estimate for a randomly chosen training set. Intuitively, an estimator with a low training set to training set variance yields estimates that typically vary little from one random training set to another.

of the mailing decisions is estimated from mailing companies' data. Specifically, an MDP model was constructed from the historical data of a catalog mailing company. The MDP states capture the customer's status, and the actions at each period are either to mail a catalog or not. First, a state space comprising 500 states was built by segmenting the customer status according to recorded customer characteristics that are widely recognized by the industry to influence customer behavior. For example, these characteristics include the purchase recency, frequency, and monetary expense of a customer.

Subsequently, an MDP model was calibrated under the key assumption that the catalog mailing company did not use any information that is not captured in the available data when choosing its mailing decisions. Otherwise, the results could be plagued by *attribution bias* (cf. the medical example for more details on attribution bias). The dynamics and reward parameters of the model were estimated using a random sample of 100,000 customers.

Finally, a dynamic mailing policy was optimized by dynamic programming using the estimated MDP model and was implemented in a field test. The field test revealed that the predicted value of the optimized mailing policy suffered from some bias and variance [44].

Assuming that the historical mailing policy of the company is known, the approach proposed in this chapter has the promise of providing unbiased and lower variance estimators of the value of different catalog mailing policies. Thus, our work may enable a better use of the available data to design more profitable dynamic mailing policies.

### **Medical decision making**

Since the seminal paper of Beck and Pauker [10] in 1983, Markov chains have been used in the medical decision making literature to model the dynamic effects of medical treatments. They are particularly useful in modeling health conditions where the timing of events is important and when important events might happen at random times or multiple times. For example, Markov models are very convenient to model

chronic diseases like depression [63, 62, 64], or to model on-going risks such as the risk of hemorrhage while on anticoagulant therapy, the risk of rupture of aortic aneurysm, or the risk of mortality (cf. the review [84] and references therein). Since the early eighties, Markov processes have been widely used to model diverse outcomes such as life expectancy, quality-adjusted life expectancy (e.g., [17, 85]), or cost-effectiveness (e.g., [35, 105]). Markov Decision Processes have also been suggested as an approach to optimize sequential medical decisions (cf. [80] for a review and references therein).

In many cases, the Markov models are simple, comprising a handful of states, and calibrated from experts' opinions and the medical literature; yet they have generated valuable medical insights. When models are calibrated using past system observations, the analysis can suffer from significant *attribution bias* [19], especially in the medical context. Indeed, most of the data are obtained from clinical data, in which health experts have selected among different treatment alternatives based on potentially unrecorded characteristics of the patient's health condition. For example, consider a practice that gives treatment  $A$  to the acute cases of a medical condition, while treatment  $B$  is given to the milder cases. If the condition acuteness of patients is not recorded, a data analysis might suggest that treatment  $B$  yields better health outcomes than treatment  $A$ , although this outcome need not be explained by the relative effectiveness of the two treatments, but by the biased allocation of patients to the two treatment options.

Randomized experiments, a.k.a. randomized controlled trials (RCT) in the biomedical literature, offer a solution to attribution bias. The experimental protocol of RCT studies makes sure that the patients are randomly assigned different treatments in order to avoid attribution bias, and thus RCTs are recognized as the premier objective comparison of treatments in medicine. There are hundreds of RCTs documented in the medical literature, but, most important for the motivation of our work, there are a growing number of multi-period RCTs. For example, the Clinical Antipsychotic Trials of Intervention Effectiveness (CATIE) study is a multi-period RCT comprising 1600 patients with schizophrenia to evaluate the clinical effectiveness of the combination/succession of nine potential drugs in the treatment of schizophrenia and

Alzheimer's disease (<http://www.catie.unc.edu/>). We will now describe in more detail an important multi-period RCT for depression treatments.

### **A multi-period Randomized Controlled Trial for depression - STAR\*D**

The Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) Study [55, 56, 57, 96, 95, 75] (<http://www.edc.pitt.edu/stard/>) is a 4000-patient randomized clinical trial to evaluate the effectiveness of different treatments for people with major depressive disorder.

Major depressive disorder is a recurring and chronic illness affecting each year 9.5 percent of the American population, or about 20.9 million adults, according to the National Institute of Mental Health [55]. "About 10 percent of men and up to 25 percent of women will experience depression in their lifetime. Depression is currently the fourth most disabling illness worldwide and is responsible for up to 70 percent of psychiatric hospitalizations and about 40 percent of suicides. As a result, the cost of depression in the United States was estimated to be \$83 billion in the year 2000."

Since only one third of the patients treated with a standard antidepressant become symptom free, some combination or succession of treatments is needed. The main goal of the STAR\*D study was to identify the best next steps for those people with depression who need to try more than one treatment. That is, which treatment strategies are the most effective for people that do not become symptom-free after one or more treatments?

The STAR\*D study comprises four stages, out of which we will describe the first two in some detail, to give a concrete illustration of the study's protocol and results. In the first stage [96, 56], 2,876 patients with depression were treated with the antidepressant citalopram for approximately three months. At the end of the first stage, about one third of the participants became symptom free ("remission"), while nine percent of participants stopped the medication because of intolerable side effects. Out of the two third of unsuccessfully treated patients in the first stage, 1,439 participated in the next stage. In the second stage [95, 75, 57], one of three options was selected:

1. switch the medication to other antidepressants (bupropion, sertraline, or ven-

lafaxine),

2. augment the current intake of citalogram with bupropion or buspirone,
3. switch to cognitive therapy, or add it to citalogram.

Depending on the outcome of the second phase treatment, patients went through a third and sometimes a fourth phase. But there were so few patients left in the fourth phase that we can focus on the outcome of the first three phases.

An accurate probabilistic model of the STAR\*D study is beyond the scope and focus of this chapter, especially because of the censoring of some participants and their rejection of some treatments. Nonetheless, a stylized model of the STAR\*D study provides a sufficient motivation for our work. We could model a patient's mental health status by an MDP. Its states are the patient partial histories and its actions are the available treatments in each state. Since all patients receive the same treatment in the first phase, the outcome of the first phase can be modeled as the initial state for the decision making problem. Hence, if we focus on Phases I to III of the study, we are dealing with a two-period MDP problem.

As a result, assessing the effectiveness of specific dynamic treatment regimes for depression in STAR\*D amounts to evaluating its value on a two-period MDP for which we observed approximately 1,500 trajectories.

### 4.1.2 Literature review

In the absence of a model of the system at hand, the performance of a candidate policy needs to be estimated from observed trajectories, which might have been sampled under another policy. Importance sampling, a.k.a. likelihood ratio estimation, (e.g. [30]) is a well-suited approach for such off-policy estimation. Importance sampling has been used to reduce the variance of simulation-based estimation by optimizing the sampling policy (e.g., [81]), and to estimate efficiently the gradient by simulation (e.g., [42, 29] and the references therein). When the sampling policy is adaptive, further variance reduction can be achieved, sometimes yielding a zero variance estimator [39,

33]. In our case, the sampling policy is not adaptive and not subject to optimization. Hence, importance sampling is simply a way to obtain unbiased off-policy estimators of the system’s performance [103, 65].

In this chapter, we use a general control variate approach [41, 43, 52] to reduce the variance of our importance sampling estimator and obtain new optimized estimators for a policy’s value and value gradient. Interestingly, this approach also sheds light on various variance reduction techniques found in reinforcement learning, especially in the context of value gradient estimation, as in Greensmith et al [31]. We will compare thoroughly our approach and the ones described in their paper, in Section 4.5.

The two problems of value and value gradient estimation, in the absence of an MDP model, are thoroughly investigated in the reinforcement learning literature. Most value estimation methods rely on Bellman’s equations. Q-learning [100, 101] estimates the value function using system trajectories, instead of a model. The related approach of temporal differences,  $TD(\lambda)$ , was first proposed by Sutton [87] and then extensively studied. The aforementioned approaches were subsequently combined with value function approximation in order to cope with the curse of dimensionality. In some cases, these methods (or adaptations of them) have been proved to converge and have some desirable approximation properties, e.g., [73, 98, 8]. In this chapter, we will also rely on function approximation methods and even temporal difference algorithms, but rather as a way to regularize the estimation of high-dimensional quantities compared to the size of the training set. Unlike temporal difference methods with function approximation, our estimators are always unbiased, even when the approximation architecture is poor.

The estimation of the value gradient has been extensively studied in the context of policy search methods [104, 45, 31, 103, 88, 14]. In this body of work, it has been observed that the estimation of the performance gradient can have high variance. A suggested approach to reduce its variance is to add to the gradient estimate a baseline, which does not add bias but impacts the estimator’s variance (e.g., [104, 45, 31, 88]).

Another class of methods that aim to reduce the variance of gradient estimates in policy search are the so-called actor-critic methods, where an estimate of the value

function is used to reduce the variance of the performance gradient [9, 40, 14, 103, 31, 89]. If the value function approximation belongs to an appropriate space tailored to the policy space, actor-critic methods provide an unbiased estimate of the performance gradient. However, it is not possible in practice to guarantee that the approximation architecture is adequate, and thus that this method does not introduce bias.

In a similar fashion to actor-critic methods, Henderson and Meyn [33] use approximate solutions to the Poisson equation (obtained from a quadratic and a fluid approximation of the value function) to reduce the variance of the estimated steady-state performance of queueing networks, without introducing a bias. However, their approach requires the knowledge of the underlying Markov model. (cf. Subsection 4.3.1 for further discussion on this paper).

Generally, the variance reduction techniques from the reinforcement learning literature exploit ad hoc ideas. Our approach, in Section 4.3, will allow us to unify and generalize some ideas seen in reinforcement learning.

In the biostatistics community, Robins studied how to evaluate the effect of a *dynamic treatment regime* (i.e., policy) from *observational data* (in the statistical literature, observational data refer specifically to data obtained from a possibly “biased” sampling policy) [69, 70]. He also uses importance sampling to perform off-policy value estimation, while controlling for attribution bias.

In order to estimate the mean response to dynamic treatment regimes using observed trajectories without dealing with attribution bias, the assumption of no unmeasured confounders (or sequential randomization (SR)) is handy [51, 68, 48]. Intuitively, this assumption says that the patients who receive a treatment at some time point conditional on some recorded information are not statistically different from other patients in terms of unobserved determinants of their conditions. Thus, this assumption is fundamental to justify the use of Markov models in conjunction with observational data. However, the assumption SR cannot be checked from observational data. Nonetheless, it can be enforced by experimental design, for example if the treatments are sequentially randomized during data collection. Under the as-

sumption SR, the estimation of the mean response to a dynamic treatment regime is equivalent to our problem. Similar to our unconstrained optimal estimator, the paper [51] provides a minimum variance estimator for the value of a given policy under the assumption SR.

Murphy [48] went further by showing that an optimal dynamic treatment regime can be estimated from observed trajectories under the assumption SR provided that the form of the “advantages” (as this word is used in reinforcement learning) is known, and Robins [68] identified a minimum variance estimator of the optimal policy under the same assumptions.

### 4.1.3 Contributions

In order to estimate a given policy’s value (or value gradient) from a training set comprising trajectories observed under a known sampling policy, we combine an importance sampling estimation method and a control variate approach to variance reduction. Our estimation procedure is based on estimators with the lowest training set to training set variance in two broad classes of unbiased estimators, namely an unconstrained and a constrained class.

In the unconstrained case, a minimum variance estimator can be characterized as the projection of a naive estimator on the set of random variables with zero action innovations. Alternatively, we provide an algebraic expression for the optimal estimator. In the case of value estimation, our estimator is the same as the minimum variance value estimator (5.3) from [51].

Similarly, in the constrained case, we characterize theoretically the optimal constrained estimators for the value and value gradient.

Our optimal value estimators require the knowledge of the “advantage” of a state-action pair, and not its Q-factor - an important nuance in some applications. Unfortunately, the advantages are unknown in practice and the advantages (or the Q-factors) need to be guessed or estimated. A salient feature of our approach is that the estimators remain unbiased for any guess and for all possible underlying MDP models (unlike the related papers [48, 68], which have to know the true form of the advantages

in order to estimate the *optimal policy*).

Since the constrained class of estimators is a subset of the unconstrained one, the best constrained estimator has higher variance than the best unconstrained estimator in theory. However, they are valuable in practice when the Q-factors need to be approximated, in particular when

- the Q-factors cannot be handled efficiently by a computer because of the excessive size of the state-action space (the curse of dimensionality in reinforcement learning),
- the training set is insufficient to build accurate estimates of the Q-factors so that regularized estimates are better.

Our approach can be expected to outperform standard reinforcement learning methods in the latter case, because the sound statistical principles of our approach exploit the available data more efficiently. In addition, our approach is less affected by the choice of a poor approximation architecture than Temporal Difference methods. We corroborate these claims by comparing numerically the practical performance of our different approaches for value estimation with competitive algorithms from the literature.

#### 4.1.4 Chapter structure

This chapter is organized as follow. In the next section, the mathematical formulation of the estimation problem is given. In Section 4.3, we introduce the concept of innovations, which is useful for our subsequent analysis. In Sections 4.4 and 4.5, we respectively characterize the optimal estimator for the value and the value gradient.

## 4.2 Problem formulation

In this section, we introduce the mathematical problem formulation, starting with a generative model for the underlying system and the observation mechanism, and concluding with the problem of efficient estimation.

### 4.2.1 Probabilistic model of the system

To avoid technicalities, we will assume that the true underlying model is an MDP with finite state and action spaces. Let  $\mathcal{S}$  be a finite state space and  $\mathcal{A}$  be a finite action space, and define the sample space  $\Omega = (\mathcal{S} \times \mathcal{A} \times \mathbb{R})^T$  where  $T$  is a finite time horizon. We endow  $\Omega$  with the product sigma-algebra  $\mathcal{F}$  generated by  $(2^{\mathcal{S}} \times 2^{\mathcal{A}} \times \mathcal{B})^T$ , where  $\mathcal{B}$  is the Borel  $\sigma$ -field of  $\mathbb{R}$ .

We will denote random variables with upper-case letters and their realizations with lower-case. Let  $S_t$  be the random state in  $\mathcal{S}$  occupied by the system,  $A_t$  the random action in  $\mathcal{A}$  chosen at time  $t$ , and  $R_t$  the associated random reward. A random trajectory of the system is realized sequentially in the following order:  $S_1, A_1, R_1, S_2, A_2, \dots, S_T, A_T, R_T$ .

We assume that the controller selects its actions according to a *known* Markovian sampling policy  $\nu$ . It is characterized by the conditional probabilities  $\nu_t(a|s)$  of choosing action  $a$  in state  $s$  at time  $t$ . We assume that the sampling policy  $\nu$  gives non-zero probability to all actions in all states, i.e.  $\nu_t(a|s) > 0$  for all  $a \in \mathcal{A}, s \in \mathcal{S}$ . Alternatively, we could restrict the action space to the actions that are selected with non-zero probability by the sampling policy  $\nu$  or that are observed in the data.

The initial state distribution  $\eta$ , the true MDP model  $K$ , which comprises the probabilistic description of the state dynamics  $K^d$  and of the reward  $K^r$ , and the sampling policy  $\nu$  induce a probability measure  $\mu$  on the sample space  $\Omega$  that describes the likelihood of each trajectory  $(S_1, A_1, R_1, S_2, A_2, \dots, S_T, A_T, R_T)$  by

$$\begin{aligned} \mu(S_1 = s_1, A_1 = a_1, R_1 \in \mathfrak{R}_1, \dots, S_T = s_T, A_T = a_T, R_T \in \mathfrak{R}_T) \\ = \eta(s_1) \prod_{t=1}^T \nu(a_t|s_t) K^d(s_{t+1}|t, s_t, a_t) K^r(R_t \in \mathfrak{R}_t|t, s_t, a_t, s_{t+1}). \end{aligned}$$

We assume that  $K^r(\cdot|t, s_t, a_t, s_{t+1})$  has finite second moment for all  $t, s_t, a_t, s_{t+1}$ , i.e.,  $E[R_t^2] < +\infty$ .

**Remark 4.2.1.** *The above factored form of  $\mu$  implies that the reward  $R_t$  at time  $t$  is independent of the past, given the state-action pair  $(S_t, A_t, S_{t+1})$ . This assumption can*

be enforced by redesigning the model's state space so that it includes all the information that influences the reward.

More generally, the state space design includes all the relevant information to the system so that the future of the system's trajectory is independent from its past given its current state. In some cases, it is necessary to include all the past history of the system in the state space. In particular, we can model a partially observed MDP (POMDP) as an MDP with the conditional state occupation probability given the past history as state.

The initial state distribution  $\eta$  and the underlying MDP model  $K$  are unknown to the decision maker but are assumed to be fixed. Hence, when we talk about the performance of any policy, we will always refer to its performance on the MDP model  $K$  with initial state distribution  $\eta$ . We will denote by  $E_\nu$  the expectations with respect to  $\mu$  to highlight the dependency on policy  $\nu$ , in contrast to  $E_\theta$ , which denotes the expectation with respect to the probability measure induced by  $\eta, K$  and policy  $\theta$ . When the context is clear, we might write  $E$  instead of  $E_\nu$  for conciseness. We define the value of policy  $\theta$  as  $V_\theta = E_\theta[R_1 + \dots + R_T]$ .

## 4.2.2 Observed data

A training set  $\mathcal{T} = \{(s_1^k, a_1^k, r_1^k, \dots, s_T^k, a_T^k, r_T^k), k = 1, \dots, n\}$  comprising  $n$  IID trajectories observed under the sampling policy  $\nu$  is an element of  $\Omega^n$  sampled according to the product probability measure  $\mu^n$  defined by

$$\mu^n(\{(s_1^k, a_1^k, \mathfrak{R}_1^k, \dots, s_T^k, a_T^k, \mathfrak{R}_T^k), k = 1, \dots, n\}) = \prod_{k=1}^n \mu(s_1^k, a_1^k, \mathfrak{R}_1^k, \dots, s_T^k, a_T^k, \mathfrak{R}_T^k).$$

The expectation with respect to  $\mu^n$  of an integrable function of the random training set  $\mathcal{T}$  will also be denoted  $E_\nu$  or  $E$ . For conciseness, we will use bold letters,  $\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T$ , to refer to the vector of the  $n$  copies in the training set  $\mathcal{T}$  of  $S_1, A_1, R_1, \dots, S_T, A_T, R_T$ , respectively. For example, we have  $\mathbf{S}_1 = (S_1^1, \dots, S_1^n)$ .

Let  $\mathcal{E}$  be the space of measurable and square integrable functions of the variables  $(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T)$  and  $\mathcal{E}_t$  be the space of square integrable measurable

functions of the variables  $(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_t, \mathbf{A}_t)$ , which we will identify with a subset of  $\mathcal{E}$ . We endow  $\mathcal{E}$  with the Euclidean norm induced by the scalar product defined, for all  $f, g \in \mathcal{E}$ , by

$$\langle f, g \rangle = E_\nu[fg] = \int_{\Omega^n} f(\omega^1, \dots, \omega^n)g(\omega^1, \dots, \omega^n)d\mu^n.$$

For any event  $F \in \mathcal{F}$ , we let  $N_F$  be the random number of times that the event  $F$  is observed in the training set  $\mathcal{T}$ , e.g., we will denote by  $N_{s_1} = \sum_{k=1}^n \mathbf{1}_{\{S_1^k = s_1\}}$  the number of times the initial state of the trajectories of  $\mathcal{T}$  is  $s_1$ . Given a training set  $\mathcal{T} = \{(s_1^k, a_1^k, r_1^k, \dots, s_T^k, a_T^k, r_T^k), k = 1, \dots, n\}$ , we define an empirical probability measure  $\mathbb{P}_n$  by

$$\begin{aligned} \mathbb{P}_n(S_1 = s_1, A_1 = a_1, R_1 \in \mathfrak{R}_1, \dots, S_T = s_T, A_T = a_t, R_T \in \mathfrak{R}_T) \\ = \hat{\eta}(s_1) \prod_{t=1}^T \hat{\nu}_t(a_t | s_t) \mathbb{K}_n^d(s_{t+1} | t, s_t, a_t) \mathbb{K}_n^r(R_t \in \mathfrak{R}_t | t, s_t, a_t, s_{t+1}), \end{aligned}$$

where

$$\begin{aligned} \hat{\eta}(s_1) &= \frac{N_{s_1}}{n}, \\ \hat{\nu}_t(A_t = a_t | s_t) &= \frac{N_{s_t, a_t}}{N_{s_t}}, \\ \mathbb{K}^d(S_{t+1} = s_{t+1} | t, s_t, a_t) &= \frac{N_{s_t, a_t, s_{t+1}}}{N_{s_t, a_t}}, \\ \mathbb{K}^r(R_t \in \mathfrak{R}_t | t, s_t, a_t, s_{t+1}) &= \frac{N_{s_t, a_t, \mathfrak{R}_t, s_{t+1}}}{N_{s_t, a_t, s_{t+1}}}. \end{aligned}$$

When the denominator in the right-hand side of these equations is zero for some  $(t, s_t, a_t, s_{t+1})$ , we let the left-hand side be an arbitrary probability distribution. Any choice yields the same joint probability distribution  $\mathbb{P}_n$  with probability one, namely

$$\begin{aligned} & \mathbb{P}_n(S_1 = s_1, A_1 = a_1, R_1 \in \mathfrak{R}_1, \dots, S_T = s_T, A_T = a_T, R_T \in \mathfrak{R}_T) \\ &= \begin{cases} 0 & \text{if } N_{s_1} \dots N_{s_T} = 0 \\ \frac{N_{s_1, a_1, \mathfrak{R}_1, s_2}}{n} \dots \frac{N_{s_{T-1}, a_{T-1}, \mathfrak{R}_{T-1}, s_T}}{N_{s_{T-1}}} \cdot \frac{N_{s_T, a_T, \mathfrak{R}_T}}{N_{s_T}} & \text{otherwise} \end{cases} \end{aligned}$$

This definition of  $\mathbb{P}_n$  ensures that it has the Markov property with respect to  $\mathcal{S}$ . We will denote  $\mathbb{E}_n[X] = \int X d\mathbb{P}_n$  the empirical expectation of a random variable  $X$  on  $(\Omega, \mathcal{F})$ , when it is well-defined.

