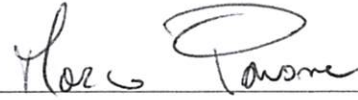


RISK-SENSITIVE AND DATA-DRIVEN SEQUENTIAL DECISION MAKING

A DISSERTATION  
SUBMITTED TO THE DEPARTMENT OF INSTITUTE OF COMPUTATIONAL &  
MATHEMATICAL ENGINEERING  
AND THE COMMITTEE ON GRADUATE STUDIES  
OF STANFORD UNIVERSITY  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF  
DOCTOR OF PHILOSOPHY

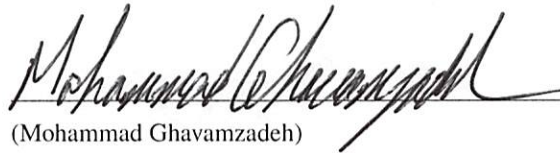
Yinlam Chow  
March 2017

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



(Marco Pavone) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



(Mohammad Ghavamzadeh)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.



(Ramesh Johari)

Approved for the Stanford University Committee on Graduate Studies

# Abstract

Markov decision processes (MDPs) provide a mathematical framework for modeling sequential decision making where system evolution and cost/reward depend on uncertainties and control actions of a decision. MDP models have been widely adopted in numerous domains such as robotics, control systems, finance, economics, and manufacturing. At the same time, optimization theories of MDPs serve as the theoretical underpinnings to numerous dynamic programming and reinforcement learning algorithms in stochastic control problems. While the study in MDPs is attractive for several reasons, there are two main challenges associated with its practicality:

- An accurate MDP model is oftentimes not available to the decision maker. Affected by modeling errors, the resultant MDP solution policy is non-robust to system fluctuations.
- The most widely-adopted optimization criterion for MDPs is represented by the risk-neutral expectation of a cumulative cost. This does not take into account the notion of risk, i.e., increased awareness of events of small probability but high consequences.

In this thesis we study multiple important aspects in risk-sensitive sequential decision making where the variability of stochastic costs and robustness to modeling errors are taken into account. First, we address a special type of risk-sensitive decision making problems where the percentile behaviors are considered. Here risk is either modeled by the conditional value-at-risk (CVaR) or the Value-at-risk (VaR). VaR measures risk as the maximum cost that might be incurred with respect to a given confidence level, and is appealing due to its intuitive meaning and its connection to chance-constraints. The VaR risk measure has many fundamental engineering applications such as motion planning, where a safety constraint is imposed to upper bound the probability of maneuvering into dangerous regimes. Despite its popularity, VaR suffers from being unstable, and its singularity often introduces mathematical issues to optimization problems. To alleviate this problem, an alternative measure that addresses most of VaR's shortcomings is CVaR. CVaR is a risk-measure that is rapidly gaining popularity in various financial applications, due to its favorable computational properties (i.e., CVaR is a coherent risk) and superior ability to safeguard a decision maker from the “outcomes that hurt the most”. As a risk that measures the conditional expected cost given that such cost is greater than or equal to VaR, CVaR accounts for the total cost of undesirable events (it corresponds to events whose associated probability is low, but the corresponding cost is high) and is therefore preferable in financial applications

such as portfolio optimization.

Second, we consider optimization problems in which the objective function involves a coherent risk measure of the random cost. Here the term coherent risk [7] denotes a general class of risks that satisfies convexity, monotonicity, translational-invariance and positive homogeneity. These properties not only guarantee that the optimization problems are mathematically well-posed, but they are also axiomatically justified. Therefore modeling risk-aversion with coherent risks has already gained widespread acceptance in engineering, finance and operations research applications, among others. On the other hand, when the optimization problem is sequential, another important property of a risk measure is time consistency. A time consistent risk metric satisfies the “dynamic-programming” style property which ensures rational decision making, i.e., the strategy that is risk-optimal at the current stage will also be deemed optimal in subsequent stages. To get the best of both worlds, the recently proposed Markov risk measures [119] satisfy both the coherent risk properties and time consistency. Thus to ensure rationality in risk modeling and algorithmic tractability, this thesis will focus on risk-sensitive sequential decision making problems modeled by Markov risk measures.

# Acknowledgements

First and foremost, I would like to thank my advisor, Prof. Marco Pavone. Marco has been an excellent teacher, an influential mentor and a dependable friend. His scientific acumen, commitment to perfection, charisma in presentation, and intellectual integrity have a continuing impact on my professional career and personal development.