**Remark 4.2.2.** *Observe that we form an empirical estimate  $\hat{\nu}_t(A_t = a_t | s_t)$  of the known sampling probability  $\nu_t(a_t | s_t)$  to define  $\mathbb{P}_n$ . If we replace  $\hat{\nu}_t(A_t = a_t | s_t)$  by  $\nu_t(a_t | s_t)$ , the probability measure  $\mathbb{P}_n$  is more arbitrary. Indeed, if an action chosen with positive probability by  $\nu$  is not observed in the data, the resulting empirical distribution  $\mathbb{P}_n$  would be arbitrary. As a result,  $\mathbb{P}_n(F)$  would not provide unbiased estimates of  $\mu(F)$  for all events  $F$ , as we will prove in Lemma 4.2.4, but only an asymptotically unbiased estimator.*

**Remark 4.2.3.** *Observe that we do not define the empirical distribution to be*

$$\begin{aligned} & \tilde{\mathbb{P}}_n(S_1 = s_1, A_1 = a_1, R_1 \in \mathfrak{R}_1, \dots, S_T = s_T, A_T = a_T, R_T \in \mathfrak{R}_T) \\ &= \frac{1}{n} \sum_{k=1}^n \mathbf{1}_{\{s_1^k = s_1, a_1^k = a_1, r_1^k \in \mathfrak{R}_1, \dots, s_T^k = s_T, a_T^k = a_T, r_T^k \in \mathfrak{R}_T\}}, \end{aligned}$$

*which is non-Markov. Nonetheless, most of the analysis of this chapter applies directly to the case where the empirical measure is  $\tilde{\mathbb{P}}_n$ . When relevant, we will discuss how the results would be modified if the empirical measure was  $\tilde{\mathbb{P}}_n$ , instead of  $\mathbb{P}_n$ . We will also compare these two possible definitions of the empirical distribution shortly.*

The following lemma justifies the use of  $\mathbb{P}_n$  as an empirical distribution approximating  $\mu$ .

**Lemma 4.2.4.** *Let  $f$  be a square integrable function on  $(\Omega, \mathcal{F}, \mu)$  and let*

$$Z(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T) = \mathbb{E}_n[f(S_1, A_1, R_1, \dots, S_T, A_T, R_T)].$$

*Then,  $Z$  is in  $\mathcal{E}$  and its expectation is*

$$E[\mathbb{E}_n[f(S_1, A_1, R_1, \dots, S_T, A_T, R_T)]] = E[f(S_1, A_1, R_1, \dots, S_T, A_T, R_T)].$$

*Furthermore, if the function  $f$  is only a function of the state-action pair  $(S_t, A_t)$  at time  $t$ ,*

$$\begin{aligned} E[\mathbb{E}_n[f(S_t, A_t)]^2] &= E[f(S_t, A_t)]^2 + \frac{1}{n} \text{var}(f(S_t, A_t)), \\ \text{var}(\mathbb{E}_n[f(S_t, A_t)]) &= \frac{1}{n} \text{var}(f(S_t, A_t)). \end{aligned} \quad (4.2.1)$$

*Proof.* For any finite  $n$ , it is easy to see that  $\mathbb{E}_n[f] \in \mathcal{E}$ . Thus, the integral of  $\mathbb{E}_n[f]$  is well-defined and finite.

We will show that  $E[\mathbb{E}_n[f]] = E[f]$  by a classic approach of proving the results for the indicator function of all sets in a  $\pi$ -system of  $(\Omega, \mathcal{F})$ . Then we will conclude by the monotone class theorem ([102], Theorem 3.14) that it is true for all measurable functions.

First, fix a measurable set  $F$  of the form  $F = (s_1, a_1, \mathfrak{R}_1, \dots, s_T, a_T, \mathfrak{R}_T) \in \mathcal{F}$ , where the sets  $\mathfrak{R}_t$  are Borel sets in  $\mathbb{R}$  for  $t = 1, \dots, T$ . Since the system model and the sampling policy generating the training sets are Markovian, we have

$$\begin{aligned} &E[\mathbb{P}_n(F)] \\ &= E\left[\hat{\eta}(s_1) \prod_{t=1}^T \hat{\nu}_t(a_t|s_t) \mathbb{K}_n^d(s_{t+1}|t, s_t, a_t) \mathbb{K}_n^r(\mathfrak{R}_t|t, s_t, a_t, s_{t+1})\right] \\ &= E\left[\hat{\eta}(s_1) \prod_{t=1}^T E[\hat{\nu}_t(a_t|s_t) | \mathbf{S}_t] E[\mathbb{K}_n^d(s_{t+1}|t, s_t, a_t) | \mathbf{S}_t, \mathbf{A}_t] E[\mathbb{K}_n^r(\mathfrak{R}_t|t, s_t, a_t, s_{t+1}) | \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}]\right]. \end{aligned}$$

To conclude that  $E[\mathbb{P}_n(F)] = \mu(F)$ , it suffices to show that

$$\begin{aligned} E[\hat{\eta}(s_1)] &= \eta(s_1), \\ E[\hat{\nu}_t(A_t = a_t | s_t) | \mathbf{S}_t] &= \nu_t(A_t = a_t | s_t), \\ E[\mathbb{K}_n^d(s_{t+1} | t, s_t, a_t) | \mathbf{S}_t, \mathbf{A}_t] &= K^d(s_{t+1} | t, s_t, a_t), \\ E[\mathbb{K}_n^r(\mathfrak{R}_t | t, s_t, a_t, s_{t+1}) | \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}] &= K^r(\mathfrak{R}_t | t, s_t, a_t, s_{t+1}). \end{aligned}$$

There are two cases to consider. When, given the conditional information,  $N_{s_t} > 0$ ,  $N_{s_t, a_t} > 0$ , and  $N_{s_t, a_t, s_{t+1}} > 0$ , respectively, we have

$$\begin{aligned} E[\hat{\eta}(s_1)] &= E\left[\frac{N_{s_1}}{n}\right] = \eta(s_1), \\ E[\hat{\nu}_t(A_t = a_t | s_t) | \mathbf{S}_t] &= E\left[\frac{N_{s_t, a_t}}{N_{s_t}} | \mathbf{S}_t\right] = \nu_t(A_t = a_t | s_t), \\ E[\mathbb{K}_n^d(s_{t+1} | t, s_t, a_t) | \mathbf{S}_t, \mathbf{A}_t] &= E\left[\frac{N_{s_t, a_t, s_{t+1}}}{N_{s_t, a_t}} | \mathbf{S}_t, \mathbf{A}_t\right] = K^d(s_{t+1} | t, s_t, a_t), \\ E[\mathbb{K}_n^r(\mathfrak{R}_t | t, s_t, a_t, s_{t+1}) | \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}] &= E\left[\frac{N_{s_t, a_t, \mathfrak{R}_t, s_{t+1}}}{N_{s_t, a_t, s_{t+1}}} | \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}\right] = K^r(\mathfrak{R}_t | t, s_t, a_t, s_{t+1}). \end{aligned}$$

When a denominator in the above expressions is zero, the corresponding empirical conditional probability is in fact arbitrary, as we observed earlier. Since this arbitrary choice has no effect on the empirical probability  $\mathbb{P}_n(F)$ , it has no effect on  $E[\mathbb{P}_n(F)]$ . Therefore, we can assume in this proof that these (arbitrary) empirical conditional probabilities are equal to the true conditional probability. For example, if  $N_{s_t} = 0$ , then  $\hat{\nu}_t(A_t = a_t | s_t)$  can be chosen (arbitrarily) to be equal to  $\nu_t(a_t | s_t)$ . This concludes the proof that

$$E[\mathbb{P}_n(s_1, a_1, \mathfrak{R}_1, \dots, s_T, a_T, \mathfrak{R}_T)] = \mu(s_1, a_1, \mathfrak{R}_1, \dots, s_T, a_T, \mathfrak{R}_T).$$

Now, we verify that the assumptions of the monotone class theorem ([102], Theorem 3.14) are satisfied with the set

$$\mathcal{H} = \{f \in L^2(\Omega, \mathcal{F}, \mu) \mid E[\mathbb{E}_n[f]] = E[f]\}.$$

- It is easy to see that  $\mathcal{H}$  is a vector space, which contains the constant 1.
- If  $f_k \in \mathcal{H}$  are such that  $f_k \uparrow f$  with  $f$  bounded, then  $\mathbb{E}_n[f_k] \uparrow \mathbb{E}_n[f]$  for all training sets by the monotone convergence theorem. Moreover,  $E[\mathbb{E}_n[f_k]] \uparrow E[f]$  and  $E[f_k] \uparrow E[f]$ . Consequently, we have  $f \in \mathcal{H}$ .
- Finally, we proved above that  $\mathbf{1}_F \in \mathcal{H}$  for all  $F$  of the form  $F = (s_1, a_1, \mathfrak{R}_1, \dots, s_T, a_T, \mathfrak{R}_T) \in \mathcal{F}$ , where the sets  $\mathfrak{R}_t$  are Borel sets in  $\mathbb{R}$  for  $t = 1, \dots, T$ . These sets form a  $\pi$ -system for  $(\Omega, \mathcal{F})$ , i.e., the  $\sigma$ -field generated by these sets is  $\mathcal{F}$ .

As a result, the monotone class theorem applies and states that  $L^2(\Omega, \mathcal{F}, \mu) = \mathcal{H}$ , which concludes the first claim of the lemma.

Now, let us prove the last claim of the lemma. Observe that for any state-action pair  $(t, s_t, a_t)$ , the marginal distributions satisfy  $\tilde{\mathbb{P}}_n(S_t = s_t, A_t = a_t) = \mathbb{P}_n(S_t = s_t, A_t = a_t)$ . Hence, if the function  $f$  is only a function of  $(S_t, A_t)$ , we have  $\mathbb{E}_n[f(S_t, A_t)] = \tilde{\mathbb{E}}_n[f(S_t, A_t)]$ . For the probability measure  $\tilde{\mathbb{P}}_n$ , it is easy to see that

$$\text{var}(\tilde{\mathbb{E}}_n[f]) = E \left[ \tilde{\mathbb{E}}_n[f]^2 \right] - E[f]^2 = \frac{1}{n} \text{var}(f).$$

As a result,

$$E \left[ \mathbb{E}_n[f]^2 \right] = E \left[ \tilde{\mathbb{E}}_n[f]^2 \right] = E[f]^2 + \frac{1}{n} \text{var}(f).$$

□

**Comparison of the two empirical probability measures** The Markovian and the non-Markovian empirical probability measures are both viable choices, but we will argue in favor of the Markovian one.

Notice that for a state-action pair  $(t, s_t, a_t)$  observed in the training set  $\mathcal{T}$ , the conditional marginal distributions satisfy  $\tilde{\mathbb{P}}_n(S_{t+1}|t, s_t, a_t) = \mathbb{P}_n(S_{t+1}|t, s_t, a_t)$ . Similarly, there holds  $\tilde{\mathbb{P}}_n(R_t \in \mathfrak{R}_t | t, s_t, a_t, s_{t+1}) = \mathbb{P}_n(R_t \in \mathfrak{R}_t | t, s_t, a_t, s_{t+1})$ .

Although the distribution  $\tilde{\mathbb{P}}_n$  is an unbiased estimator of the true distribution  $\mu$ , using  $\tilde{\mathbb{P}}_n$  as an empirical distribution would be an inefficient use of the observations since we know that the true probabilistic model is Markovian with respect to the

state space  $\mathcal{X}$ . Let us illustrate this point with a simple example. Consider an MDP model such that  $T = 2$ ,  $\mathcal{S} = \{A, B\}$  and  $\mathcal{A} = \{u, d\}$ , where the state transitions and immediate rewards are deterministic. Assume that we observe only two trajectories, reported in the following table

$S_1$	$A_1$	$R_1$	$S_2$	$A_2$	$R_2$
$A$	$u$	$r_{1,A,u}$	$A$	$u$	$r_{2,A,u}$
$B$	$u$	$r_{1,B,u}$	$A$	$d$	$r_{2,A,d}$

Let  $\theta$  be the policy that chooses first the action  $u$  and then  $d$ . With the two above observed trajectories, the distribution  $\tilde{\mathbb{P}}_n$  does not give an estimate of the value of policy  $\theta$ , from the initial state  $A$ . In contrast, the empirical distribution  $\mathbb{P}_n$  combines the first transition of the first trajectory with the second transition of the second trajectory, to obtain a trajectory that can be used to estimate the performance of policy  $\theta$ .

Let us consider another situation to compare the two empirical probability measures. Two biased coins are tossed independently in a sequence. Denote respectively by  $p$  and  $q$  the probability of heads for the first and second coin. The empirical probability of seeing two heads is  $\tilde{\mathbb{P}}_n(HH) = \frac{N_{HH}}{n}$ , and has mean  $pq$  and variance  $\frac{1}{n}pq(1-pq)$ . On the other hand, the Markovian empirical probability of the same event,  $\mathbb{P}_n(HH) = \frac{N_{H1}}{n} \frac{N_{H2}}{n}$ , is also an unbiased estimator since  $E[\mathbb{P}_n(HH)] = E\left[\frac{N_{H1}}{n}\right] E\left[\frac{N_{H2}}{n}\right] = pq$ . Moreover, its second moment is the product of the second moments of Binomial random variables, divided by  $n^4$ , that is

$$\begin{aligned}
 E[\mathbb{P}_n(HH)^2] &= \frac{1}{n^4} E[N_{H1}^2] E[N_{H2}^2] \\
 &= np(1-p+np) \cdot nq(1-q+nq)/n^4.
 \end{aligned}$$

Hence, the variance of  $\mathbb{P}_n(HH)$  is

$$\begin{aligned} \text{var}(\mathbb{P}_n(HH)) &= E[\mathbb{P}_n(HH)^2] - E[\mathbb{P}_n(HH)]^2 \\ &= \frac{1}{n}pq(q(1-p) + p(1-q)) + \frac{1}{n^2}pq(1-p)(1-q). \end{aligned}$$

If we replace the factor  $1/n^2$  by  $1/n$ , the above expression would become equal to the variance of  $\tilde{\mathbb{P}}_n(HH)$  since  $(q(1-p) + p(1-q)) + (1-p)(1-q) = 1 - pq$ . Hence, the variance of  $\mathbb{P}_n(HH)$  is always smaller than the variance of  $\tilde{\mathbb{P}}_n(HH)$ , in this example. Furthermore, we can compare the leading term of each variance expression, that is the coefficient of  $1/n$ . For the sake of illustration assume that  $p = q = 0.1$  (which makes the probability of observing the event  $HH$  rather unlikely), then the coefficient associated with  $\tilde{\mathbb{P}}_n(HH)$  is  $0.01(1 - 0.01) \simeq 0.01$ . Comparatively, the coefficient associated with  $\mathbb{P}_n(HH)$  is  $0.01(0.09 + 0.09) \simeq 0.002$ , which is five times less than the variance of  $\tilde{\mathbb{P}}_n(HH)$ .

### 4.2.3 Estimation problem

The controller considers a set of policies indexed by a parameter  $\theta \in \Theta$ . We assume that the policy space  $\Theta$  contains only policies that are Markovian with respect to the state space  $\mathcal{S}$ . This condition can sometimes be enforced by redesigning the state space  $\mathcal{S}$  so that each state contains the information required by every policy of interest. Any Markovian policy  $\theta \in \Theta$  can be identified with the mapping  $(t, s) \mapsto \pi_t(a|s, \theta)$ , where  $\pi_t(a|s, \theta)$  is the probability that policy  $\theta$  chooses action  $a \in \mathcal{A}$  in state  $s \in \mathcal{S}$  at time  $t \in \{1, \dots, T\}$ . It will be handy to introduce the notational convention  $\pi_0(A_0|S_0, \theta) = \nu_0(A_0|S_0) = 1$  for all  $S_0, A_0, \theta$ . Furthermore, we require that each policy  $\theta \in \Theta$  is such that  $(\pi_t(a|s, \theta) > 0 \Rightarrow \nu_t(a|s) > 0)$ .

In Section 4.5, we will add some regularity assumptions on the set  $\Theta$  and on the mapping  $\theta \in \Theta \mapsto (\pi_t(a|s, \theta))_{s \in \mathcal{X}, a \in \mathcal{A}}$  so that the performance gradient is well-defined on the policy space  $\Theta$ .

An estimator is a random variable in  $\mathcal{E}$ . Note that this includes random variables given by formulas that involve the unknown model parameters. Clearly such esti-

mators are not practical. An estimator  $\hat{V}_\theta$  is an *unbiased* estimator of  $V_\theta$  if  $\hat{V}_\theta \in \mathcal{E}$  and  $E[\hat{V}_\theta] = V_\theta$ . An estimator  $\hat{V}_\theta$  is an *asymptotically unbiased* estimator of  $V_\theta$  if  $\hat{V}_\theta \in \mathcal{E}$  and  $E[\hat{V}_\theta]$  converges to  $V_\theta$  as the number  $n$  of trajectories in each training set increases to infinity.

Consider an unbiased estimator  $\hat{V}_\theta$  of the value  $V_\theta$ . Every training set  $\mathcal{T} = (\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T)$  maps to a different estimated value  $\hat{V}_\theta(\mathcal{T})$ , with an expectation equal to  $V_\theta$  (since the estimator is unbiased). The training set to training set variance (in the sequel, we will say training set variance)  $\text{var}(\hat{V}_\theta(\mathcal{T}))$  captures the variability of the estimated value  $\hat{V}_\theta(\mathcal{T})$  when the training set  $\mathcal{T}$  is randomly chosen. In this chapter, we would like to find unbiased estimators of the value  $V_\theta$  and the value derivatives  $\frac{\partial V_\theta}{\partial \theta}$  for any fixed policy  $\theta \in \Theta$ , with low training set variance.

### 4.3 Method of control variates, innovations of a random variable and geometry

In this section, we introduce some well-known statistical methods, which we tailor to our problem. First, we explain briefly the method of control variate, which is the method that we will use to find low variance estimators. Then we define the notion of action innovation of a random variable on a multi-period sample space. Finally, we characterize minimum variance estimators of the mean of a random variable in  $\mathcal{E}$  (with the action innovations as control variates) as orthogonal projections. These results will be specialized in Sections 4.4 and 4.5 to yield minimum variance estimators for the value and value derivative, respectively.

#### 4.3.1 Method of control variates

The method of control variates is a well-known method to reduce the variance of simulation estimates (e.g., [41, 43, 52] and references therein).

Consider a random variable  $Y$  with finite mean and variance, which we can sim-

ulate. The empirical mean  $\hat{Y}_n$  of  $n$  independent samples  $Y_1, \dots, Y_n$  of  $Y$ , i.e.,

$$\hat{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i,$$

is an unbiased estimator of the mean  $E[Y]$  of  $Y$ , with variance  $\text{var}(\hat{Y}_n) = \frac{1}{n} \text{var}(Y)$ .

Let  $Z$  be another random variable on the same probability space with finite mean and variance. By subtracting from  $Z$  its mean, we can assume without loss of generality that  $Z$  has zero mean. From  $n$  samples  $(Y_1, Z_1), \dots, (Y_n, Z_n)$ , we can form an unbiased estimator of  $E[Y]$  by

$$\widehat{Y + Z}_n = \frac{1}{n} \sum_{i=1}^n (Y_i + Z_i).$$

Its variance is  $\text{var}(\widehat{Y + Z}_n) = \frac{1}{n} \text{var}(Y + Z) = \frac{1}{n} [\text{var}(Y) + \text{var}(Z) + 2 \text{cov}(Y, Z)]$ . Hence, if  $Y$  and  $Z$  are negatively correlated, the estimator  $\widehat{Y + Z}_n$  has the potential to have lower variance than the empirical mean  $\hat{Y}_n$ . We refer to  $Z$  as a control variate.

In general, the objective is to identify a control variate  $Z$  that will minimize the variance of the estimator  $\widehat{Y + Z}_n$ .

The idea of control variate is used by Meyn and Henderson [33] in order to reduce the variance of simulation-based estimate of the steady-state mean number of customers  $\alpha$  in multi-class queueing networks. In their paper, the samples  $Y_n$ , which represent the number of customers in the system at time  $n$ , are generated from the trajectories of a Markov chain, in contrast with the simpler situation mentioned above where the samples were independent and identically distributed. Nonetheless, the principle is the same.

They start from a simple estimator  $\alpha(n) = \frac{1}{n} \sum_{i=0}^{n-1} Y_i$ , which is the mean number of customers in the network up to time  $n - 1$ . They consider control variates that are inspired from the Poisson equation (and also bear resemblance to the control variate of the baseline approach, which is detailed in Subsection 4.5.2). Specifically, for an arbitrary measurable real-valued function  $h$ , they define  $\Delta_h(y)$  the difference in expectation between the current value  $h(y)$  and the subsequent value  $h(Y_{n+1})$

given that  $Y_n = y$ . In order to compute  $\Delta_h$ , the transition kernel of the underlying Markov chain needs to be known. When  $Y_n$  is distributed according to its steady-state distribution,  $\Delta_h(Y_n)$  has zero expectation for all  $h$ . Thus, they introduce the control variate  $Z_n = \Delta_h(Y_n)$  and define the estimator  $\alpha^c(n) = \alpha(n) + \beta/n \sum_{i=0}^{n-1} \Delta_h(Y_i)$ .

If  $h$  is a solution to the Poisson equation and  $\beta = 1$ , then  $\alpha^c(n) = \alpha$  with probability one. Inspired by this observation, Meyn and Henderson investigate how to obtain an approximate solution of the Poisson equation by fluid and quadratic approximations, and how to choose the parameter  $\beta$  in order to have a low variance estimate of  $\alpha$ . They illustrate their algorithms with numerical experiments, but do not provide any theoretical results linking the quality of the approximation of a solution to Poisson's equation with the estimator variance.

In this chapter, we also tackle an estimation problem based on Markovian systems using the method of control variates. However, we use a different family of control variates, which are based on action innovations -- a notion that we introduce in the next subsection. This family of control variates is better suited to our setting since we do not know the underlying MDP model  $K$ , but we know the sampling policy  $\nu$ . Similar to [33], the optimal control variate is related to the Q-factors, the solution to Bellman's equations. In addition, we characterize the variance of the estimators that use a suboptimal control variate.

### 4.3.2 Innovations of a random variable

We define the notion of action and dynamics innovations of a random variable in  $\mathcal{E}$  and establish basic properties of innovations.

This concept is not new in the field of probability and statistics. It is closely related to the notion of martingale increments.

We can write any random variable  $Z \in \mathcal{E}$  as

$$Z = E[Z|\mathbf{S}_1] + (E[Z|\mathbf{S}_1, \mathbf{A}_1] - E[Z|\mathbf{S}_1]) + (E[Z|\mathbf{S}_1, \mathbf{A}_1, \mathbf{S}_2] - E[Z|\mathbf{S}_1, \mathbf{A}_1]) \quad (4.3.1) \\ + \dots + (Z - E[Z|\mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_T, \mathbf{A}_T]).$$

This expression for  $Z$  makes apparent the information that is revealed progressively as the trajectory gets realized. Formally, let

$$I_t^a[Z] = E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t] \in \mathcal{E}_t, \quad t = 1, \dots, T,$$

be the *action innovation*, which is associated with the realization of the actions  $A_t$  and

$$I_t^d[Z] = E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}] - E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \in \mathcal{E}_{t+1}, \quad t = 1, \dots, T-1,$$

be the *dynamics innovation*, which is associated with the realization of the new states  $S_{t+1}$ . We also let

$$I_T^d[Z] = Z - E[Z|\mathbf{S}_1, \dots, \mathbf{S}_T, \mathbf{A}_T] \in \mathcal{E}$$

and

$$I_0^d[Z] = E[Z|\mathbf{S}_1] \in \mathcal{E}_1.$$

Observe that all the innovations  $I_t^a[\cdot]$  and  $I_t^d[\cdot]$  are linear functionals from  $\mathcal{E}$  into  $\mathcal{E}_t$ . Furthermore, they allow us to write succinctly the decomposition (4.3.1) of  $Z$  as

$$Z = \sum_{t=0}^T I_t^d[Z] + \sum_{t=1}^T I_t^a[Z]. \quad (4.3.2)$$

The innovations have interesting properties summarized in the following lemma.

**Lemma 4.3.1.**

(a) *All the innovations of  $Z \in \mathcal{E}$  are uncorrelated. Thus, the variance of  $Z$  is the sum of the variances of its innovations, i.e.,*

$$\text{var}(Z) = \sum_{t=0}^T \text{var}(I_t^d[Z]) + \sum_{t=1}^T \text{var}(I_t^a[Z]).$$

(b) For all random variable  $Z \in \mathcal{E}$  and  $t, \tau = 1, \dots, T$ , there holds

$$I_\tau^\alpha [I_t^\alpha [Z]] = \begin{cases} I_t^\alpha [Z] & \text{if } t = \tau \\ 0 & \text{otherwise} \end{cases}$$

(c) For  $t = 1, \dots, T$ ,  $\tau = 0, \dots, T$  and  $Z \in \mathcal{E}$ , we have

$$I_\tau^\alpha [I_t^\alpha [Z]] = 0.$$

(d) For any  $Y, Z \in \mathcal{E}$ ,  $t = 1, \dots, T$  and  $\tau = 0, \dots, T$ , the innovations  $I_t^\alpha [Z]$  and  $I_\tau^\alpha [Y]$  are uncorrelated.

(e) For any  $X, Y \in \mathcal{E}$  and  $t \neq \tau$ , the action innovations  $I_t^\alpha [X]$  and  $I_\tau^\alpha [Y]$  are uncorrelated.

*Proof.* (a) It is easy to check that all the innovations in (4.3.1) are uncorrelated (orthogonal) to each other. For example,

$$\begin{aligned} & \text{cov}(E[Z|\mathbf{S}_1]; E[Z|\mathbf{S}_1, \mathbf{A}_1] - E[Z|\mathbf{S}_1]) \\ &= E \left[ E[Z|\mathbf{S}_1] E[E[Z|\mathbf{S}_1, \mathbf{A}_1] - E[Z|\mathbf{S}_1]|\mathbf{S}_1] \right] - E[E[Z|\mathbf{S}_1]] E[E[Z|\mathbf{S}_1, \mathbf{A}_1] - E[Z|\mathbf{S}_1]] \\ &= E[ E[Z|\mathbf{S}_1] \cdot 0 ] - E[E[Z|\mathbf{S}_1]] \cdot 0 = 0. \end{aligned}$$

(b)

- If  $t = \tau$ , we have

$$I_\tau^\alpha [I_t^\alpha [Z]] = E[I_t^\alpha [Z]|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[I_t^\alpha [Z]|\mathbf{S}_1, \dots, \mathbf{S}_t] = I_t^\alpha [Z].$$

- If  $t < \tau$ , we have

$$I_\tau^\alpha [I_t^\alpha [Z]] = E[I_t^\alpha [Z]|\mathbf{S}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] - E[I_t^\alpha [Z]|\mathbf{S}_1, \dots, \mathbf{S}_\tau] = I_t^\alpha [Z] - I_t^\alpha [Z] = 0.$$

- If  $t > \tau$ , we have

$$I_\tau^\alpha [I_t^\alpha [Z]] = E[I_t^\alpha [Z] | \mathbf{S}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] - E[I_t^\alpha [Z] | \mathbf{S}_1, \dots, \mathbf{S}_\tau] = 0 - 0 = 0.$$

(c) Fix  $t \in \{1, \dots, T\}$ . By definition, we have  $I_0^d [I_t^\alpha [Z]] = E[I_t^\alpha [Z] | S_1] = 0$ . The action innovation at time  $T$ ,  $I_T^d [I_t^\alpha [Z]]$ , is zero since  $I_t^\alpha [Z]$  is a function of  $(\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t)$ .

To conclude the proof, we show that, for  $\tau = 1, \dots, T - 1$ , we have

$$I_\tau^d [I_t^\alpha [Z]] = E[I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau, \mathbf{S}_{\tau+1}] - E[I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] = 0.$$

For all  $\tau < t$ , we have

$$E[I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] = E[I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau, \mathbf{S}_{\tau+1}] = 0.$$

On the other hand, for  $\tau \geq t$ ,

$$E [I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] = I_t^\alpha [Z] = E [I_t^\alpha [Z] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau, \mathbf{S}_{\tau+1}].$$

Therefore,

$$I_\tau^d [I_t^\alpha [X_t]] = E[I_t^\alpha [X_t] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau, \mathbf{S}_{\tau+1}] - E[I_t^\alpha [X_t] | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_\tau, \mathbf{A}_\tau] = 0.$$

(d) Since all the action innovations  $I_t^\alpha [Z]$  and  $I_\tau^d [Y]$  have zero expectation for  $t = 1, \dots, T$ , it suffices to show that  $E [I_t^\alpha [Z] I_\tau^d [Y]] = 0$ . There are four cases to consider.

- If  $\tau = 0$ ,  $E [I_t^\alpha [Z] I_\tau^d [Y]] = E [E [I_t^\alpha [Z] | S_1] I_0^d (Y)] = 0$ .
- If  $\tau = T$ ,  $E [I_t^\alpha [Z] I_\tau^d [Y]] = E [I_t^\alpha [Z] E [I_T^d [Y] | \mathbf{S}_1, \dots, \mathbf{A}_T]] = 0$ .
- If  $t > \tau > 0$ ,  $E [I_t^\alpha [Z] I_\tau^d [Y]] = E [E [I_t^\alpha [Z] | \mathbf{S}_1, \dots, \mathbf{S}_{\tau+1}] I_\tau^d [Y]] = 0$ .
- If  $t \leq \tau < T$ ,  $E [I_t^\alpha [Z] I_\tau^d [Y]] = E [I_t^\alpha [Z] E [I_\tau^d [Y] | \mathbf{S}_1, \dots, \mathbf{A}_\tau]] = 0$ .

(e) Since all the action innovations have zero expectation, it suffices to show that  $E[I_t^a[Z]I_\tau^a[Y]] = 0$ . Without loss of generality, assume  $t < \tau$ . We have

$$E[I_t^a[Z]I_\tau^a[Y]] = E[I_t^a[Z]E[I_\tau^a[Y]|\mathbf{S}_1, \dots, \mathbf{S}_\tau]] = 0.$$

□

**Definition 4.3.2.** Starting with the random variables  $Z \in \mathcal{E}$ ,  $X_t \in \mathcal{E}_t$ ,  $t = 1, \dots, T$ , we define the random variable  $Z[X_1, \dots, X_T]$  by

$$Z[X_1, \dots, X_T] = Z - \sum_{t=1}^T I_t^a[X_t]. \quad (4.3.3)$$

Note that  $Z[0, \dots, 0] = Z$ .

Notice that  $Z$  and  $Z[X_1, \dots, X_T]$  have the same expectation so that the random variable  $Z[X_1, \dots, X_T]$  can be thought of as an unbiased estimator of  $E[Z]$  for any choice of  $X_t \in \mathcal{E}$ ,  $t = 1, \dots, T$ .

### 4.3.3 Minimum variance estimator: the unconstrained case

In this subsection, we will show that a minimum variance element in the family  $\{Z[X_1, \dots, X_T], X_t \in \mathcal{E}_t\}$  is the orthogonal projection of  $Z$  on a suitable space. This characterization will be central to the identification of optimal unconstrained estimators of the value and value gradient in Sections 4.4 and 4.5.

Define the projection operator  $\mathbf{\Pi}$  on the linear subspace  $\mathcal{E}^d = \{Z \in \mathcal{E} \mid I_t^a[Z] = 0, t = 1, \dots, T\}$  by

$$Z \in \mathcal{E} \mapsto \mathbf{\Pi}(Z) = \sum_{t=0}^T I_t^d[Z].$$

Equivalently, from Equation (4.3.2) we have

$$\mathbf{\Pi}(Z) = Z - \sum_{t=1}^T I_t^a[Z]. \quad (4.3.4)$$

Intuitively,  $\mathbf{\Pi}(Z)$  is an approximation of  $Z$  on  $\mathcal{E}^d$ , in the sense that its dynamics innovations match those of  $Z$  at all time, i.e.,  $I_t^d(\mathbf{\Pi}(Z)) = I_t^d[Z]$  for  $t = 1, \dots, T$ , but  $\mathbf{\Pi}(Z)$  has zero action innovations, with probability one.

The next proposition shows that the projection  $\mathbf{\Pi}(Z)$  of the random variable  $Z$  has minimum variance among the random variables of the form  $Z[X_1, \dots, X_T]$  with  $X_t \in \mathcal{E}_t$ . But first we establish two useful lemmas.

**Lemma 4.3.3.** *The operator  $\mathbf{\Pi}$  is an orthogonal projection on  $\mathcal{E}^d$  for the natural scalar product on  $\mathcal{E}$ . As a result, the variance of any  $Z \in \mathcal{E}$  decomposes as*

$$\text{var}(Z) = \text{var}(\mathbf{\Pi}(Z)) + \text{var}(Z - \mathbf{\Pi}(Z)).$$

*Proof.* First, thanks to Lemma 4.3.1 there holds  $I_t^a[I_\tau^d[Z]] = 0$  for all  $\tau = 0, \dots, T$  and all  $t = 1, \dots, T$ . Thus, we have  $I_t^a[\mathbf{\Pi}(Z)] = 0$ . Consequently,  $\mathbf{\Pi}(Z) \subset \mathcal{E}^d$  for all  $Z \in \mathcal{E}$ .

Moreover, it is easy to verify that  $\mathbf{\Pi}$  is linear and  $\mathbf{\Pi}^2 = \mathbf{\Pi}$ . Hence, we have checked that  $\mathbf{\Pi}$  is a linear projection operator on  $\mathcal{E}^d$ .

Now, we show it is an orthogonal projection.

Since the image  $\mathbf{\Pi}(Z)$  of  $Z$  is simply the sum of all the dynamics innovations,  $\mathbf{\Pi}(Z)$  and  $Z - \mathbf{\Pi}(Z)$  are uncorrelated. Consequently, the variance of  $Z$  is the sum of their variances. □

**Lemma 4.3.4.** *For any  $Z \in \mathcal{E}$  and  $X_t \in \mathcal{E}_t$ ,  $t = 1, \dots, T$ , we have*

$$\mathbf{\Pi}(Z[X_1, \dots, X_T]) = \mathbf{\Pi}(Z).$$

*Proof.* We have

$$\mathbf{\Pi}(Z[X_1, \dots, X_T]) = \mathbf{\Pi}(Z) - \sum_{t=1}^T \sum_{\tau=0}^T I_\tau^d [I_t^a[X_t]].$$

We conclude the proof by observing that the last term is zero by the part (c) of Lemma 4.3.1. □

**Proposition 4.3.5.** For any given  $Z \in \mathcal{E}$ , the random variable  $\mathbf{\Pi}(Z)$  can be written as

$$\mathbf{\Pi}(Z) = Z[X_1^*, \dots, X_T^*],$$

with  $X_t^*(\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t) = -E[Z|\mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \in \mathcal{E}_t$ . Alternatively, we can let  $X_t^* = I_t^a[Z]$ .

Moreover,  $\mathbf{\Pi}(Z)$  has the minimum variance in the set  $\{Z[X_1, \dots, X_T], X_1 \in \mathcal{E}_1, \dots, X_T \in \mathcal{E}_T\}$ .

*Proof.* By definition, we have

$$\mathbf{\Pi}(Z) = \sum_{t=0}^T I_t^d[Z] = Z - \sum_{t=1}^T I_t^a[Z] = Z[X_1^*, \dots, X_T^*],$$

with  $X_t^* = I_t^a[Z]$ . It is easy to check that we can also let  $X_t^*(\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t) = -E[Z|\mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t]$ .

By Lemma 4.3.3,

$$\text{var}(Z[X_1, \dots, X_T]) = \text{var}\left(\mathbf{\Pi}(Z[X_1, \dots, X_T])\right) + \text{var}\left(Z[X_1, \dots, X_T] - \mathbf{\Pi}(Z[X_1, \dots, X_T])\right).$$

From Lemma 4.3.4, there holds  $\mathbf{\Pi}(Z[X_1, \dots, X_T]) = \mathbf{\Pi}(Z)$  so that the first term in the expression of the variance does not depend on  $X_1, \dots, X_T$ .

The second term is non-negative and equals to zero when  $X_t = X_t^*$ .  $\square$

#### 4.3.4 Minimum variance estimator: the constrained case

Similar to the unconstrained case, we show that a minimum variance element in the family  $\{Z[X_1, \dots, X_T], X_t \in \mathcal{X}_t\}$  is obtained by the orthogonal projection of  $Z$  on a suitable space. The result of this subsection will be useful to characterize the optimal constrained estimators of the value and value gradient in Sections 4.4 and 4.5, respectively.

Let  $\mathcal{X}_t$  be a closed linear subspace of  $\mathcal{E}_t$  for  $t = 1, \dots, T$ , and define

$$\mathcal{I}_t = \{I_t^\alpha[X_t], X_t \in \mathcal{X}_t\} \subset \mathcal{E}_t.$$

The set  $\mathcal{I}_t$  is a non-empty linear subspace of  $\mathcal{E}$ . Since the state and action space are finite,  $\mathcal{I}_t$ , which contains only functions of  $\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t$ , is finite-dimensional, and thus closed in  $\mathcal{E}$ .

Let  $\mathbf{\Pi}_t$  be the orthogonal projection from  $\mathcal{E}$  onto  $\mathcal{I}_t$ , and let  $\mathcal{E}^c = \{Z \in \mathcal{E} \mid \mathbf{\Pi}_t(I_t^\alpha[Z]) = 0, t = 1, \dots, T\}$ . Define the operator  $\mathbf{\Pi}^c$  on  $\mathcal{E}$  by

$$Z \in \mathcal{E} \mapsto \mathbf{\Pi}^c(Z) = Z - \sum_{t=1}^T \mathbf{\Pi}_t(I_t^\alpha[Z]).$$

**Lemma 4.3.6.** *The operator  $\mathbf{\Pi}^c$  is an orthogonal projection from  $\mathcal{E}$  onto  $\mathcal{E}^c$ .*

*Proof.* Let  $Z \in \mathcal{E}$ . There holds

$$\begin{aligned} \mathbf{\Pi}^c(\mathbf{\Pi}^c(Z)) &= \mathbf{\Pi}^c(Z) - \sum_{\tau=1}^T \mathbf{\Pi}_\tau(I_\tau^\alpha[\mathbf{\Pi}^c(Z)]) \\ &= \mathbf{\Pi}^c(Z) - \sum_{\tau=1}^T \mathbf{\Pi}_\tau \left[ I_\tau^\alpha \left[ \sum_{t=1}^T I_t^d[Z] + \sum_{t=1}^T (I - \mathbf{\Pi}_t)(I_t^\alpha[Z]) \right] \right]. \end{aligned}$$

The last equality follows by replacing

$$\begin{aligned} \mathbf{\Pi}^c(Z) &= Z - \sum_{t=1}^T \mathbf{\Pi}_t(I_t^\alpha[Z]) \\ &= \sum_{t=1}^T I_t^d[Z] + \sum_{t=1}^T (I - \mathbf{\Pi}_t)(I_t^\alpha[Z]) + \sum_{t=1}^T \mathbf{\Pi}_t(I_t^\alpha[Z]) - \sum_{t=1}^T \mathbf{\Pi}_t(I_t^\alpha[Z]) \\ &= \sum_{t=1}^T I_t^d[Z] + \sum_{t=1}^T (I - \mathbf{\Pi}_t)(I_t^\alpha[Z]). \end{aligned}$$

Notice that the terms  $I_\tau^\alpha[I_t^d[Z]] = 0$  by Lemma 4.3.1. Let us look at the last terms  $I_\tau^\alpha((I - \mathbf{\Pi}_t)[I_t^\alpha[Z]])$ . Since  $\mathbf{\Pi}_t(I_t^\alpha[Z]) \in \mathcal{I}_t$ , there is, by definition of  $\mathcal{I}_t$ ,  $Y_t \in \mathcal{X}_t$  such that  $\mathbf{\Pi}_t(I_t^\alpha[Z]) = I_t^\alpha[Y_t]$ . Hence,  $I_\tau^\alpha[\mathbf{\Pi}_t(I_t^\alpha[Z])] = I_\tau^\alpha[I_t^\alpha[Y_t]]$ . By Lemma 4.3.1, we conclude that for  $t \neq \tau$ ,  $I_\tau^\alpha(\mathbf{\Pi}_t(I_t^\alpha[Z])) = 0$ , and for  $t = \tau$ ,  $I_\tau^\alpha(\mathbf{\Pi}_t(I_t^\alpha[Z])) = I_\tau^\alpha[I_t^\alpha[Y_t]] =$

$\mathbf{\Pi}_t(I_t^a[Z])$ . Consequently, we have, for all  $t, \tau$ ,

$$I_\tau^a [(I - \mathbf{\Pi}_t)(I_t^a[Z])] = I_\tau^a [I_t^a[Z]] - I_\tau^a [\mathbf{\Pi}_t(I_t^a[Z])] = 0.$$

This implies that  $\mathbf{\Pi}^c(\mathbf{\Pi}^c(Z)) = \mathbf{\Pi}^c(Z)$ . This concludes the proof that  $\mathbf{\Pi}^c$  is a projection on  $\mathcal{E}^c$ .

It is easy to see that the projection  $\mathbf{\Pi}^c$  is orthogonal.  $\square$

The following proposition generalizes Proposition 4.3.5 to the constrained case.

**Proposition 4.3.7.** *Let  $X_t^* = I_t^a[Z] \in \mathcal{E}_t$ , and let  $Y_t \in \mathcal{E}_t$ . The variance of the estimator of  $E[Z]$  defined by  $Z[X_1^* - Y_1, \dots, X_T^* - Y_T]$  decomposes as*

$$\text{var}(Z[X_1^* - Y_1, \dots, X_T^* - Y_T]) = \text{var}(\mathbf{\Pi}(Z)) + \sum_{t=1}^T \text{var}(I_t^a[Y_t]).$$

*A minimum variance element in the set  $\{Z[X_1, \dots, X_T] \mid X_t \in \mathcal{X}_t, t = 1, \dots, T\}$  is*

$$\mathbf{\Pi}^c(Z) = Z[W_1^c, \dots, W_T^c],$$

where  $W_t^c = \arg \min_{W_t \in \mathcal{X}_t} E[(X_t^* - I_t^a[W_t])^2]$ .

*Proof.* We have

$$\begin{aligned} Z[X_1^* - Y_1, \dots, X_T^* - Y_T] &= Z - \sum_{t=1}^T I_t^a[X_t^* - Y_t] \\ &= \sum_{t=0}^T I_t^d[Z] + \sum_{t=1}^T I_t^a[Y_t] \\ &= \mathbf{\Pi}(Z) + \sum_{t=1}^T I_t^a[Y_t]. \end{aligned}$$

All the terms of this expression are uncorrelated from Lemma 4.3.1. Hence, the variance of the left-hand side is the sum of the variance of the terms in the right-hand side. This shows the first claim of the proposition.

We would like that  $Y_t \in \mathcal{E}_t$  minimize the variance  $\text{var}(I_t^a[Y_t])$  under the constraint

that  $X_t^* - Y_t \in \mathcal{X}^t$ . Using the variable  $W_t = X_t^* - Y_t$ , a variance minimizing choice corresponds to

$$W_t^c = \arg \min_{W_t \in \mathcal{X}_t} E [I_t^a [X_t^* - W_t]^2] = \arg \min_{W_t \in \mathcal{X}_t} E [(X_t^* - I_t^a [W_t])^2].$$

In the second part of the proposition, it remains to show that  $\Pi^c(Z) = Z [W_1^c, \dots, W_T^c]$ .

By definition of  $W_t^c$ , we have  $I_t^a [W_t^c] = \Pi_t(X_t^*) = \Pi_t(I_t^a [Z])$ . On the other hand,

$$\begin{aligned} Z [W_1^c, \dots, W_T^c] &= Z - \sum_{t=1}^T I_t^a [W_t^c] \\ &= Z - \sum_{t=1}^T \Pi_t(I_t^a [Z]) \\ &= \Pi^c(Z). \end{aligned}$$

□

The next two sections exploit the results established in this section in order to characterize optimal estimators for the value and the value gradient.

## 4.4 Estimation of the value of a policy

In this section, we define two classes of unbiased estimators of the value of policy  $\theta \in \Theta$ , which are based on trajectories observed while following a known sampling policy  $\nu$ : the unconstrained and the constrained estimators. Using the concepts of Section 4.3, we identify estimators in these classes with minimum training set variance and propose algorithms that take advantages of these theoretical insights. In practice, the best algorithm relies on the optimal unconstrained or constrained estimator, depending on the availability of observations compared to the dimensionality of the underlying MDP.

Recall that the training set  $\mathcal{T}$  comprises  $n$  IID trajectories sampled according to the probability measure  $\mu^n$ , which depends on the initial state distribution  $\eta$ , on the unknown transition kernel  $K$  of the MDP and on the sampling policy  $\nu$ . But we

would like to estimate the expectation of the total reward according to the probability measure associated with  $\eta$ ,  $K$  and policy  $\theta$ . This difference can be overcome using the idea of importance sampling.

The probability measure associated with policy  $\theta$  has a Radon-Nikodym derivative with respect to  $\mu^n$ , since the sampling policy  $\nu$  gives non-zero probability to all actions. Furthermore, its derivative (called the likelihood ratio in the importance sampling literature) is

$$\mathbf{L}_\theta(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T) = \frac{\pi_1(\mathbf{A}_1|\mathbf{S}_1, \theta)}{\nu_1(\mathbf{A}_1|\mathbf{S}_1)} \dots \frac{\pi_T(\mathbf{A}_T|\mathbf{S}_T, \theta)}{\nu_T(\mathbf{A}_T|\mathbf{S}_T)},$$

where  $\pi_t(\mathbf{A}_t|\mathbf{S}_t, \theta) = \prod_{k=1}^n \pi_t(A_t^k|S_t^k, \theta)$  and  $\nu_t(\mathbf{a}_t|\mathbf{S}_t) = \prod_{k=1}^n \nu_t(a_t^k|S_t^k)$ . Thus, the Radon-Nikodym derivative  $\mathbf{L}_\theta(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T)$  does not depend on the transition kernel  $K$ , nor on the initial state distribution  $\eta$ . Define also the likelihood ratio for one trajectory by  $L_\theta(S_1, A_1, R_1, \dots, S_T, A_T, R_T) = \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_T(A_T|S_T, \theta)}{\nu_T(A_T|S_T)}$ .

**Remark 4.4.1.** *In the case where the true model is non-Markov, i.e.,*

$$\begin{aligned} & \mu(S_1 = s_1, A_1 = a_1, R_1 \in \mathfrak{R}_1, \dots, S_T = s_T, A_T = a_t, R_T \in \mathfrak{R}_T) \\ &= \eta(s_1) \prod_{t=1}^T \nu(a_t|s_t) K^d(s_{t+1}|t, s_1, \dots, s_t, a_t) K^r(R_t \in \mathfrak{R}_t|t, s_1, \dots, s_t, a_t, s_{t+1}), \end{aligned}$$

the likelihood ratio between trajectories generated by policies  $\theta$  and  $\nu$  is the same as in the Markovian case, that is,  $\mathbf{L}_\theta(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T) = \frac{\pi_1(\mathbf{A}_1|\mathbf{S}_1, \theta)}{\nu_1(\mathbf{A}_1|\mathbf{S}_1)} \dots \frac{\pi_T(\mathbf{A}_T|\mathbf{S}_T, \theta)}{\nu_T(\mathbf{A}_T|\mathbf{S}_T)}$ . However, we will need the true model to be Markovian in the sequel - for example to have well-defined  $Q$ -factors.

The value  $V_\theta$  of policy  $\theta$  given the initial state distribution  $\eta$  and MDP model  $K$  can be written as an expectation with respect to  $\mu$ :

$$\begin{aligned} V_\theta &= E_\theta[R_1 + \dots + R_T] = E_\nu[L_\theta(R_1 + \dots + R_T)] \\ &= E_\nu \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} R_t \right]. \end{aligned}$$

As established in Lemma 4.2.4, we can approximate the true expectation with respect to  $\mu$  by the (random) empirical expectation  $\mathbb{E}_n$  to obtain an unbiased estimator  $\hat{V}_\theta$  of  $V_\theta$  defined by

$$\hat{V}_\theta(\mathbf{S}_1, \mathbf{A}_1, \mathbf{R}_1, \dots, \mathbf{S}_T, \mathbf{A}_T, \mathbf{R}_T) = \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} R_t \right]. \quad (4.4.1)$$

Indeed, under our assumptions,  $\sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} R_t \in L^2(\Omega, \mathcal{F}, \mu)$  and by Lemma 4.2.4, the estimator  $\hat{V}_\theta$  belongs to  $\mathcal{E}$  and has expectation

$$E_\nu \left[ \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} R_t \right] \right] = E_\nu \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} R_t \right] = V_\theta.$$

Observe that the estimator  $\hat{V}_\theta$  is random only through the empirical measure, which depends on the random training set  $\mathcal{T}$ .

On the downside, the estimator  $\hat{V}_\theta$  can have a large training set variance, especially if the number of samples  $n$  is small compared to the number of states  $|\mathcal{X}|$ . Hence, we would like to reduce the variance of  $\hat{V}_\theta$  using a control variate approach.

**Remark 4.4.2.** *When the likelihood ratios take very large values, the variance of the estimator  $\hat{V}_\theta$  could become particularly large. This situation might occur if the time horizon is very long, or if the sampling policy  $\nu$  picks some actions with very little probability, whereas policy  $\pi$  selects them with higher probability.*

*On the other hand, observe that a large state space does not affect directly the scale of the likelihood ratios.*

For all  $X_t \in \mathcal{E}_t$ , the law of iterated expectations states that

$$E_\nu[X_t | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t] = \sum_{\mathbf{a}_t} \nu_t(\mathbf{a}_t | \mathbf{S}_t) E_\nu[X_t | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{a}_t],$$

or equivalently  $E[I_t^\alpha[X_t]] = 0$ . Consequently, the estimator

$$\hat{V}_\theta[X_1, \dots, X_T] = \hat{V}_\theta - \sum_{t=1}^T I_t^\alpha[X_t] \quad (4.4.2)$$

is an unbiased estimator of the value  $V_\theta$  of any policy  $\theta \in \Theta$  for all  $X_1 \in \mathcal{E}_1, \dots, X_T \in \mathcal{E}_T$ .

**Remark 4.4.3.** *Since we consider the situation where the sampling policy  $\nu$  is known, the action innovation at time  $t$  of a known function  $X_t(\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t)$  can be computed. In contrast, evaluating dynamic innovations requires the knowledge of the marginal distribution of the next state, which is unknown. Therefore, we only use action innovations as control variates, and not dynamics innovations.*

The concept of innovations provides insights in the value estimator  $\hat{V}_\theta[X_1, \dots, X_T]$ : according to Lemma 4.3.1,  $\hat{V}_\theta[X_1, \dots, X_T]$  has the same dynamics innovation as the naive estimator  $\hat{V}_\theta$ , but its action innovations are  $I_t^a[\hat{V}_\theta[X_1, \dots, X_T]] = I_t^a[\hat{V}_\theta] - I_t^a[X_t]$ . This chapter focuses on this class of estimators only, but this class is quite rich and includes many of the estimators proposed in the reinforcement learning literature, as we will see later.

Now the question is: what choice for  $X_1, \dots, X_T$  minimizes the training set variance of the unbiased estimator  $\hat{V}_\theta[X_1, \dots, X_T]$  given a known sampling policy  $\nu$ ? We will identify the optimal value estimator when the functions  $X_t$  can be arbitrary elements of  $\mathcal{E}_t$  (unconstrained case), and when  $X_t$  is constrained to belong to a closed convex subspace  $\mathcal{X}_t \subset \mathcal{E}_t$  for  $t = 1, \dots, T$ , in Subsections 4.4.1 and 4.4.2, respectively.

#### 4.4.1 Unconstrained value estimator

Now, we will find a value estimator of the form  $\hat{V}_\theta[X_1, \dots, X_T]$  with minimum variance. Geometrically, we will show that the projection of the naive estimator  $\hat{V}_\theta$  by  $\Pi$  is such an optimal estimator and we will provide an algebraic expression for it.

##### Characterization of the optimal unconstrained value estimator

Define the tail cost of policy  $\theta$  on the trajectory  $(S_1, A_1, R_1, \dots, S_T, A_T, R_T)$  by

$$C_\theta^t = R_t + \frac{\pi_{t+1}(A_{t+1}|S_{t+1}, \theta)}{\nu_{t+1}(A_{t+1}|S_{t+1})} R_{t+1} + \dots + \frac{\pi_{t+1}(A_{t+1}|S_{t+1}, \theta)}{\nu_{t+1}(A_{t+1}|S_{t+1})} \dots \frac{\pi_T(A_T|S_T, \theta)}{\nu_T(A_T|S_T)} R_T.$$

The conditional expectation of  $C_\theta^t$  with respect to  $\mu$  given a state-action pair is known as the Q-factor  $Q_\theta^t$  of policy  $\theta$  at time  $t$  in the MDP model  $K$ , i.e.,

$$E_\nu[C_\theta^t | S_t = s, A_t = a] = E_\theta[R_t + \dots + R_T | S_t = s, A_t = a] = Q_\theta^t(s, a),$$

while its expectation given a state is known as the value  $V_\theta^t$  of the state, i.e.,

$$V_\theta^t(s) = E_\nu[C_\theta^t | S_t = s] = E_\theta[R_t + \dots + R_T | S_t = s] = V_\theta^t(s).$$

For  $t = 1, \dots, T$ , denote by  $B_t \in \mathcal{E}$  the (random) Bellman error at time  $t$ , that is

$$B_t = R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1} | S_{t+1}) Q_\theta^{t+1}(S_{t+1}, a_{t+1}) - Q_\theta^t(S_t, A_t),$$

where  $Q_\theta^{T+1} = 0$ . Notice that it is a function only of  $S_t$ ,  $A_t$ ,  $S_{t+1}$ , and  $R_t$ .

**Proposition 4.4.4.**

(a) *The minimum variance estimator of  $V_\theta$  in the family*

$$\left\{ \hat{V}_\theta[X_1, \dots, X_T], X_t \in \mathcal{E}_t \right\}$$

is

$$\begin{aligned} \hat{V}_\theta^* &= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1 | S_1, \theta) Q_\theta^1(S_1, a_1) \right] \\ &+ \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} B_t \right]. \end{aligned} \quad (4.4.3)$$

(b) *Given the Markovian sampling policy  $\nu$ , the variance of the optimal estimator  $\hat{V}_\theta^*$  satisfies*

$$\text{var}_\nu(\hat{V}_\theta^*) = \frac{1}{n} \text{var}_\eta(V_\theta(S_1)) + \frac{1}{n} \sum_{t=1}^T \text{var} \left( \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} B_t \right). \quad (4.4.4)$$

*Proof.* (a) Thanks to Proposition 4.3.5, a minimal variance estimator of the form

$\hat{V}_\theta[X_1, \dots, X_T]$  with  $X_t \in \mathcal{E}_t$  is  $\Pi(\hat{V}_\theta)$ .

Furthermore, we know  $\Pi(\hat{V}_\theta) = \hat{V}_\theta[X_1^*, \dots, X_T^*]$  with

$$\begin{aligned}
X_t^* &= -E[\hat{V}_\theta | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \\
&= -E \left[ \mathbb{E}_n \left[ \sum_{\tau=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_\tau(A_\tau|S_\tau, \theta)}{\nu_\tau(A_\tau|S_\tau)} R_\tau \right] \middle| \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t \right] \\
&= -\mathbb{E}_n \left[ \sum_{\tau=1}^{t-1} \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_\tau(A_\tau|S_\tau, \theta)}{\nu_\tau(A_\tau|S_\tau)} E[R_\tau | S_1, A_1, \dots, S_t, A_t] \right] \\
&\quad - \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} E[C_\theta^t | S_1, A_1, \dots, S_t, A_t] \right] \\
&= -\mathbb{E}_n \left[ \sum_{\tau=1}^{t-1} \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_\tau(A_\tau|S_\tau, \theta)}{\nu_\tau(A_\tau|S_\tau)} E[R_\tau | S_\tau, A_\tau, S_{\tau+1}] \right] \\
&\quad - \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} Q_\theta^t(S_t, A_t) \right].
\end{aligned}$$

Indeed, the third equality is justified since we have

$$\begin{aligned}
&E \left[ \mathbb{E}_n \left[ \sum_{\tau=t}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_\tau(A_\tau|S_\tau, \theta)}{\nu_\tau(A_\tau|S_\tau)} R_\tau \right] \middle| \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t \right] \\
&= \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} E \left[ \mathbb{E}_n [C_\theta^t] \middle| \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t \right] \right] \\
&= \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} E[C_\theta^t | S_1, A_1, \dots, S_t, A_t] \right].
\end{aligned}$$

And the last equality follows from the independence of the reward  $R_t$  from the past, conditional on  $(S_t, A_t, S_{t+1})$ , and the Markovianity of policies  $\theta$  and  $\nu$ .

Consequently, for  $t = 1, \dots, T$ ,

$$\begin{aligned}
I_t^\alpha[X_t^*] &= E[X_t^* | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - \sum_{\mathbf{a}_t} \nu_t(\mathbf{a}_t | \mathbf{S}_t) E[X_t^* | \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{a}_t] \\
&= -\mathbb{E}_n \left[ \frac{\pi_0(A_0 | S_0, \theta)}{\nu_0(A_0 | S_0)} \cdots \frac{\pi_{t-1}(A_{t-1} | S_{t-1}, \theta)}{\nu_{t-1}(A_{t-1} | S_{t-1})} \left( \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) - \sum_{a_t} \pi_t(a_t | S_t, \theta) Q_\theta^t(S_t, a_t) \right) \right] \\
&= E \left[ -\mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) \mid \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{A}_t \right] \right] \\
&\quad - \sum_{a_t} \nu_t(a_t | S_t) E \left[ -\mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) \mid \mathbf{S}_1, \mathbf{A}_1, \dots, \mathbf{S}_t, \mathbf{a}_t \right] \right] \\
&= I_t^\alpha \left( -\mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) \right] \right).
\end{aligned}$$

Therefore, letting  $X_t = -\mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) \right] \in \mathcal{E}_t$  for  $t = 1, \dots, T$ , also minimizes the variance of the estimator  $\hat{V}_\theta[X_1, \dots, X_T]$ .

Plugging these optimal values in (4.4.2) yields

$$\begin{aligned}
\hat{V}_\theta[X_1^*, \dots, X_T^*] &= \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} R_t \right] \\
&\quad - \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_{t-1}(A_{t-1} | S_{t-1}, \theta)}{\nu_{t-1}(A_{t-1} | S_{t-1})} \left( \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} Q_\theta^t(S_t, A_t) - \sum_{a_t} \pi_t(a_t | S_t, \theta) Q_\theta^t(S_t, a_t) \right) \right] \\
&= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1 | S_1, \theta) Q_\theta^1(S_1, a_1) \right] \\
&\quad + \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \cdots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} \left( R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1} | S_{t+1}) Q_\theta^{t+1}(S_{t+1}, a_{t+1}) - Q_\theta^t(S_t, A_t) \right) \right].
\end{aligned}$$

(b) Now, we show that the minimum variance has the form claimed in the proposition.

From Lemma 4.3.4, the variance of the optimal estimator  $\hat{V}_\theta^*$  is the sum of the variances of the dynamics innovations of the naive estimator  $\hat{V}_\theta$ , i.e.,

$$\text{var}(\hat{V}_\theta^*) = \sum_{t=0}^T \text{var} \left( I_T^d[\hat{V}_\theta] \right).$$

For  $t = 0$ , we have  $I_T^d[\hat{V}_\theta] = E[\hat{V}_\theta | \mathbf{S}_1] = \mathbb{E}_n[V_\theta(S_1)]$ , which does not depend on the sampling policy  $\nu$ . Moreover, we have

$$\text{var}(\mathbb{E}_n[V_\theta(S_1)]) = \frac{1}{n} \text{var}(V_\theta(S_1)).$$

On the other hand, for  $t = 1, \dots, T-1$ ,

$$\begin{aligned} I_t^d[\hat{V}_\theta] &= E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1}] - E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \\ &= \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \dots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} (E[R_t | S_t, A_t, S_{t+1}] + V_\theta^{t+1}(S_{t+1}) - Q_\theta^t(S_t, A_t)) \right]. \end{aligned}$$

The dynamics innovations for  $t = 1, \dots, T-1$  have zero mean and their variance is

$$\text{var} \left[ I_t^d[\hat{V}_\theta] \right] = E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \dots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} (E[R_t | S_t, A_t, S_{t+1}] + V_\theta^{t+1}(S_{t+1}) - Q_\theta^t(S_t, A_t)) \right]^2 \right].$$

For  $t = T$ , we have

$$\begin{aligned} I_T^d[\hat{V}_\theta] &= \hat{V}_\theta - E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_T, \mathbf{A}_T] \\ &= \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \dots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} (R_t - E[R_t | S_t, A_t, S_{t+1}]) \right]. \end{aligned}$$

Since the terms  $R_t - E[R_t | S_t, A_t, S_{t+1}]$  have zero expectation and are uncorrelated, we have

$$\begin{aligned} \text{var} \left( I_T^d[\hat{V}_\theta] \right) &= E_\nu \left[ \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \dots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} (R_t - E[R_t | S_t, A_t, S_{t+1}]) \right]^2 \right] \\ &= \sum_{t=1}^T E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1 | S_1, \theta)}{\nu_1(A_1 | S_1)} \dots \frac{\pi_t(A_t | S_t, \theta)}{\nu_t(A_t | S_t)} (R_t - E[R_t | S_t, A_t, S_{t+1}]) \right]^2 \right]. \end{aligned}$$

Since the rewards  $\mathbf{R}_t$  are independent from the system dynamics given  $(\mathbf{S}_t, \mathbf{A}_t, \mathbf{S}_{t+1})$ , we have

$$\begin{aligned}
& E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} (E[R_t|S_t, A_t, S_{t+1}] + V_\theta^{t+1}(S_{t+1}) - Q_\theta^t(S_t, A_t)) \right]^2 \right. \\
& \quad \left. + \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} (R_t - E[R_t|S_t, A_t, S_{t+1}]) \right]^2 \mid S_1, \dots, S_t, A_t, S_{t+1} \right] \\
& = E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} (R_t + V_\theta^{t+1}(S_{t+1}) - Q_\theta^t(S_t, A_t)) \right]^2 \mid S_1, \dots, S_t, A_t, S_{t+1} \right] \\
& = E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} B_t \right]^2 \mid S_1, \dots, S_t, A_t, S_{t+1} \right]
\end{aligned}$$

Collecting all the terms of  $\text{var}(\hat{V}_\theta^*)$  using the above identity yields

$$\text{var}_\nu(\hat{V}_\theta^*) = \frac{1}{n} \text{var}_\eta(V_\theta(S_1)) + \sum_{t=1}^T E_\nu \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} B_t \right]^2 \right].$$

Since the last terms have zero mean, we can replace their square mean with their variance. Finally, the result (b) of Lemma 4.2.4 yields the desired expression.  $\square$

The variance of the optimal estimator  $\mathbf{\Pi}(\hat{V}_\theta)$  comprises the variance of the immediate rewards and of the system's state dynamics, but does not include the variance of the action innovations, which is  $\text{var}(\hat{V}_\theta - \mathbf{\Pi}(\hat{V}_\theta))$ . If a suboptimal choice  $X_t^* - Y_t$  of the variate  $X_t$  is made, the action innovation variance of the estimator  $\hat{V}_\theta$  is not zero and the estimator variance is (cf. Proposition 4.3.7)

$$\text{var}_\nu(\hat{V}_\theta[X_1^* - Y_1, \dots, X_T^* - Y_T]) = \text{var}_\nu(\hat{V}_\theta^*) + \sum_{t=1}^T E_\nu [I_t^a [Y_t]^2].$$

For example, the naive estimator has an additional variance compared to the optimal estimator amounting to  $\sum_{t=1}^T E_\nu [I_t^a(\hat{V}_\theta)^2]$ .

### Algorithm

We cannot use directly the expression (4.4.3) for the optimal estimator found in Proposition 4.4.4, because we do not know the Q-factors of policy  $\theta$  in the underlying

MDP model  $K$ . Nonetheless, a salient feature of our approach is that we can use any guess for the value of the Q-factors without introducing a bias in the estimator. Indeed,  $\hat{V}_\theta \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \tilde{Q}_\theta^1(S_1, A_1) \right], \dots, \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_T(A_T|S_T, \theta)}{\nu_T(A_T|S_T)} \tilde{Q}_\theta^t(S_T, A_T) \right] \right]$  is an unbiased value estimator for any Q-factor guess  $\tilde{Q}_\theta^t$ . For example, say in medical applications, experts could help design a good guess of the Q-factors from their quantitative and qualitative experience of different treatments. Another approach would be to approximate the Q-factors  $Q_\theta$  by the Q-factors of a related MDP, which is better known. For example, the infinite-horizon value is sometimes easier to assess and approximate. Also, in a queueing network, the analysis of its fluid limit can be used as an approximation to its high-load performance [33].

A general approach to approximate the Q-factors from observed trajectories is to solve the empirical Bellman equation, that is to solve recursively for the look-up table  $\hat{Q}_t(s, a)$  the following system of equations

$$\begin{aligned} \hat{Q}_\theta^T(s_T, a_T) &= \mathbb{E}_n[R_T | S_T = s_T, A_T = a_T], \\ \hat{Q}_\theta^t(s_t, a_t) &= \mathbb{E}_n \left[ R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1} | S_{t+1}, \theta) \hat{Q}_\theta^{t+1}(S_{t+1}, a_{t+1}) \mid S_t = s_t, A_t = a_t \right]. \end{aligned}$$

When  $\hat{Q}_\theta$  is used, instead of  $Q_\theta$ , in the expression of the optimal value estimator (4.4.3), we obtain

$$\hat{V}_\theta \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \hat{Q}_\theta^1, \dots, \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_T(A_T|S_T, \theta)}{\nu_T(A_T|S_T)} \hat{Q}_\theta^T \right] = \sum_{a_1} \pi_1(a_1 | s_1, \theta) \hat{Q}_\theta^1(s_1, a_1),$$

which is a standard estimator based on Bellman equations. Indeed, by the Markovianity of the empirical probability measure  $\mathbb{P}_n$ , the term corresponding to time  $t$  in

Equation (4.4.2) is zero since

$$\begin{aligned}
& \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}) \hat{Q}_\theta^{t+1}(S_{t+1}, a_{t+1}) - \hat{Q}_\theta^t(S_t, A_t) \right) \right] \\
&= \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \mathbb{E}_n \left[ R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}) \hat{Q}_\theta^{t+1}(S_{t+1}, a_{t+1}) - \hat{Q}_\theta^t(S_t, A_t) \mid S_t, A_t \right] \right] \\
&= 0.
\end{aligned}$$

In this case, our analysis does not yield a new estimation method, but it provides a new interpretation for the very common estimator  $\sum_{a_1} \pi_1(a_1|s_1, \theta) \hat{Q}_\theta^1(s_1, a_1)$ .

**Remark 4.4.5.** *Interestingly, this estimator need not have a lower variance than  $\hat{V}_\theta$  as illustrated in Subsection 4.4.4, because it is not the optimal estimator of Proposition 4.4.4 since the true  $Q$ -factor  $Q_\theta$  has been replaced by estimates  $\hat{Q}_\theta$ .*

**Remark 4.4.6.** *Our analysis does not prove that this estimator is unbiased, because  $\hat{Q}_\theta$  is not a fixed function since it is chosen from the training set. However, as the number of observed trajectories increases to infinity, the estimates  $\hat{Q}_\theta$  converge with probability one to  $Q_\theta$ . Hence, our analysis implies that this estimator is asymptotically unbiased and its training set variance converges to the training set variance of the true optimal estimator.*

This estimator of the value from state  $s_1$  requires the observation of at least one trajectory for all actions  $a_1$  that are selected with positive probability in state  $s_1$ . Otherwise, the corresponding  $\hat{Q}_\theta^1(s_1, a_1)$  and consequently the value estimator are not even defined. We will see in Subsection 4.4.4 that this estimator requires a fair amount of observed trajectories to yield good value estimate.

### Optimal sampling policy

We assumed that the sampling policy  $\nu$  is given, but in this part we make a short digression to assess what would be a good sampling policy. Proposition 4.4.4 allows

us to characterize a fixed Markovian policy  $\nu^*$  yielding the optimal estimator  $\hat{V}_\theta^*$  with the minimum variance when using our optimal estimators. Notice, however, that we are not considering adaptive sampling policies (i.e., sampling policies that change as realized trajectories are observed), even though they could reduce the variance of our estimate further - even sometimes make the variance of the estimation procedure equal to zero [39] (provided that some typically unknown characteristics of the system are available to the sampler).

Let  $\mathcal{V}_{t,s}$  be a non-empty closed set of probability distribution over  $\mathcal{A}$ . If we constrain the sampling policy  $\nu$  to be such that  $\nu_t(\cdot|s) \in \mathcal{V}_{t,s}$  for all  $t, s$ , the optimal sampling policy  $\nu^*$  can be found by dynamic programming as shown in the next proposition.

**Proposition 4.4.7.** *Define recursively  $\Delta(t, s)$  for  $t = 1, \dots, T$  and  $s \in \mathcal{S}$  by*

$$\begin{aligned} \Delta(T, s) &= \min_{\nu \in \mathcal{V}_{T,s}} E_K \left[ \frac{\pi_T(A_T|S_T, \theta)}{\nu(A_T)} B_T^2 \mid S_T = s \right], \quad s \in \mathcal{S} \\ &= \min_{\nu \in \mathcal{V}_{T,s}} \sum_{a \in \mathcal{A}} \frac{\pi_T(a|s, \theta)^2}{\nu(a)} E_\theta [(R_T - E[R_T|s, a])^2 \mid S_T = s, A_T = a] \\ \Delta(t, s) &= \min_{\nu \in \mathcal{V}_{t,s}} E_\theta \left[ \frac{\pi_t(A_t|S_t, \theta)}{\nu(A_t)} (B_t^2 + \Delta(t+1, S_{t+1})) \mid S_t = s \right], \quad s \in \mathcal{S} \\ &= \min_{\nu \in \mathcal{V}_{t,s}} \sum_{a \in \mathcal{A}} \frac{\pi_t(a|s, \theta)^2}{\nu(a)} E_K \left[ B_t^2 + \Delta(t+1, S_{t+1}) \mid S_t = s, A_t = a \right] \end{aligned}$$

Then  $\Delta(1, s) = n \cdot \text{var}_\nu(\hat{V}_\theta^*) - \text{var}(V_\theta(S_1))$ .

Furthermore, let  $\nu^*$  be a Markovian policy such that  $\nu_t^*(\cdot|s)$  achieves the minimum in the above recursion. The policy  $\nu^*$  is a sampling policy in the constraint set that yields the optimal estimator  $\hat{V}_\theta^*$  with the minimum training set variance.

*Proof.* From Proposition 4.4.4, we know that the variance of the optimal estimator for a given Markovian sampling policy  $\nu$  is

$$\text{var}_\nu(\hat{V}_\theta^*) = \frac{1}{n} \text{var}(V_\theta(S_1)) + \frac{1}{n} E_\nu \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} B_t^2 \right].$$

It is immediate to check by induction that

$$\Delta(\tau, s) = E_\nu \left[ \sum_{t=\tau}^T \frac{\pi_\tau(A_\tau|S_\tau, \theta)}{\nu_\tau(A_\tau|S_\tau)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} B_t^2 \mid S_\tau = s \right].$$

Thus, the rest of the proposition follows from a classical dynamic programming argument.  $\square$

However, this result is not quite practical as is, because it requires the knowledge of the underlying MDP.

To summarize this subsection, Proposition 4.4.4 characterizes the minimum variance value estimator as a function of the Q-factors, which are unknown in practice. This result suggests to use an educated guess for the Q-factors to achieve a low variance estimator. Oftentimes, this guess is an estimate of the Q-factors. Such estimates are meaningful only if there are enough observed data compared to the model complexity. When the state space is intractably large, or when there are few observed trajectories compared to the state space's size, a good estimate of the Q-factors will not be available. In this case, it makes sense to estimate more accurately a low-dimensional approximation of the Q-factors than to use a poor estimate of the exact Q-factors. This approach is investigated in the next subsection through the concept of constrained value estimators.

## 4.4.2 Constrained value estimator

As mentioned in the paragraph above, there are some situations where it is beneficial to restrict the family of estimators under consideration in order to have an algorithm that performs better in practice. In this subsection, we show that the projection  $\Pi^c(\hat{V}_\theta)$  of the naive estimator has the minimum variance among value estimators of the form  $\hat{V}_\theta[X_1, \dots, X_T]$  with  $X_t \in \mathcal{X}_t$ , and we characterize it algebraically.

The following proposition is the analogous of Proposition 4.4.4 in the constrained case.

**Proposition 4.4.8.**

(a) The estimator  $\hat{V}_\theta^c = \mathbf{\Pi}^c(\hat{V}_\theta)$  belongs to the family of estimators

$$\left\{ \hat{V}_\theta[X_1, \dots, X_T], X_t \in \mathcal{X}_t \right\} \quad (4.4.5)$$

Furthermore, it has the minimum variance in this family.

(b) Let  $\mathcal{Q}_\theta^t$  be a linear subspace of  $\mathcal{E}_t$  of measurable functions of  $(S_t, A_t)$  for  $t = 1, \dots, T$ . Then a minimum-variance estimator in the set defined by (4.4.5) with

$$\mathcal{X}_t = \left\{ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} Y_t(S_t, A_t) \right], Y_t \in \mathcal{Q}_\theta^t \right\}$$

is the projection by  $\mathbf{\Pi}^c$  of the naive estimator  $\hat{V}_\theta$ , that is,

$$\begin{aligned} \hat{V}_\theta^c = & \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \tilde{Q}_\theta^1(S_1, a_1) \right] \\ & + \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}) \tilde{Q}_\theta^{t+1}(S_{t+1}, a_{t+1}) - \tilde{Q}_\theta^t(S_t, A_t) \right) \right], \end{aligned} \quad (4.4.6)$$

with  $\tilde{Q}_\theta^t$  defined as

$$\arg \min_{Q \in \mathcal{Q}_\theta^t} E \left[ \mathbb{E}_n \left[ \frac{\pi_0(A_0|S_0, \theta)}{\nu_0(A_0|S_0)} \dots \frac{\pi_{t-1}(A_{t-1}|S_{t-1}, \theta)}{\nu_{t-1}(A_{t-1}|S_{t-1})} I_t^a \left[ \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} (Q_\theta^t(S_t, A_t) - Q(S_t, A_t)) \right] \right]^2 \right]. \quad (4.4.7)$$

*Proof.* (a) This is a direct application of Proposition 4.3.7.

(b) It is clear that  $\mathcal{X}_t$  defined in the second part of the proposition is a closed linear subspace of  $\mathcal{E}_t$ . From Proposition 4.3.7, we have

$$\hat{V}_\theta^c = \hat{V}_\theta[W_1^c, \dots, W_T^c] = \hat{V}_\theta - \sum_{t=1}^T I_t^a[W_t^c],$$

where  $W_t^c$  is defined by

$$W_t^c = \arg \min_{W_t \in \mathcal{X}_t} E \left[ (I_t^a[\hat{V}_\theta] - I_t^a[W_t])^2 \right].$$

Observe that

$$\begin{aligned}
I_t^a[\hat{V}_\theta] &= E \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} Q_\theta^t(S_t, A_t) \right] \mid \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t \right] \\
&- E \left[ \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} Q_\theta^t(S_t, A_t) \right] \mid \mathbf{S}_1, \dots, \mathbf{S}_t \right] \\
&= \mathbb{E}_n \left[ \frac{\pi_0(A_0|S_0, \theta)}{\nu_0(A_0|S_0)} \dots \frac{\pi_{t-1}(A_{t-1}|S_{t-1}, \theta)}{\nu_{t-1}(A_{t-1}|S_{t-1})} I_t^a \left[ \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} Q_\theta^t(S_t, A_t) \right] \right].
\end{aligned}$$

Since  $W_t^c \in \mathcal{X}_t$ , there exists a function  $\tilde{Q}_\theta^t \in \mathcal{Q}_\theta^t$  such that

$$\tilde{Q}_\theta^t = \arg \min_{Q \in \mathcal{Q}_\theta^t} E \left[ \mathbb{E}_n \left[ \frac{\pi_0(A_0|S_0, \theta)}{\nu_0(A_0|S_0)} \dots \frac{\pi_{t-1}(A_{t-1}|S_{t-1}, \theta)}{\nu_{t-1}(A_{t-1}|S_{t-1})} I_t^a \left[ \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} (Q_\theta^t(S_t, A_t) - Q(S_t, A_t)) \right] \right]^2 \right]$$

and such that  $W_t^c$  defined above takes the form

$$W_t^c = \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \dots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \tilde{Q}_\theta^t(S_t, A_t) \right].$$

Rearranging the terms of  $\hat{V}_\theta[W_1^c, \dots, W_T^c]$  yields the expression (4.4.6) for the optimal constrained estimator.  $\square$

### 4.4.3 Algorithms

Now, we leverage Proposition 4.4.8 to propose practical algorithms for value estimation. To simplify the notation and to ease the comparison with alternative approaches, we consider the on-policy case, i.e., the sampling policy  $\nu = \theta$ .

Furthermore, we will use a linear approximation architecture for the subspaces  $\mathcal{Q}_\theta^t$ , i.e.,

$$\mathcal{Q}_\theta^t = \{ \phi(s, a) \beta_t, \beta_t \in \mathbb{R}^m \},$$

where  $\phi = (\phi_1, \dots, \phi_r)$  is a row vector of features, which are functions from  $\mathcal{S} \times \mathcal{A}$  into  $\mathbb{R}$ . Here, we assumed for the notation's clarity that the features are time independent, but the subsequent analysis extends readily to the case where the features depend on time.

When  $\theta = \nu$ , the optimal on-policy estimator of Proposition 4.4.8 is

$$\hat{V}_\nu^c = \mathbb{E}_n \left[ \sum_{a_1} \nu_1(a_1|S_1) \tilde{Q}_\nu^1(S_1, a_1) \right] + \sum_{t=1}^T \mathbb{E}_n \left[ R_t + \sum_{a_{t+1}} \nu_{t+1}(a_{t+1}|S_{t+1}) \tilde{Q}_\nu^{t+1}(S_{t+1}, a_{t+1}) - \tilde{Q}_\nu^t(S_t, A_t) \right],$$

where Equation (4.4.7) becomes

$$\tilde{Q}_\nu^t = \arg \min_{Q \in \mathcal{Q}_\nu^t} E_\nu \left[ \mathbb{E}_n \left[ I_t^a [Q_\nu^t(S_t, A_t) - Q(S_t, A_t)] \right]^2 \right].$$

Since  $I_t^a[Q] = Q(S_t, A_t) - E[Q|S_t] = Q(S_t, A_t) - \sum_{a_t} \pi_t(a_t|S_t)Q(S_t, a_t)$  is only a function of  $(S_t, A_t)$  and has zero expectation, we can invoke Equation (4.2.1) in Lemma 4.2.4 to define  $\tilde{Q}_\nu^t$  equivalently as

$$\tilde{Q}_\nu^t = \arg \min_{Q \in \mathcal{Q}_\nu^t} \frac{1}{n} E_\nu \left[ I_t^a [Q_\nu^t(S_t, A_t) - Q(S_t, A_t)]^2 \right]. \quad (4.4.8)$$

In practice, we know neither  $E_\nu$ , nor the true Q-factors  $Q_\theta^t$  so that they need to be estimated/approximated. The method of Temporal Differences (TD) (cf. [73] for an introduction to TD) provides a natural method to approximate  $\tilde{Q}_\nu^t(s, a)$  from Equation (4.4.7). In this section, when we refer to the method of TD or Q-learning, we always refer to the version of these method with function approximations by a linear combination of the features  $\phi(s, a)\beta_t$ .

Even though the method of Temporal Differences  $TD(\lambda)$  depends continuously on a parameter  $\lambda$  that trades-off the bias and variance of the approximate Q-factor, we will consider the two extreme cases of  $TD(0)$  (Q-learning) and  $TD(1)$  in this chapter because they capture the full extent of the spectrum. (Of course,  $TD(\lambda)$  could also be used to approximate the Q-factors.)

(a)  $TD(1)$  is a smoothing of the empirical Q-factors  $\hat{Q}_\theta^t$ . More precisely,  $TD(1)$

solves for the weights  $\beta_1$  such that

$$\beta_1 = \arg \min_{\beta} \mathbb{E}_n \left[ \left( \hat{Q}_{\theta}^1(S_1, A_1) - \phi(S_1, A_1)\beta \right)^2 \right]. \quad (4.4.9)$$

Thus, our constrained estimator becomes

$$\begin{aligned} \hat{V}_{\theta}^{c,1} &= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \phi(S_1, a_1) \beta_1 \right] \\ &+ \mathbb{E}_n \left[ R_1 + \sum_{a_2} \pi_2(a_2|S_2, \theta) \hat{Q}_{\theta}^2(S_2, a_2) - \phi(S_1, A_1) \beta_1 \right]. \end{aligned}$$

In the sequel, we will denote  $\hat{Q}_{\theta,1}^t = \phi\beta_t$ , where  $\beta_t$  is obtained by  $TD(1)$ .

- (b) Q-learning ( $TD(0)$ ) in a finite horizon problem [98, 49] computes the weights  $\beta_t$  by solving recursively

$$\beta_T = \arg \min_{\beta \in \mathbb{R}^m} \mathbb{E}_n \left[ (R_T - \phi(S_T, A_T)\beta)^2 \right] \quad (4.4.10)$$

$$\beta_t = \arg \min_{\beta \in \mathbb{R}^m} \mathbb{E}_n \left[ \left( R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) \phi(S_{t+1}, a_{t+1}) \beta_{t+1} - \phi(S_t, A_t)\beta \right)^2 \right].$$

In the sequel, we will denote  $\hat{Q}_{\theta,0}^t = \phi\beta_t$ , where  $\beta_t$  is obtained by  $TD(1)$ . Using the weights  $\beta_t$  computed by Q-learning, we obtain a practical estimator

$$\begin{aligned} \hat{V}_{\pi}^{c,0} &= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \phi(S_1, a_1) \beta_1 \right] \\ &+ \sum_{t=1}^T \mathbb{E}_n \left[ R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) \phi(S_{t+1}, a_{t+1}) \beta_{t+1} - \phi(S_t, A_t) \beta_t \right]. \end{aligned}$$

There are two interesting points to observe about these algorithms:

1. The empirical probability provides a natural weighting of the states so that the approximation focuses on the more likely states.
2. The approximation error of Q-learning with function approximation grows as

the square root of the time horizon, instead of exponentially for arbitrary for arbitrarily loss function. [98].

**Remark 4.4.9.** *Since the control variates are chosen as a function of the training set, our theoretical analysis does not prove that the estimator  $\hat{V}_\pi^{c,0}$  and  $\hat{V}_\pi^{c,1}$  are unbiased. However, in the limit of large training set, the recursions (4.4.10) and (4.4.9) converge with probability one to some approximate  $Q$ -factors. Consequently, our theoretical results imply that the estimators  $\hat{V}_\pi^{c,0}$  and  $\hat{V}_\pi^{c,1}$  are asymptotically unbiased.*

**Remark 4.4.10.** *Proposition 4.4.8 shows that the optimal constrained value estimator calls for an approximation of  $I_t^a[\hat{V}_\theta]$ . But the innovation  $I_t^a[Q_\pi^t](s, a) = Q_\pi^t(s, a) - \sum_\alpha \pi_t(\alpha|s, \theta)Q_\pi^t(s, \alpha)$  can be interpreted as the advantage  $A_t(s, a)$  of action  $a$  in state  $s$  compared to the other actions on average. Thus, it is slightly different from the  $Q$ -factor of the state-action pair  $(s, a)$ . Some work has been dedicated to estimating and/or approximating the advantages [7, 32] and to illustrate the benefit of looking at the advantages rather than the  $Q$ -factors. At a general level, there is little difference between learning the  $Q$ -factors and the advantages since Bellman's equations for the advantages also involve the value functions. However, in specific applications, there can be a substantial benefit of working with the advantages (e.g. [7, 32]).*

As long as the Bellman error term (i.e. the second term in the above expressions) can be computed, these estimators are practical. This is possible when the number  $n$  of observed trajectories in the training set is not too large, which is the regime this chapter is mostly focused on.

The following lemma studies the computational complexity of our estimators.

**Lemma 4.4.11.** (a) *The computation of the optimal weights  $\beta_t$ ,  $t = 1, \dots, T$  in  $TD(0)$  (resp.  $TD(1)$ ), can be done  $O(Tm^3 + Tm^3n|\mathcal{A}|)$  time.*

(b) *Given the weights  $\beta_t$ , the evaluation of the estimator  $\hat{V}_\pi^{c,0}$  (resp.  $\hat{V}_\pi^{c,1}$ ) takes  $O(TAmn)$  time.*

*Proof.* In the proof, we consider the case of  $TD(0)$ . The case of  $TD(1)$  can be done similarly.

(a) For each time index  $t = 1, \dots, T$ , we solve a linear system that corresponds to letting the derivative with respect to  $\beta$  of (4.4.10) equal to zero in order to find the optimal weight  $\beta_t \in \mathbb{R}^m$ . To form the corresponding  $m \times m$  matrix, it takes  $O(n|\mathcal{A}|m)$  time to compute each of the  $m^2$  coefficients. Considering that the inversion of the  $m \times m$  matrix can be done in  $O(m^3)$  time concludes the proof.

(b) For each time  $t = 1, \dots, T$ , the expression of the term of  $\hat{V}_\pi^{c,0}$  corresponding to time  $t$  is done in  $O(nm|\mathcal{A}|)$ . The claim result follows easily.  $\square$

#### 4.4.4 Numerical experiments

In this subsection, we compare the on-policy estimation accuracy of our approach with temporal difference methods, namely  $TD(0)$  (Q-learning) and  $TD(1)$ . Specifically, we will consider seven estimation methods:

- “Naive estimator:” the importance sampling value estimator  $\hat{V}_\theta$ , which reduces to the sample mean of the trajectory rewards in the on-policy case,
- “Optimal estimator:” the true optimal unconstrained estimator  $\hat{V}_\theta^* = \hat{V}_\theta[Q_\theta^1, \dots, Q_\theta^T]$  (which knows the true factors  $Q_\theta^t$ ),
- “Empirical Q estimator:” the optimal unconstrained estimator where the Q-factors are approximated by the empirical Q-factors  $\hat{Q}_\theta^t$ , i.e.,

$$\hat{V}_\theta^q = \hat{V}_\theta[\hat{Q}_\theta^1, \dots, \hat{Q}_\theta^T] = \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \hat{Q}_\theta^1(S_1, a_1) \right],$$

- “Q-learning estimator:” the estimator

$$\hat{v}_\theta^0 = \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \tilde{Q}_{\theta,0}^1(S_1, a_1) \right],$$

where  $\tilde{Q}_{\theta,0}^1$  are approximated Q-factors obtained from Q-learning (TD(0)) by (4.4.10).

- “TD(1) estimator:” the estimator

$$\hat{v}_\theta^1 = \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \tilde{Q}_{\theta,1}^1(S_1, a_1) \right],$$

where  $\tilde{Q}_{\theta,1}^1$  are approximated Q-factors obtained from TD(1) by (4.4.9),

- “Q-learning control variate:” the optimal estimator where the Q-factors are replaced by the approximated Q-factors  $\hat{Q}_{\theta,0}^t$  of Q-learning, i.e.,

$$\hat{V}_\theta^{c,0} = \hat{V}_\theta[\tilde{Q}_{\theta,0}^1, \dots, \tilde{Q}_{\theta,0}^T],$$

- “TD(1) control variate:” the optimal estimator where the Q-factors are replaced by the approximated Q-factors  $\hat{Q}_{\theta,1}^t$  of TD(1), i.e.,

$$\hat{V}_\theta^{c,1} = \hat{V}_\theta[\tilde{Q}_{\theta,1}^1, \dots, \tilde{Q}_{\theta,1}^T].$$

The performances of the different estimators will be compared numerically as a function of the cardinality of the training sets on two simple MDP models described thereafter. Unless specified otherwise, the performance of a value estimator is judged on the basis of the empirical mean squared estimation error evaluated from many training sets. Thus, both bias and variance are penalized.

We will illustrate three points:

- When there are enough observed trajectories, the variance of the naive estimator  $\hat{V}_\theta$  is significantly larger than the variance of the estimators  $\hat{V}_\theta[\tilde{Q}_\theta^1, \dots, \tilde{Q}_\theta^T]$ , where  $\tilde{Q}_\theta^t$  is some estimate of the true Q-factors  $Q_\theta^t$  (e.g., the empirical Q-factors  $\hat{Q}_\theta^t$ , or the approximated factors  $\hat{Q}_{\theta,0}^t$  or  $\hat{Q}_{\theta,1}^t$  obtained from  $TD(0)$  and  $TD(1)$ ). Whereas Proposition 4.4.4 shows that the optimal estimator  $\hat{V}_\theta[Q_\theta^1, \dots, Q_\theta^T]$  has lower variance than the naive estimator  $\hat{V}_\theta$ , the unknown Q-factors  $Q_\theta^t$  need to be estimated, and the noise in their estimation could worsen the estimation error of  $\hat{V}_\theta[\tilde{Q}_\theta^1, \dots, \tilde{Q}_\theta^T]$ . In fact, we will see examples where the estimates  $\tilde{Q}_\theta^t$  of

the Q-factors are so noisy that the resulting estimators  $\hat{V}_\theta[\tilde{Q}_\theta^1, \dots, \tilde{Q}_\theta^T]$  are worse than the naive one, but this situation occurs when there is little data available to estimate the Q-factors.

- When the number of observed trajectories is small, it is more efficient to estimate approximate (regularized) Q-factors than to estimate less accurately a look-up table estimate of the Q-factors (provided the approximate space is not poorly chosen). Thus, the constrained estimator can add value in practice when the training set is small.
- The temporal difference methods  $TD(\lambda)$  with function approximation, a fortiori Q-learning and  $TD(1)$ , can estimate the Q-factors with few observed trajectories, but they achieve this by trading-off a lower variance with potential bias. In fact, if the features  $\phi$  are not appropriately chosen, the  $TD(\lambda)$  estimators can have an arbitrarily large bias. Hence,  $TD(\lambda)$  with function approximation is not appropriate for accurate value estimation. In contrast, our constrained estimators can take advantage of the biased estimates of the Q-factors generated by  $TD(\lambda)$  to yield an unbiased value estimate with smaller variance than the naive estimator, even when there is little available observations.

### Line MDP

Let us consider a “linear” MDP comprising  $(S + 1)$  states,  $\mathcal{S} = \{0, 1, \dots, S\}$ . The state  $S$  is absorbing. In state  $k = 0, \dots, S - 1$ , the controller can either choose the action  $I$  and stay in state  $k$  with probability one or choose the action  $M$  to move to the state  $k + 1$  with probability one. All actions have a reward of one, except the actions in the absorbing state  $S$ , which have zero cost. In addition, the system is initialized at state 0. A graphical illustration of the MDP with  $S = 3$  is shown in Figure 4-1.

Let  $\theta = \nu$  be the policy that chooses action  $I$  and  $M$  with probability 1/2 in all states.

In our experiments with the line MDP, we used two features. The first is  $\phi_1(k, a) =$

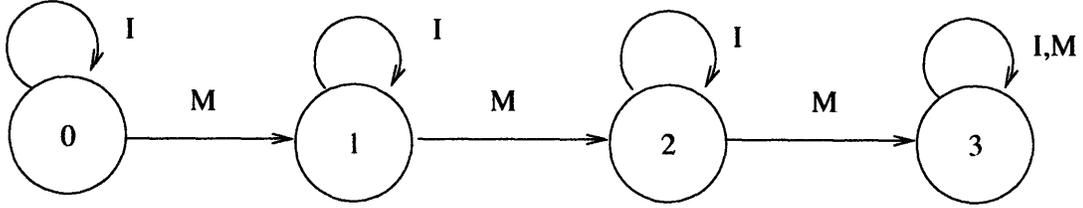


Figure 4-1: Graphical representation of the line MDP with  $S = 3$ .

$S - k$  and the second is  $\phi_2(k, a) = (-1)^{\mathbf{1}\{a=I\}}$ . When the time horizon is large, the finite horizon value functions are well-approximated by the infinite horizon ones. By solving Bellman's equations, we obtain the infinite-horizon value functions  $\bar{V}_v(k) = 2(S - k)$  for  $k \in \mathcal{S}$ . We deduce that the corresponding advantages are  $A(k, I) = 1$  and  $A(k, M) = -1$  for  $k = 0, \dots, S - 1$ . As a result, the proposed features allow a good linear approximation of the Q-factors.

In this example, the optimal estimator  $\hat{V}_\theta[Q_\theta^1, \dots, Q_\theta^T]$ , which uses the unknown Q-factors  $Q_\theta^t$ , is equal to the value  $V_\theta$  with probability one since the state dynamics and the rewards are deterministic given the current state-action pair (and thus, all Bellman errors  $R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) Q_\theta^{t+1}(S_{t+1}, a_{t+1}) - Q_\theta^t(s_t, a_t)$  are zero with probability one). Consequently, it has zero squared error independently from the number of trajectories in the training set. Hence, the estimation error of our algorithms needs to be entirely explained by the noise in the estimation of the Q-factors.

When the time horizon  $T = 6$  and  $S = 2$ , Figures 4-2 and 4-3 compare the empirical squared error between different unbiased value estimates and the true value for different numbers of observed trajectories in the training set.

When there are few observed trajectories (illustrated by Figure 4-3), Q-learning provides a reasonable estimate  $\tilde{Q}_{\theta,0}^t$  of approximate Q-factors the fastest. As a result, the mean squared estimation error of the value estimator  $\hat{V}_\theta[\tilde{Q}_{\theta,0}^1, \dots, \tilde{Q}_{\theta,0}^T]$  decreases very quickly. In the present example, this estimator using 6 trajectories has two times less variance than the naive estimator, and with 20 trajectories it has five times less variance. On the other hand, the estimator  $\hat{V}_\theta[\hat{Q}_\theta^1, \dots, \hat{Q}_\theta^T]$  without Q-function

approximation requires at least 30 trajectories in order to have a lower mean squared error than the naive estimator. However, as the number of trajectories in the training set increases, Figure 4-2 shows that the look-up table estimation of the Q-factors improves and eventually (for  $n \geq 75$ ) yields a better estimate than the other methods that rely on approximated Q-factors.

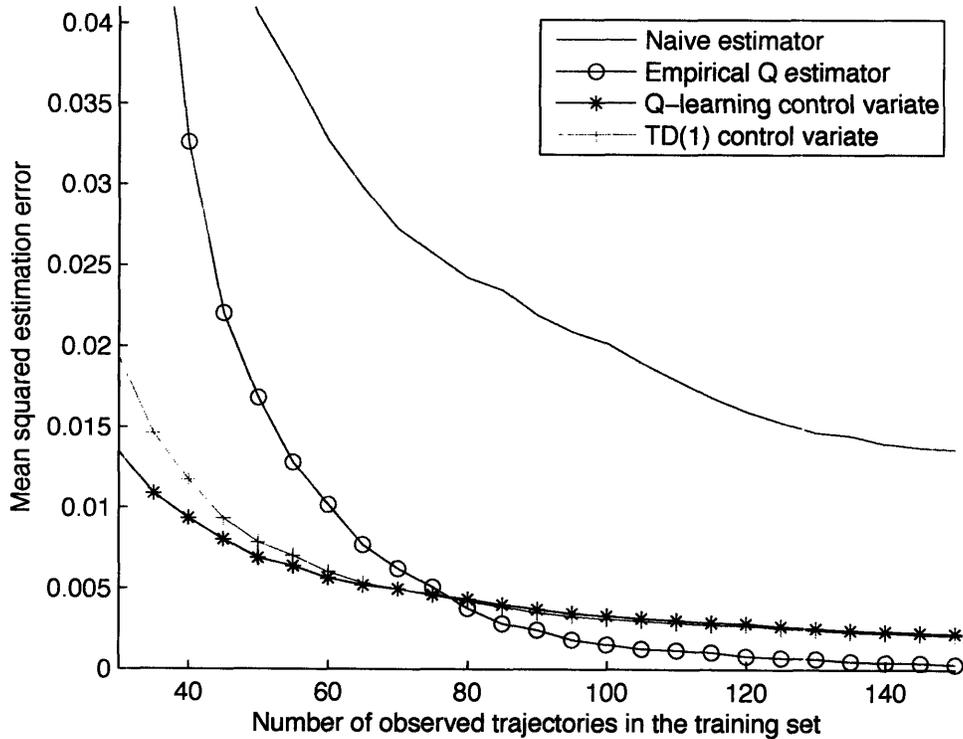


Figure 4-2: Empirical squared error between different value estimators and the true value on the line MDP, as a function of the number of observed trajectories per training set, estimated with 1000 random training sets. The true value function is  $V_\theta = 4.5625$  for  $T = 6$  and  $S = 3$ .

Now, we illustrate with the line MDP that the estimator based on Q-learning and  $TD(1)$ , namely  $\hat{v}_\theta^0$  and  $\hat{v}_\theta^1$ , can have arbitrarily large bias. In order to make the shortcomings of  $TD(\lambda)$  obvious, we restrict the approximate Q-factors computed to be a linear function of  $\phi_2$ . Since the policy  $\theta$  selects action  $I$  and  $M$  with equal probability and that  $\phi_2(k, M) = -\phi_2(k, I)$ , the  $TD(\lambda)$  estimators will always yield

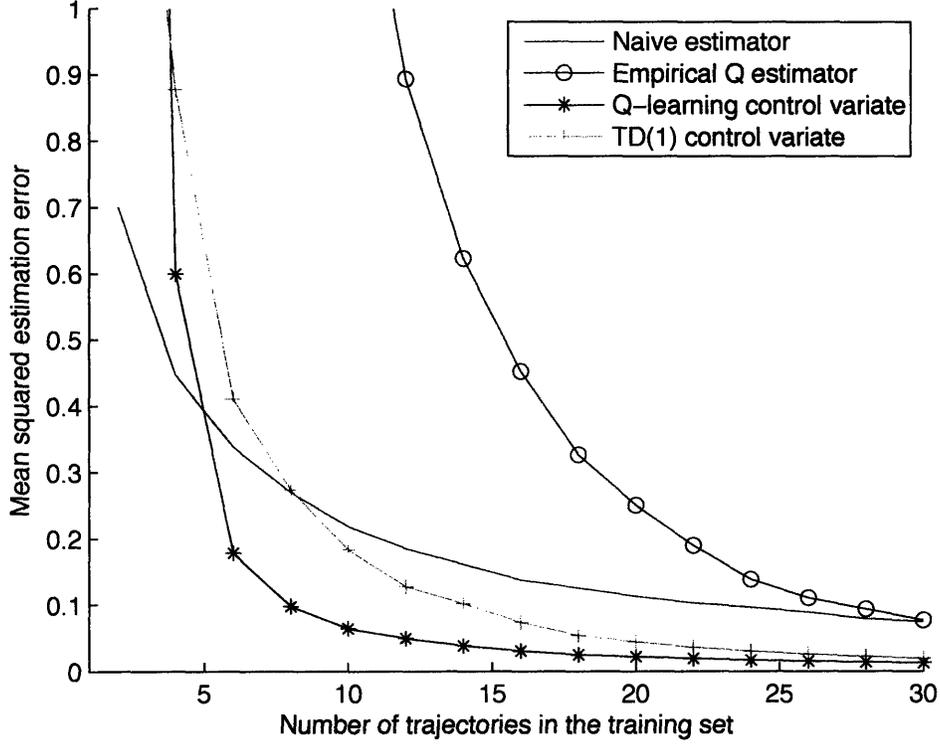


Figure 4-3: Empirical squared error between different value estimators and the true value on the line MDP, as a function of the number of observed trajectories per training set, estimated with 1000 random training sets.

zero for all training sets, e.g.,

$$\begin{aligned} \hat{v}_\theta^0 &= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \tilde{Q}_{\theta,0}^1(S_1, a_1) \right] \\ &= \frac{1}{2} \phi_2(0, I) \beta - \frac{1}{2} \phi_2(0, I) \beta = 0. \end{aligned}$$

while the true value is  $V_\theta = 4.5625$ . In contrast, our estimators are unbiased and eventually converge to the true value  $V_\theta$  as the training set increases to infinity.

## Inventory problem

The second MDP we consider models a simple inventory problem over a finite time horizon  $T = 5$ . A distributor sells a single product at a fixed unit price,  $p = 2$ . The numbers of units demanded at each period are random IID, with an unknown distribution with support  $\{0, \dots, D\}$ . Demand is satisfied by the distributor as long as he has inventory, which is replenished at the beginning of each period at a unit cost  $c = 1$ . If a demanded unit cannot be supplied by the distributor, then the demand is lost. Let  $M = 5$  be the maximum inventory the distributor can store. This multi-period newsvendor problem can be modeled by an MDP, where the state in  $\{0, \dots, M\}$  is the inventory at the beginning of the period and the action in  $\{0, \dots, D\}$  is the replenishment order size.

In the numerical experiments, the demand is 0, 1, 2, or 3 with probability 0.1, 0.5, 0.35, and 0.05, respectively. We let the policy  $\theta = \nu$  choose all actions with equal probability (or up to the inventory capacity  $M$  if the order is too large). The initial state has zero inventory. The value is  $V_\theta = -6.6739$ . Furthermore, we use the two features  $\phi_1(s, a) = -a$  and  $\phi_2(s, a) = M - s$  to approximate the Q-factors.

Figure 4-4 compares the empirical squared error of our different estimators as a function of the number of trajectories in the training set.

- The optimal estimator  $\hat{V}_\theta[Q_\theta^1, \dots, Q_\theta^T]$ , which knows the true Q-factors, reduces dramatically the estimation error, but the unconstrained estimator  $\hat{V}_\theta[\hat{Q}_\theta^1, \dots, \hat{Q}_\theta^T] = \sum_{a=0}^3 \pi_1(a|0) \hat{Q}_\theta^1(0, a)$ , which uses the empirical Q-factors  $\hat{Q}_\theta^t$ , has the highest estimation error, even much higher than the naive importance sampling estimator. The unconstrained estimator requires at least 200 trajectories per training set to have a similar estimator error as the naive estimator.
- Our constrained estimators using the estimates of approximate Q-factors from Q-learning and  $TD(1)$  have a much lower error than the naive estimator. Since the true Q-factors are close to the approximation space, the constrained estimators perform almost as well as the optimal estimator.

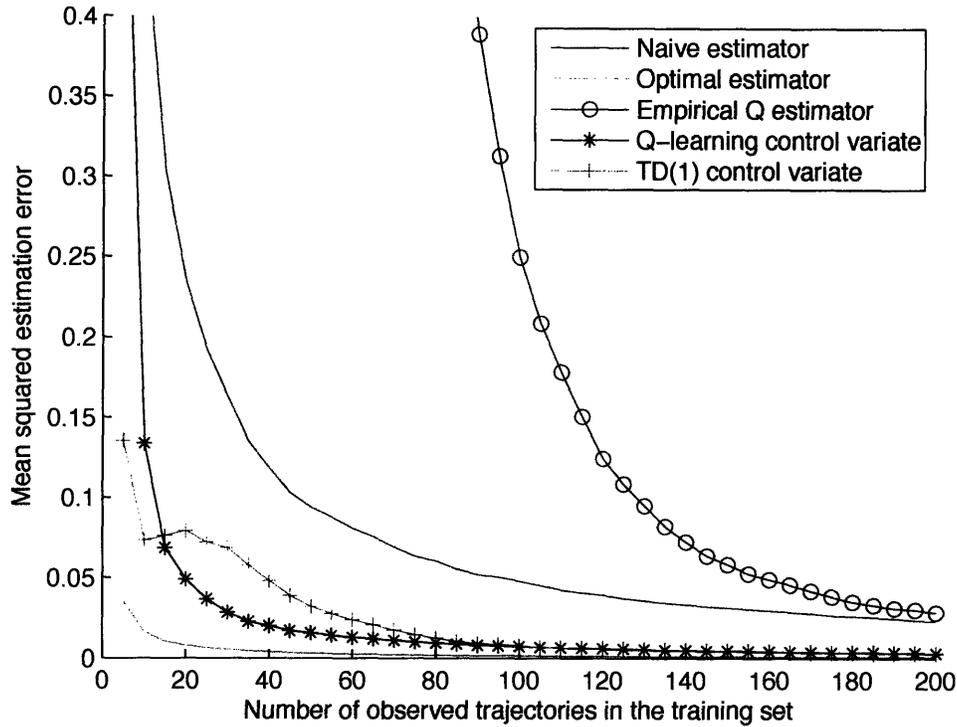


Figure 4-4: Empirical squared error between different value estimators and the true value of the inventory problem, as a function of the number of observed trajectories per training set, estimated with 1000 random training sets. By comparison, the true value is  $V_\theta = -6.6$ .

Figure 4-5 plots the Q-learning and TD(1) estimators in addition to the other estimators plotted in Figure 4-4. We can see that they perform quite poorly: the Q-learning estimator performs similarly to the naive estimator, while the TD(1) estimator follows the unconstrained estimator, which is the worst by far in these experiments. Interestingly, our constrained estimators, which relies on the Q-learning and TD(1) estimation of approximate Q-factors, are performing much better.

Nonetheless, the Q-learning and TD(1) estimators have a squared estimation error decreasing to almost zero because the true Q-factors are well approximated by our features. When we use only the feature  $\phi_1$  to approximate the Q-factors, the bias introduced by the inexact approximation architecture becomes more apparent as seen on Figure 4-5. The mean squared estimation error is lower bounded by the bias

squared. The squared bias of Q-learning (resp. of  $TD(1)$ ) is approximately 1.2 (resp. 0.75) in this example.

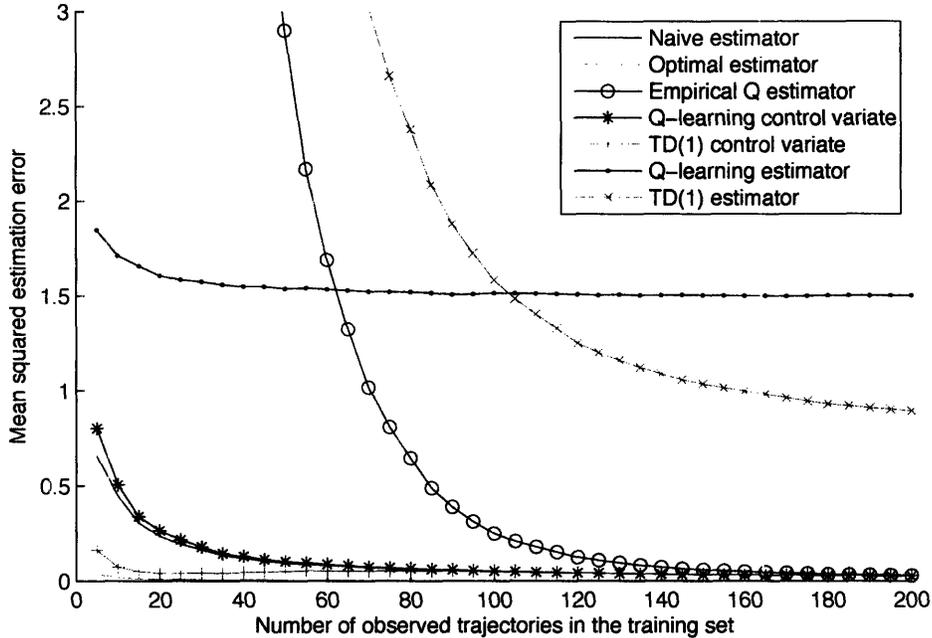


Figure 4-5: Empirical squared error between different value estimators and the true value of the inventory problem, as a function of the number of observed trajectories per training set, estimated with 1000 random training sets. In this experiment, the Q-factors are approximated only using the feature  $\phi_1$ .

In this section, we characterized theoretically optimal value estimators in the unconstrained and constrained cases. When there is little data available, we illustrated numerically that our constrained estimators outperform competing approaches by taking advantage of regularized estimates of the Q-factors to yield lower variance value estimate without being biased.

## 4.5 Estimation of the gradient of policy value

In this section, we define an unconstrained and a constrained class of unbiased estimators of the value gradient and we identify a minimum-variance estimator in each

class. In addition, we compare our approach with other ones from the reinforcement learning literature.

We will consider the estimation of the derivative of policy value with respect to one real-valued parameter  $\theta$ . The results of this section are also relevant for problems where the policy depends on more than one parameter, because they can be applied to estimate component-wise or directional derivatives. However, our setting will not capture the covariance between derivative estimates component by component.

We add the following technical assumptions in this section. Among others, these assumptions ensure that the value function is continuously differentiable on  $\Theta$ .

**Assumption 4.5.1.**

1. The set  $\Theta = (\theta_0, \theta_1)$ ,  $\theta_0 < \theta_1$  is an open interval of the real line  $\mathbb{R}$ .
2. The functions  $\log \pi_t(a|s, \theta)$  are continuously differentiable for all  $t, s, a$ , as a function of  $\theta \in \Theta$ . Consequently, for all fixed training sets,  $\hat{V}_\theta$  is continuously differentiable as a function of  $\theta \in \Theta$ .
3. The random function  $\frac{\partial \hat{V}_\theta}{\partial \theta}$  of the training set is integrable for all  $\theta \in \Theta$ , and  $E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right]$  is a continuous function of  $\theta \in \Theta$ , so that we have

$$\frac{\partial}{\partial \theta} \left( \int_{\theta_0}^{\theta} E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right] d\theta \right) = E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right].$$

In addition, there holds

$$\int_{\theta_0}^{\theta_1} E_\nu \left[ \left| \frac{\partial \hat{V}_\theta}{\partial \theta} \right| \right] d\theta < +\infty.$$

We will assume that the same holds for the conditional expectations  $E_\nu[\cdot | \mathbf{s}_1, \dots, \mathbf{s}_t]$  and  $E_\nu[\cdot | \mathbf{s}_1, \dots, \mathbf{s}_t, \mathbf{a}_t]$ , for all  $t$  and  $\mathbf{s}_1, \dots, \mathbf{s}_T, \mathbf{a}_T$ .

### 4.5.1 Characterization of optimal gradient estimators

Under the assumptions stated at the beginning of the section, the estimator  $\hat{V}_\theta$  defined in Equation (4.4.1) is continuously differentiable for all training sets. Hence, we can define the estimator of the value derivative  $\widehat{\frac{\partial V_\theta}{\partial \theta}}$  by

$$\begin{aligned} \widehat{\frac{\partial V_\theta}{\partial \theta}} &= \frac{\partial \hat{V}_\theta}{\partial \theta} \\ &= \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \cdots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) R_t \right]. \end{aligned} \quad (4.5.1)$$

The estimator  $\widehat{\frac{\partial V_\theta}{\partial \theta}}$  is an unbiased estimator of the value gradient  $\frac{\partial V_\theta}{\partial \theta}$  for all  $\theta \in \Theta$ . Indeed, by Fubini's theorem (under Assumption 4.5.1), we have

$$\int_{\theta_0}^{\theta} E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right] d\theta = E_\nu \left[ \int_{\theta_0}^{\theta} \frac{\partial \hat{V}_\theta}{\partial \theta} d\theta \right] = E_\nu \left[ \hat{V}_\theta \right]. \quad (4.5.2)$$

Consequently,

$$\frac{\partial V_\theta}{\partial \theta} = \frac{\partial E_\nu[\hat{V}_\theta]}{\partial \theta} = \frac{\partial}{\partial \theta} \left( \int_{\theta_0}^{\theta} E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right] d\theta \right) = E_\nu \left[ \frac{\partial \hat{V}_\theta}{\partial \theta} \right] = E_\nu \left[ \widehat{\frac{\partial V_\theta}{\partial \theta}} \right],$$

where the first equality follows from the unbiasedness of the value estimator  $\hat{V}_\theta$ , the second from (4.5.2), and the fourth from the definition of the gradient estimator  $\widehat{\frac{\partial V_\theta}{\partial \theta}}$ .

Nonetheless, the gradient estimator  $\widehat{\frac{\partial V_\theta}{\partial \theta}}$  might have large variance and we would like to reduce it using control variates. Similar to the previous section, we want to find an estimator of the derivative of the value function with minimal variance, within the set of unbiased estimators

$$\left\{ \frac{\partial \widehat{V}_\theta}{\partial \theta}[X_1, \dots, X_T], X_t \in \mathcal{E}_t \right\} \text{ or } \left\{ \frac{\partial \widehat{V}_\theta}{\partial \theta}[X_1, \dots, X_T], X_t \in \mathcal{X}_t \right\} \quad (4.5.3)$$

for the unconstrained and constrained case, respectively. We assume that the set  $\mathcal{X}_t \subset \mathcal{E}_t$  is a closed linear subspace of  $\mathcal{E}_t$ , for  $t = 1, \dots, T$ , in the constrained case.

**Remark 4.5.2.** *In this section, we still use the notations  $Z[X_1, \dots, X_t]$  introduced*

in Section 4.3. Now, the random variables are often defined as derivatives of other random variables, making the notations more prone to confusion. Sometimes, we will use parenthesis to ease the parsing of the expression,

**Proposition 4.5.3.** *In both unconstrained and constrained cases, the optimal estimator of the value gradient is the gradient of the optimal value estimator.*

*In the unconstrained case we have*

$$\frac{\partial \Pi(\hat{V}_\theta)}{\partial \theta} = \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) \left[ \left( \frac{\partial X_1^*}{\partial \theta} \right), \dots, \left( \frac{\partial X_T^*}{\partial \theta} \right) \right] = \Pi \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right),$$

where  $(X_t^*)_{1 \leq t \leq T}$  are such that  $\Pi(\hat{V}_\theta) = \hat{V}_\theta[X_1^*, \dots, X_T^*]$ .

*In the constrained case, if the constraint sets  $\mathcal{X}_t$  are independent of  $\theta$ , we have*

$$\frac{\partial \Pi^c(\hat{V}_\theta)}{\partial \theta} = \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) \left[ \left( \frac{\partial X_1^c}{\partial \theta} \right), \dots, \left( \frac{\partial X_T^c}{\partial \theta} \right) \right] = \Pi^c \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right),$$

where  $(X_t^c)_{1 \leq t \leq T}$  are such that  $\Pi^c(\hat{V}_\theta) = \hat{V}_\theta[X_1^c, \dots, X_T^c]$ .

*Proof.* From Propositions 4.3.5 and 4.3.7, we deduce that the optimal estimators exist and are obtained as the image of the naive estimator  $\widehat{\frac{\partial V_\theta}{\partial \theta}}$  by the projection  $\Pi$  in the unconstrained case and by the projection  $\Pi^c$  in the constrained case.

First, we prove that

$$\Pi \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) = \frac{\partial \Pi(\hat{V}_\theta)}{\partial \theta}$$

and

$$\Pi^c \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) = \frac{\partial \Pi^c(\hat{V}_\theta)}{\partial \theta}.$$

It is enough to prove that we can interchange the projection and differentiation operators.

Recall that the operator  $\Pi$  is defined by  $\Pi(Z) = Z - \sum_{t=1}^T (E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t])$ . Using Fubini's theorem as we did at the beginning of the subsection, we can interchange the conditional expectations and the differentiation with respect

to  $\theta$ . Thus, it follows that we can interchange the operators  $\mathbf{\Pi}$  and  $\frac{\partial}{\partial\theta}$  evaluated on  $\hat{V}_\theta$ .

In the case of the constrained operator  $\mathbf{\Pi}^c$ , the argument needs to be adapted slightly. Recall that  $\mathbf{\Pi}^c(Z) = Z - \sum_{t=1}^T \mathbf{\Pi}_t (E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[Z|\mathbf{S}_1, \dots, \mathbf{S}_t])$ .

We establish now that

$$\frac{\partial \mathbf{\Pi}_t \left( E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \right)}{\partial \theta} = \mathbf{\Pi}_t \left( \frac{\partial E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t]}{\partial \theta} \right).$$

The orthogonal projections  $\mathbf{\Pi}_t$  are non-expansive for the  $L_2$ -norm and a fortiori they are continuous for the  $L_2$ -norm. Hence, if we have for the  $L_2$ -norm that

$$\frac{\partial E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t]}{\partial \theta} = \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left( E[\hat{V}_{\theta+\epsilon} | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \right), \quad (4.5.4)$$

then we can conclude by  $L_2$ -continuity of  $\mathbf{\Pi}_t$  that

$$\begin{aligned} \mathbf{\Pi}_t \left( \frac{\partial E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t]}{\partial \theta} \right) &= \mathbf{\Pi}_t \left( \lim_{\epsilon \rightarrow 0} \frac{1}{\epsilon} \left( E[\hat{V}_{\theta+\epsilon} | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \right) \right) \\ &= \lim_{\epsilon \rightarrow 0} \mathbf{\Pi}_t \left( \frac{1}{\epsilon} \left( E[\hat{V}_{\theta+\epsilon} | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] - E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \right) \right) \\ &= \frac{\partial \mathbf{\Pi}_t \left( E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t] \right)}{\partial \theta}. \end{aligned}$$

Under Assumption 4.5.1, Equation (4.5.4) holds with certainty, i.e., for all  $\mathbf{s}_1, \dots, \mathbf{s}_t, \mathbf{a}_t$ . Since the state and action spaces are finite and the training set as a finite number of trajectories, the random variable  $E[\hat{V}_\theta | \mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t]$  belongs to the finite-dimensional subspace of  $\mathcal{E}$  of functions of  $\mathbf{S}_1, \dots, \mathbf{S}_t, \mathbf{A}_t$  alone. Therefore, the limit in Equation (4.5.4) also holds in a  $L_2$ -sense on this subspace, and thus we can interchange differentiation and  $\mathbf{\Pi}_t$  on this subspace.

As a result, we can also interchange the operators  $\mathbf{\Pi}^c$  and  $\frac{\partial}{\partial\theta}$  evaluated on  $\hat{V}_\theta$ .

Now, we show that  $\left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) \left[ \left( \frac{\partial X_1^*}{\partial \theta} \right), \dots, \left( \frac{\partial X_T^*}{\partial \theta} \right) \right] = \mathbf{\Pi} \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right)$ , where the functions

$X_t^*$  are such that  $\Pi(\hat{V}_\theta) = \hat{V}_\theta[X_1^*, \dots, X_T^*]$ . We have

$$\begin{aligned} \Pi\left(\frac{\partial \hat{V}_\theta}{\partial \theta}\right) &= \frac{\partial \Pi(\hat{V}_\theta)}{\partial \theta} = \frac{\partial \hat{V}_\theta}{\partial \theta} - \sum_{t=1}^T \frac{\partial I_t^a[X_t^*]}{\partial \theta} \\ &= \frac{\partial \hat{V}_\theta}{\partial \theta} - \sum_{t=1}^T I_t^a\left(\frac{\partial X_t^*}{\partial \theta}\right) \\ &= \left(\frac{\partial \hat{V}_\theta}{\partial \theta}\right) \left[\left(\frac{\partial X_1^*}{\partial \theta}\right), \dots, \left(\frac{\partial X_T^*}{\partial \theta}\right)\right]. \end{aligned}$$

A analogous argument takes care of the constrained case.  $\square$

**Corollary 4.5.4.** *In the unconstrained case, the optimal estimator for the gradient of the value is*

$$\begin{aligned} \frac{\partial \widehat{V}_\theta^*}{\partial \theta} &= \sum_{t=1}^T \mathbb{E}_n \left[ E_\nu \left[ \sum_{a_t} \pi_t(a_t|S_t, \theta) \frac{\partial \log \pi_t(a_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, a_t)|S_1 \right] \right] \quad (4.5.5) \\ &+ \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \dots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) B_t \right] \\ &+ \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \times \right. \\ &\left. \left( \sum_{\tau=t+1}^T \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) E_\nu \left[ \sum_{a_\tau} \pi_\tau(a_\tau|S_\tau, \theta) \frac{\partial \log \pi_\tau(a_\tau|S_\tau, \theta)}{\partial \theta} Q_\theta^\tau(S_\tau, a_\tau)|S_{t+1}, a_{t+1} \right] \right. \right. \\ &\left. \left. - \sum_{\tau=t+1}^T E_\nu \left[ \sum_{a_\tau} \pi_\tau(a_\tau|S_\tau, \theta) \frac{\partial \log \pi_\tau(a_\tau|S_\tau, \theta)}{\partial \theta} Q_\theta^\tau(S_\tau, a_\tau)|S_t, A_t \right] \right) \right]. \end{aligned}$$

**Remark 4.5.5.** *The first term in (4.5.5) approximates directly the derivative of the value function  $V_\theta$ , which is*

$$\frac{\partial V_\theta}{\partial \theta} = E_\theta \left[ \sum_{t=1}^T \sum_{a_t} \pi_t(a_t|S_t, \theta) \frac{\partial \log \pi_t(a_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, a_t) \right].$$

*The other terms have zero mean when the  $Q$ -factors are known and by Lemma 4.2.4 their variance is order of  $O(1/n)$ . Thus, as  $n$  increases to infinity, these terms converge to zero.*

*Proof.* Differentiating Equation (4.4.3) yields

$$\begin{aligned} \frac{\partial (\mathbf{\Pi}(\hat{V}_\theta))}{\partial \theta} &= \mathbb{E}_n \left[ \sum_{a_1} \pi_1(a_1|S_1, \theta) \left( \frac{\partial \log \pi_1(a_1|S_1, \theta)}{\partial \theta} + \frac{\partial Q_\theta^1(S_1, a_1)}{\partial \theta} \right) \right] \\ &+ \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \cdots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) B_t + \frac{\partial B_t}{\partial \theta} \right) \right]. \end{aligned}$$

Differentiating Bellman's equations

$$Q_\theta^t(s, a) = E_\nu \left[ R_t + \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) Q_\theta^{t+1}(S_{t+1}, a_{t+1}) \mid S_t = s, A_t = a \right]$$

yields, since the expectation  $E_\nu[R_t|S_t, A_t]$  is independent of  $\theta$ ,

$$\begin{aligned} \frac{\partial Q_\theta^t(s, a)}{\partial \theta} &= E_\nu \left[ \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) \left( \frac{\partial \log \pi_{t+1}(a_{t+1}|S_{t+1}, \theta)}{\partial \theta} Q_\theta^{t+1}(S_{t+1}, a_{t+1}) \right. \right. \\ &\quad \left. \left. + \frac{\partial Q_\theta^{t+1}(S_{t+1}, a_{t+1})}{\partial \theta} \right) \mid S_t = s, A_t = a \right]. \end{aligned}$$

By induction, this expression yields

$$\frac{\partial Q_\theta^t(s, a)}{\partial \theta} = \sum_{\tau=t+1}^T E_\nu \left[ \sum_{a_\tau} \pi_\tau(a_\tau|S_\tau, \theta) \frac{\partial \log \pi_\tau(a_\tau|S_\tau, \theta)}{\partial \theta} Q_\theta^\tau(S_\tau, a_\tau) \mid S_t = s, A_t = a \right].$$

Substituting this expression in the derivative of the Bellman error  $B_t$ , which is

$$\begin{aligned} \frac{\partial B_t}{\partial \theta} &= \sum_{a_{t+1}} \pi_{t+1}(a_{t+1}|S_{t+1}, \theta) \left( \frac{\partial \log \pi_{t+1}(a_{t+1}|S_{t+1}, \theta)}{\partial \theta} Q_\theta^{t+1}(S_{t+1}, a_{t+1}) + \frac{\partial Q_\theta^{t+1}(S_{t+1}, a_{t+1})}{\partial \theta} \right) \\ &\quad - \frac{\partial Q_\theta^t(S_t, A_t)}{\partial \theta}, \end{aligned}$$

yields the claimed result.  $\square$

Furthermore, when we approximate the expectation with respect to  $\mu$  in (4.5.5) by the Markovian empirical expectation  $\mathbb{E}_n$ , the last term is equal to zero (in contrast to case of the non-Markovian empirical distribution  $\tilde{\mathbb{P}}_n$ ) so that the optimal estimator

is

$$\begin{aligned} & \sum_{t=1}^T \mathbb{E}_n \left[ E_\theta \left[ \sum_{a_t} \pi_t(a_t|S_t, \theta) \frac{\partial \log \pi_t(a_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, a_t) | S_1 \right] \right] \\ & + \sum_{t=1}^T \mathbb{E}_n \left[ \frac{\pi_1(A_1|S_1, \theta)}{\nu_1(A_1|S_1)} \cdots \frac{\pi_t(A_t|S_t, \theta)}{\nu_t(A_t|S_t)} \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \cdots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) B_t \right]. \end{aligned}$$

In the next two subsections, we will specialize our results to the on-policy case, that is  $\theta = \nu$ , in order to be in the same setting as the paper of Greensmith et al [31], which will serve as a benchmark for our approach.

## 4.5.2 Comparison with the baseline approach

The baseline method is also a control variate method to reduce the variance of the gradient estimates (cf. [31] and references therein). It relies on the observation that, since the probabilities  $\pi_t(a|s, \theta)$  add to one for all states  $s \in \mathcal{S}$ ,

$$E_\theta \left[ \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} | S_t \right] = \sum_a \pi_t(a|S_t, \theta) \frac{\partial \pi_t(a|S_t, \theta)}{\partial \theta} / \pi_t(A_t|S_t, \theta) = \frac{\partial \sum_a \pi_t(a|S_t, \theta)}{\partial \theta} = 0.$$

As a result, one of the approaches in [31] uses  $\frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \beta_t(S_t)$  as a control variate for the on-policy estimation of the value gradient

$$\nabla V_\theta = E_\theta \left[ \sum_{t=1}^T \pi_t(A_t|S_t, \theta) \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, A_t) \right]$$

and they refer to  $\beta_t$  as the baseline. In this subsection, we show that the baseline method is essentially a special case of our control variate approach to reduce the variance of gradient estimators; yet the baseline method suffers from two unnecessary limitations:

- (a) the baseline variate has a restricted form of dependency on the current action  $A_t$  compared to our control variate,

(b) the initial gradient estimator in [31] is

$$\frac{\partial V_\theta}{\partial \theta} = E_\theta \left[ \sum_{t=1}^T \pi_t(A_t|S_t, \theta) \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, A_t) \right],$$

which can be biased when the unknown Q-factors are not properly guessed.

Now, let us look more closely at these two points.

(a) For all baselines  $\beta_t$  we can write

$$\frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \beta_t(S_t) = I_t^a[Y_t]$$

with  $Y_t(s, a) = \frac{\partial \log \pi_t(a|s, \theta)}{\partial \theta} \beta_t(s)$ . Indeed,  $Y_t$  has zero conditional expectation given  $S_t$ , since we have for all  $s \in \mathcal{S}$  and for all  $t$ , by conditioning on  $A_t$ ,

$$E[Y_t(S_t, A_t)|S_t = s] = \sum_a \pi_t(a|s, \theta) \frac{\partial \log \pi_t(a|s, \theta)}{\partial \theta} \beta_t(s) = \beta_t(s) \cdot 0 = 0.$$

Consequently,  $I_t^a[Y_t] = E[Y_t|S_t, A_t] - E[Y_t|S_t] = Y_t = \frac{\partial \log \pi_t(a|s, \theta)}{\partial \theta} \beta_t$ , which is the control variate induced by baseline  $\beta_t$ . Thus, the baseline approach is a special case of our control variate approach with a dependency on the action that is proportional to  $\frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta}$ .

(b) The basic estimator of the value gradient used in [31] is

$$\mathbb{E}_n \left[ \sum_{t=1}^T \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, A_t) \right]. \quad (4.5.6)$$

Within our class of estimators,  $\left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) [W_1, \dots, W_T]$  with

$$W_t = \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \dots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) Q_\theta^t(S_t, A_t), \quad (4.5.7)$$

differs from the expression (4.5.6) only by a Bellman error term, since we have

$$\begin{aligned} \left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) [W_1, \dots, W_T] &= \mathbb{E}_n \left[ \sum_{t=1}^T \sum_{a_t} \frac{\partial \pi_t(a_t | S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, a_t) \right] \\ &+ \mathbb{E}_n \left[ \sum_{t=1}^T \frac{\partial \log \pi_t(A_t | S_t, \theta)}{\partial \theta} B_t \right]. \end{aligned}$$

When the unknown Q-factors  $Q_\theta^t$  are improperly guessed, the estimator (4.5.6) can be biased. In contrast, the Bellman error term makes sure that our estimator  $\left( \frac{\partial \hat{V}_\theta}{\partial \theta} \right) [W_1, \dots, W_T]$  is always unbiased. In the subsequent discussion, we will informally “equate” the naive estimator (4.5.6) with the estimator  $\frac{\partial \hat{V}_\theta}{\partial \theta} [W_1, \dots, W_T]$  where  $W_t$  is defined in Equation (4.5.7) so that we can focus on the difference of the optimal estimators between the baseline control variate and our more general variance reduction technique.

Now, we can compare the baseline method to our control variate approach. When the baseline is unrestricted, Theorem 8 in [31] asserts that the best baseline  $\beta_t$  to be used in the estimator of the value gradient

$$\mathbb{E}_n \left[ \sum_{t=1}^T \left( \frac{\partial \pi_t(A_t | S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, A_t) - \frac{\partial \log \pi_t(A_t | S_t, \theta)}{\partial \theta} \beta_t(S_t) \right) \right].$$

is given by

$$\beta_t^*(s) = \frac{E_\theta \left[ \left( \frac{\partial \log \pi_t(A_t | S_t, \theta)}{\partial \theta} \right)^2 Q_\theta^t(s, A_t) | S_t = s \right]}{E_\theta \left[ \left( \frac{\partial \log \pi_t(A_t | S_t, \theta)}{\partial \theta} \right)^2 | S_t = s \right]}.$$

In contrast, we know from Proposition 4.5.3 that the minimum variance unconstrained gradient estimator is

$$\frac{\partial \widehat{V}_\theta^*}{\partial \theta} = \frac{\partial \hat{V}_\theta}{\partial \theta} \left[ \frac{\partial X_1^*}{\partial \theta}, \dots, \frac{\partial X_T^*}{\partial \theta} \right],$$

where

$$\frac{\partial X_t^*}{\partial \theta} = -\mathbb{E}_n \left[ \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \dots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) Q_\theta^t(S_t, A_t) + \frac{\partial Q_\theta^t(S_t, A_t)}{\partial \theta} \right].$$

Equivalently, we have

$$\frac{\widehat{\partial V_\theta}^*}{\partial \theta} = \frac{\partial \hat{V}_\theta}{\partial \theta} \left[ W_1 + \frac{\partial Q_\theta^1(S_1, A_1)}{\partial \theta}, \dots, W_T + \frac{\partial Q_\theta^T(S_T, A_T)}{\partial \theta} \right].$$

When the term  $\frac{\partial Q_\theta^t(S_t, A_t)}{\partial \theta}$  cannot be written in the form of a baseline, that is, in the form  $\frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \beta_t(S_t)$  for some function  $\beta_t$ , the baseline approach has a higher variance compared to our method.

### 4.5.3 Comparison with actor-critic approaches

Actor-critic methods use a “value function” to generate a better estimate of the value gradient. This function is typically chosen so that the gradient estimate is unbiased and has low variance. Furthermore, it is often constrained to lie in a low-dimensional subspace when the state space is large.

Recall that the gradient of the true value  $V_\theta$  is

$$\frac{\partial V_\theta}{\partial \theta} = E_\theta \left[ \sum_{t=1}^T \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, A_t) \right],$$

where  $Q_\theta^t(s, a)$  is the Q-factor of policy  $\theta$  at time  $t$  in state-action pair  $(s, a)$ . If we use, instead of the Q-factors  $Q_\theta^t(s, a)$ , the functions  $(Q_\theta^t - Z_t)$  in the gradient expression, we obtain an estimator of the value gradient with a bias equal to

$$E_\theta \left[ \sum_{t=1}^T \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Z_t(S_t, A_t) \right].$$

This bias can be zero if the functions  $Z_t$  are chosen properly; for example if for

$t = 1, \dots, T$ , there holds

$$E_\theta \left[ \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} Z_t(S_t, A_t) \right] = 0.$$

Intuitively, actor-critic methods (e.g. [40]) require that the value function approximation is tailored to the policy parametrization so that there is no bias introduced in the gradient when using the functions  $(Q_\theta^t - Z_t)$ , instead of the true Q-factors  $Q_\theta^t$  in the gradient formula. However, since the generative MDP model is unknown, it is usually not possible to guarantee that approximated Q-factors do not bias the gradient.

By comparison, we suggest to use as a gradient estimator

$$\begin{aligned} & \mathbb{E}_n \left[ E_\theta \left[ \sum_{a_t} \pi_t(a_t|S_t, \theta) \frac{\partial \log \pi_t(a_t|S_t, \theta)}{\partial \theta} Q_\theta^t(S_t, a_t) | S_1 \right] \right] \\ & + \sum_{t=1}^T \mathbb{E}_n \left[ \left( \frac{\partial \log \pi_1(A_1|S_1, \theta)}{\partial \theta} + \dots + \frac{\partial \log \pi_t(A_t|S_t, \theta)}{\partial \theta} \right) B_t \right]. \end{aligned}$$

This estimator is unbiased even if the Q-factors are replaced with an erroneous guess, thanks to the Bellman error terms. Provided that the size  $n$  of the training set is not too large (which is the regime of interest in this chapter), the Bellman error term can be computed in practice.

## 4.6 Conclusion

### 4.6.1 Key findings

In this chapter, we characterized minimum variance estimators in a general class of unconstrained and constrained estimators of the policy value and policy value gradient. Our main findings are:

- Even though our optimal estimators are characterized in terms of the unknown Q-factors, wrong guesses for some unknown characteristics of the system still correspond to estimators in our class, which are unbiased, unlike many methods

in reinforcement learning whose unbiasedness relies on conditions that cannot be verified in practice. Our estimators bear resemblance with some classic approaches from reinforcement learning, but a key feature of our estimators that guarantees their unbiasedness is the presence of an empirical Bellman error term in the expressions (4.4.3) and (4.4.6). Since our work is precisely motivated by applications with a small number of observations, the computations of the Bellman error term is not time-consuming.

- Our value estimators with low variance hinge on the estimation of the advantages of the state-action pairs, not their Q-factors. In some applications, it may be possible to exploit this difference to our advantage.
- When there are few observed trajectories in the training set compared to the cardinality of the state space, it is better (in the sense that we obtain lower variance estimators) to estimate more accurately approximated (regularized) Q-factors rather to estimate loosely the full look-up table for the Q-factors, as we illustrated numerically in Subsection 4.4.4.

## 4.6.2 Discussion of the motivating examples

Now we comment briefly on how this chapter’s results could be applied to the two motivating examples described in the introduction.

### Catalog mailing problem

The MDP model for the catalog mailing problem that was mentioned in the introduction comprises approximately 500 states with two actions per state (namely mail a catalog or not). On the other hand, some catalog mailing companies observe the behavior of more than 100,000 customers over time. This suggests that the catalog mailing problem is rich in observations compared to the size of the underlying MDP (provided that attribution bias does plague the study in [86]).

However, if the sampling policy explores too little, that is if the selection probability of an action  $a$  in some state  $s$  is small, it is possible that there are only a few

observed trajectories such that action  $a$  has been selected in state  $s$ , despite a large training set.

Besides, even though the numerical analysis of the bias and variance of value estimator of the catalog mailing problem in [44] suggests that the available data allows for reasonably accurate estimate of the value function, the use of empirical Q-factors might not yield an accurate estimate of the advantages, which are critical to obtain a good value estimate. The use of method focused on advantages [7, 32, 8] or the use of the constrained estimators  $\hat{V}_\theta^{c,0}$  or  $\hat{V}_\theta^{c,1}$  of Subsection 4.4.2 with an approximation architecture for the Q-factors that will capture well the advantages could potentially improve on the value estimates obtained by previous approaches to the catalog mailing problem.

### **STAR\*D clinical trial**

Unlike the catalog mailing application, multi-period randomized clinical studies tend to lack observations compared to the underlying MDP's dimensionality. Nonetheless, medical experts have some good insights about the features of patients' health status that are relevant for depression. Consequently, the constrained estimator approach of Subsection 4.4.2 is well suited to the analysis of the STAR\*D clinical trial.

In the STAR\*D trial, only 1,500 patients stayed in the study after the first phase. At each patient's visit there are tens of recorded variables (discrete or continuous). In addition, since the patient's history is presumably important, the state definition should encode the past history over the two or three periods. As a result, an MDP model might require a very large number of states, at least a few thousands. Notwithstanding, psychiatrists think that the effectiveness of a depression treatment could be well-approximated by only 20 features [50]. Hence, the algorithms "Q-learning control variate" and "TD(1) control variate", proposed in Subsection 4.4.3, would estimate only 40 weights (one for each of the 20 features and for each of the 2 periods) with 1,500 trajectories - a ratio that seems within reason to obtain meaningful estimates.

### 4.6.3 Concluding remarks

In situations where it is difficult or costly to collect data, as in many marketing surveys or clinical trials, using estimators with minimum training set variance, instead of naive estimators, can make the difference between a meaningful and an irrelevant estimate. We expect the results of this chapter to be most relevant in these situations.

Finally, this chapter did not focus on the search of good policies using a training set of observed trajectories. Nonetheless, the potential benefits of using the better estimators exposed in this chapter in policy search is a promising direction of investigation to estimate policies with high true value. More precisely, we could estimate a policy given a training set by solving

$$\sup_{\theta \in \Theta} \hat{V}_{\theta}^*.$$

If  $\hat{\theta}$  is a policy maximizing  $\hat{V}_{\hat{\theta}}^*$ , it should be a good candidate to have a high true value  $V_{\hat{\theta}}$ , although this chapter did not investigate this question.



# Chapter 5

## Concluding remarks

In this thesis, we investigate three important questions about the control of MDPs when the underlying MDP model is unknown or uncertain.

In the first part of this thesis, we draw a comprehensive picture of the computational complexity of different formulations for the control of uncertain MDPs. Specifically, we account for the model uncertainty with three approaches: the expected utility, the worst-case model, and the maximum regret approach; in addition we consider different objective functions and types of uncertainty. We show that most formulations are plagued by the curse of uncertainty: out of the sixty analyzed problems, forty-four are at least NP-hard. Considering our complexity assessment, the worst-case model formulation with state-rectangular uncertainty seems attractive for its computational “tractability” and its hard performance guarantees.

In the second part, we motivate and define the notion of Markovian dynamically consistent convex risk measure, and we show that finding a risk-minimizing policy is equivalent to solving a zero-sum Markov game. Thus, our notion of Markovian risk allows to deal efficiently with sequential decision problems with large or even infinite time horizon. Our perspective not only guarantees that the robust control of MDPs proposed in the literature is sound from a decision-theoretic perspective, but it also suggests to mitigate the conservativeness of the worst-case robust control of uncertain MDPs by adding a penalty to “unlikely” parameters in a principled fashion. An interesting direction for further investigation is to suggest penalty functions that

are both meaningful to decision makers and well-motivated statistically.

Finally, we take a different perspective in the last chapter. Instead of separating model estimation and decision making, we exploit system's observations to estimate directly the value (resp. the value gradient) of policies on an unknown MDP. We define a broad class of unbiased estimators of the value (resp. the value gradient) and identify an estimator in this class with the lowest training set variance. Such an optimal estimator is characterized in terms of unknown characteristics of the system, which we need to estimate in order to use our theoretical characterization. We discuss different approaches to do so and we compare numerically the resulting algorithms with usual approaches from the literature. In the numerous applications where there are a few observations compared to the MDP size, our constrained estimators outperform the other algorithms by reducing substantially the estimation variance with regularized estimates of Q-factors, while being unbiased.

A major research effort on MDPs is to cope with the curse of dimensionality, which plagues many engineering applications. This thesis suggests another research direction focused on the curse of uncertainty, which is even more debilitating than the curse of dimensionality as we showed in Chapter 2. We believe that MDPs could have more impact, notably in social and medical sciences, if there were data-driven approaches to MDP control that would exploit efficiently the available information (typically in the form of observed system trajectories), rather than rely exclusively on models. An interesting prospect for future research is to investigate practical ways to exploit both structural and observational information in order to recommend better policies for practical applications.

# Bibliography

- [1] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis*. Springer-Verlag, 1994.
- [2] E. Altman. Constrained Markov Decision Processes. Technical Report 2574, INRIA, May 1995.
- [3] Ph. Artzner, F. Delbaen, J-M. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):2003–228, July 1999.
- [4] Ph. Artzner, F. Delbaen, J.M Eber, D. Heath, and H. Ku. Coherent multiperiod risk adjusted values and Bellman’s principle. Technical report, Université Louis Pasteur, 2003.
- [5] K.J. Aström. Theory and applications of adaptive control - A survey. *Automatica*, 19:471–486, 1983.
- [6] J.P. Aubin and I. Ekeland. *Applied nonlinear analysis*. Pure and Applied Mathematics. John Wiley and Sons, 1984.
- [7] L.C. Baird. Advantage learning. Technical report, US Air Force Academy, 1995.
- [8] L.C. Baird. Residual algorithms: reinforcement learning with function approximation. In A. Prieditis and S. Russel, editors, *Machine learning: proceedings of the twelfth international conference*, San Francisco, CA, 1995. Morgan Kaufman.

- [9] A.G. Barto, R.S. Sutton, and C.W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:834–846, 1983.
- [10] J.R. Beck and G.P. Pauker. The Markov process in medical prognosis. *Medical Decision Making*, 3(4), 1983.
- [11] D.P. Bertsekas. *Dynamic control and optimal control*, volume 1. Athena Scientific, second edition, 2001.
- [12] D.P. Bertsekas. *Dynamic control and optimal control*, volume 2. Athena Scientific, second edition, 2001.
- [13] D.P. Bertsekas and J.N. Tsitsiklis. An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595, August 1991.
- [14] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.
- [15] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming*, 98:49–71, 2003.
- [16] P. Billingsley. *Probability and measure*. Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 3rd edition, 1995.
- [17] D.J. Bruinvels, J. Kievit, and R.J. Eijkemans. Using age-specific rates in a Markov model. *Medical Decision Making*, 2:169, 1993.
- [18] 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, February 1999.
- [19] D.R. Cox. Causality: some statistical aspects. *Journal of the Royal Statistical Society - Series A*, 155:291–301, 1992.

- [20] D.P. de Farias and B. Van Roy. A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *submitted for publication*, 2004.
- [21] F. Delbaen. Coherent risk measures on general probability spaces. Technical report, ETH Zürich, 2000.
- [22] K. Detlefsen and G. Scandolo. Conditional and dynamic convex risk measures. Technical report, Humbolt Universität, 2005.
- [23] D. Ellsberg. Risk, ambiguity, and the Savage axioms. *Quarterly Journal of Economics*, 75:643–669, 1961.
- [24] L.G. Epstein and M. Schneider. Recursive multiple priors. *Journal of Economic Theory*, 113:1–31, 2003.
- [25] H. Föllmer and A. Schied. Convex measures of risk and trading constraints. *Finance and Stochastics*, 6:429–447, 2002.
- [26] M. Garey and D. Johnson. *Computers and intractability: A guide to the theory of NP-completeness*. W.H. Freeman and company, 2002.
- [27] L. El Ghaoui and H. Lebret. Robust solutions to least-square problems with uncertain data matrices. *SIAM Journal Matrix Analysis and Applications*, 18:1035–1064, 1997.
- [28] L. El Ghaoui, F. Oustry, and H. Lebret. Robust solutions to uncertain semi-definite programs. *SIAM Journal of Optimization*, 9:33–52, 1998.
- [29] P.W. Glynn. Likelihood ratio gradient estimation: an overview. In *Proceedings of the 19th conference on Winter simulation*, pages 366–375, 1987.
- [30] P.W. Glynn and D.L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

- [31] E. Greensmith, P.L. Bartlett, and J. Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.
- [32] M.E. Harmon, L.C. Baird, and A.H. Klopf. Advantage updating applied to a differential game. Neural Information Processing Conference, 1994.
- [33] S. G. Henderson and S. P. Meyn. Variance reduction for simulation in multiclass queueing networks. *IIE Transactions on Operations Engineering*, To appear, 2005.
- [34] D. Hernández-Hernández and S.I. Marcus. Risk-sensitive control of Markov processes in countable state space. *Systems and Control Letters*, 29(3):147–155, November 1996.
- [35] B. E. Hillner, T.J. Smith, and C.E. Desch. Efficacy and cost-effectiveness of autologous bone marrow transplantation in metastatic breast cancer. Estimates using decision analysis while awaiting clinical trial results. *Journal of American Medical Association*, 267(15), April 1992.
- [36] R.A. Howard and J.E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.
- [37] P.J. Huber. *Robust Statistics*. Wiley, 1981.
- [38] G. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30, 2005.
- [39] C. Kollman, K. Baggerly, D. Cox, and R. Picard. Adaptive importance sampling on discrete Markov chains. *Annals of Applied Probability*, 9(2):391–412, 1999.
- [40] V.R. Konda and J.N. Tsitsiklis. On actor-critic algorithms. *SIAM Journal of Control and Optimization*, 42(4):1143–1166, 2003.

- [41] S.S. Lavenberg and P. D. Welch. A perspective on the use of control variables to increase the efficiency of Monte Carlo simulations. *Management Science*, 27:322–335, 1981.
- [42] P. L’Ecuyer. A unified view of the IPA, SF, and LR gradient estimation techniques. *Management Science*, 36(11):1364–1383, November 1990.
- [43] W.W. Loh. *On the method of control variates*. PhD thesis, Stanford, 1994.
- [44] S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 2005.
- [45] P. Marbach and J.N. Tsitsiklis. Approximate gradient methods in policy-space optimization of Markov reward processes. *Journal of Discrete Event Dynamical Systems*, 13:111–148, 2003.
- [46] S.I. Marcus, E. Fernandez-Gaucherand, D. Hernandez-Hernandez, S. Coraluppi, and P. Fard. Risk sensitive Markov decision processes. *Progress in Systems and Control Theory*, 22:263–280, 1997.
- [47] M. Mundhenk. The complexity of optimal small policies. *Mathematics of Operations Research*, 25(1):118–129, February 2000.
- [48] S.A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society, Series B*, 65(2):331–366, 2003.
- [49] S.A. Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005.
- [50] S.A. Murphy and L. Gunter. Personal communication, November 2006.
- [51] S.A. Murphy, M.J. van der Laan, J. Robins, and CPPRG. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96:1410–1423, 2001.
- [52] B.L. Nelson. Control variate remedies. *Operations Research*, 38:974–992, 1990.

- [53] J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university, 1953.
- [54] A. Nilim and L. El Ghaoui. Robust solutions to Markov decision problems with uncertain transition matrices. *Operations Research*, 53(5), 2005.
- [55] National Institute of Mental Health. Questions and answers about the NIMH Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) study Background. [http://www.nimh.nih.gov/healthinformation/stard\\_qa\\_general.cfm](http://www.nimh.nih.gov/healthinformation/stard_qa_general.cfm), January 2006.
- [56] National Institute of Mental Health. Questions and answers about the NIMH Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) study Level 1 results. *American Journal of Psychiatry*, [http://www.nimh.nih.gov/healthinformation/stard\\_qa\\_level1.cfm](http://www.nimh.nih.gov/healthinformation/stard_qa_level1.cfm), January 2006.
- [57] National Institute of Mental Health. Questions and answers about the NIMH Sequenced Treatment Alternatives to Relieve Depression (STAR\*D) study Level 2 results. [http://www.nimh.nih.gov/healthinformation/stard\\_qa\\_level2.cfm](http://www.nimh.nih.gov/healthinformation/stard_qa_level2.cfm), March 2006.
- [58] C.H. Papadimitriou and J.N. Tsitsiklis. The complexity of Markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, August 1987.
- [59] I.Ch. Paschalidis and J.N. Tsitsiklis. Congestion-dependent pricing of network services. *IEEE/ACM Transactions on Networking*, 8(2):171–184, April 2000.
- [60] S.D. Patek. *Stochastic Shortest Path Games: Theory and Algorithms*. PhD thesis, Massachusetts Institute of Technology, 1997.
- [61] S.D. Patek. On terminating Markov decision processes with a risk-averse objective function. *Automatica*, 37:1379–1386, 2001.

- [62] S.B. Patten. Markov models of major depression for linking psychiatric epidemiology to clinical practice. *Clinical Practice and Epidemiology in Mental Health*, 1(2), 2005.
- [63] S.B. Patten. A major depression prognosis calculator based on episode duration. *Clinical Practice and Epidemiology in Mental Health*, 2(13), 2006.
- [64] S.B. Patten and R.C. Lee. Refining estimates of major depression incidence and episode duration in Canada using a Monte Carlo Markov model. *Medical Decision Making*, 24:351–358, July-August 2004.
- [65] L. Peshkin and S. Mukherjee. Bounds on sample size for policy evaluation in Markov environments. In D. Helmbold and B. Williamson (Eds.), editors, *COLT/EuroCOLT 2001*, 2001.
- [66] E.L. Porteus. *Foundations of Stochastic Inventory Management*. Stanford University Press, 2002.
- [67] F. Riedel. Dynamic coherent risk measures. *Stochastic Processes and their Applications*, 112(2), 2004.
- [68] J. Robins. Optimal structural nested models fo optimal sequential decisions. In D.Y. Lin and P. Heagerty, editors, *Proceedings of the Second Seattle Symposium on Biostatistics*. Springer, 2004.
- [69] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor. *Computer and Mathematics with Applications*, 14:1393- 1512, 1986.
- [70] J.M. Robins. *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics*, chapter Causal Inference from Complex Longitudinal Data, pages 69–117. Springer Verlag, 1997.
- [71] T. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 26:1443–1471, 2002.

- [72] B. Roorda, H. Schumacher, and J. Engwerda. Coherent acceptability measures in multiperiod models. *Mathematical Finance*, 2004.
- [73] B. Van Roy. *Learning and value function approximation in complex decision processes*. PhD thesis, M.I.T., 1998.
- [74] T. Runolfsson. Risk-sensitive control of Markov chains and differential games. *Proceedings of the 32nd IEEE Conference on Decision and Control*, 4:3377–3378, 1993.
- [75] A.J. Rush, M.H. Trivedi, S.R. Wisniewski, J.W. Stewart, A.A. Nierenberg, M.E. Thase, L. Ritz, M.M. Biggs, D. Warden, J.F. Luther, K. Shores-Wilson, G. Niederehe, and M. Fava. STAR\*D Study Team. Bupropion-SR, sertraline, or venlafaxine-XR after failure of SSRIs for depression. *New England Journal of Medicine*, 354(12):1231–1242, March 2006.
- [76] A. Ruszczyński and A. Shapiro. Conditional risk mappings. Technical Report 0404002, Economics Working Paper Archive EconWPA, 2004. Available at <http://ideas.repec.org/p/wpa/wuwpri/0404002.html>.
- [77] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. Technical Report 0404001, Economics Working Paper Archive EconWPA, 2004. Available at <http://ideas.repec.org/p/wpa/wuwpri/0404001.html>.
- [78] J. Satia and R.E. Lave. Markov decision processes with uncertain transition probabilities. *Operations Research*, 21(3):728–740, May-June 1973.
- [79] L.J. Savage. *Foundations of statistics*. John Wiley and sons, New York, 1954.
- [80] A.J. Schaefer, M. Bailey, S. Shechter, and M. Roberts. *Handbook of Operations Research/Management Science Applications in Health Care*, chapter Medical decisions using Markov Decision Processes, pages 597–616. Kluwer Academic Publishers.

- [81] P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352, 1994.
- [82] L.S Shapley. Stochastic games. In *Proceedings of the National Academy of Science*, volume 39, pages 1095–1100, 1953.
- [83] D. I. Simester, P. Sun, and J. N. Tsitsiklis. Dynamic catalog mailing policies. *Management Science*, to appear.
- [84] F.A. Sonnenberg and J.R. Beck. Markov models in medical decision making. *Medical Decision Making*, 13:322–339, 1993.
- [85] F.A. Sonnenberg and J.B. Wong. Commentary: fine-tuning life-expectancy calculations using Markov processes. *Medical Decision Making*, 2:170–172, 1993.
- [86] P. Sun. *Constructing learning models from data: the catalog mailing problem*. PhD thesis, Massachusetts Institute of Technology, 2003.
- [87] R.S. Sutton. Learning to predict by the methods of temporal difference. *Machine Learning*, 3:9–44, 1988.
- [88] RS. Sutton and AG. Barto. *Reinforcement Learning*. MIT Press, 1998.
- [89] R.S. Sutton, D.A. McAllester, S.P. Singh, and Y. Mansour. Policy gradient methods for Reinforcement Learning with function approximation. In *NIPS*, pages 1057–1063, 1999.
- [90] A. Ben Tal, L. El Ghaoui, and A. Nemirovski. *Semidefinite programming and applications*. Kluwer Academic Publishers, 2000.
- [91] A. Ben Tal and A. Nemirovski. Robust truss topology design via semidefinite programming. *SIAM Journal of Optimization*, 7:991–1016, 1997.
- [92] A. Ben Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23:769–805, 1998.

- [93] A. Ben Tal and A. Nemirovski. Robust solutions of uncertain quadratic and conic-quadratic problems. *SIAM Journal of Optimization*, to appear.
- [94] T.J. Teisberg. A dynamic programming model of the US strategic petroleum reserve. *Bell Journal of Economics*, 12(2):526–546, 1981.
- [95] M.H. Trivedi, M. Fava, S.R. Wisniewski, M.E. Thase, F. Quitkin, D. Warden, L. Ritz, A.A. Nierenberg, B.D Lebowitz, M.M. Biggs, J.F. Luther, K. Shores-Wilson, and A.J. Rush. STAR\*D Study Team. Medication augmentation after the failure of SSRIs for depression. *New England Journal of Medicine*, 354(12):1243–1252, March 2006.
- [96] M.H. Trivedi, A.J. Rush, S.R. Wisniewski, A.A. Nierenberg, D. Warden, L. Ritz, G. Norquist, R.H. Howland, B.D Lebowitz, P.J. McGrath, K. Shores-Wilson, M.M. Biggs, G.K. Balasubramani, and M. Fava. Evaluation of outcomes with Citalopram for depression using measurement-based care in STAR\*D: Implications for Clinical Practice. *American Journal of Psychiatry*, 163(1):28–40, January 2006.
- [97] P. Tseng. Solving H-horizon, stationary Markov decision problems in time proportional to  $\log(H)$ . *Operations Research Letters*, 9(5):287–297, September 1990.
- [98] J.N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex American-style options. *IEEE Transactions on Neural Networks*, 12(4):694–703, July 2001.
- [99] T. Wang. Conditional preferences and updating. *Journal of Economic Theory*, 108, 2003.
- [100] C.J.C.H. Watkins. *Learning from delayed rewards*. PhD thesis, University of Cambridge, 1989.
- [101] C.J.C.H. Watkins and P. Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.

- [102] D. Williams. *Probability with Martingales*. Cambridge university press, 8 edition, 2004.
- [103] J.L. Williams, J.W. Fisher III, and A.S. Willsky. Importance sampling actor-critic algorithms. In *American Control Conference*, June 2006.
- [104] R.J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Journal Machine Learning*, 8(3-4):229–256, May 1992.
- [105] J.B. Wong, W.G. Bennett, R.S. Koff, S.G. Pauker, E. Hillner, T.J. Smith, and C.E. Desch. Pretreatment evaluation of chronic hepatitis C: Risks, benefits, and costs. *Journal of American Medical Association*, 280(24), December 1998.