I would like to sincerely thank Dr. Mohammad Ghavamzadeh at Adobe Research and Prof. Shie Mannor at Technion for their collaborations on several topics in this thesis. Throughout my PhD curriculum, Shie's profound knowledge, astuteness, and openness to innovative ideas have shaped my research paths. Mohammad's rich knowledge and interests have broadened my horizons. Moreover, he has always been extremely supportive and has spent much-appreciable effort in advising my career development. I am deeply grateful to have both of you to be my personal mentors and friends. I am looking forward to work together and stay in touch in the years to come.

I would like to express my appreciation to my defense committee, who provided a lot of thoughtful comments for improving my work. I would like to thank Prof. Ramesh Johari, who worked with me on the motivation of risk-sensitive decision using knowledge from game theory. Last but not least, I would like to thank my committee chair Prof. Benjamin Van Roy, who offered additional insights to my research from the operations research perspective. Your work was among the fundamental ones that initiated my interests in the area of reinforcement learning.

I am thankful to my co-authors for all the collaborations: Stefano Carpin, Mohammad Ghavamzadeh, Lucas Janson, Summet Katariya, Anirudha Majumdar, Alan Malek, Shie Mannor, Marek Petrik, Junjie Qin, Sumeet Singh, Aviv Tamar, Jiyan Yang, Jiayuan Yu. It has been grateful working with all of you on various research topics ranging from stochastic optimal control theories, reinforcement learning and sequential hypothesis testing, to applications in energy systems and robotic platforms. You have all shared with me brilliant insights and creative ideas. The valuable lessons learned from these experiences sharpen my skill sets to conduct independent, high-quality research in the future.

To my friends in the ASL and in Stanford, including: Joseph Starek, Ashley Clark, Ed Schmerling, Sumeet Singh, Ross Allen, Zach Sunberg, Brian Ichter, Federico Rossi, Ben Hockman, Rick Zhang, Jiyan Yang, Junjie Qin, Tomas Tinoco De Rubira and Rob Wang – thank you for making my 5 years of Stanford experience memorable. To my lab mates, it has been enjoyable working together since the early stage of

ASL. Thank you for your support during all the ups and downs. Your creativity and striving for excellence have always inspired me.

I would like to extend my gratitude to Croucher foundation for their generous financial support on my PhD studies and research.

I am grateful to my parents, especially my mother Catherine, for their unconditional love. It has been difficult for not being able to stay in touch at all times. On top of that, I would thank my beloved Suidan for her support throughout these fulfilling, yet stressful years. Thanks to her unwavering encouragement, which allows me to continue the pursuit of passion even during the downtimes. I owe everything to their dedication and sacrifice.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>Acknowledgements</b>	<b>vi</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Problem Description . . . . .	2
1.2 Markov Decision Processes (MDPs) . . . . .	3
1.3 Overview of Risk Measures . . . . .	4
1.3.1 Sources of Uncertainty in MDPs . . . . .	4
1.3.2 Entropic Risk and its Limitations . . . . .	5
1.3.3 Percentile Risk Metrics . . . . .	7
1.3.4 Static, Coherent Measures of Risk . . . . .	8
1.3.5 Dynamic, Time-Consistent Measures of Risk . . . . .	9
1.4 Existing Solution Approaches and Limitations . . . . .	12
1.4.1 Time Inconsistency in Risk-aware Planning . . . . .	12
1.4.2 Limitation 2: Complexity of solution approaches . . . . .	14
1.5 Risk-sensitive Decision Making Versus Reward Shaping . . . . .	15
1.6 Thesis Contributions and Outline . . . . .	16
<b>2 Risk-Sensitive Decision Making: A CVaR Optimization Approach</b>	<b>18</b>
2.1 Introduction . . . . .	18
2.1.1 Risk Sensitive Decision Making with CVaR . . . . .	18
2.1.2 Chapter Contribution . . . . .	19
2.1.3 Chapter Organization . . . . .	19
2.2 Problem Formulation and Motivation . . . . .	20
2.2.1 Problem Formulation . . . . .	20
2.2.2 Motivation - Robustness to Modeling Errors . . . . .	20
2.3 Bellman Equation for CVaR . . . . .	22
2.4 Value Iteration with Linear Interpolation . . . . .	24

2.5	CVaR $Q$ -learning with Linear Interpolation . . . . .	27
2.5.1	Synchronous CVaR $Q$ -learning . . . . .	28
2.5.2	Asynchronous CVaR $Q$ -learning . . . . .	29
2.6	Extension to Mean-CVaR MDP . . . . .	30
2.6.1	Bellman Equation . . . . .	31
2.7	Experiments . . . . .	33
2.8	Conclusion . . . . .	34
<b>3</b>	<b>Risk-Constrained Reinforcement Learning with Percentile Risk</b>	<b>35</b>
3.1	Introduction . . . . .	35
3.1.1	Risk Sensitive Reinforcement Learning . . . . .	35
3.1.2	Chapter Contribution . . . . .	36
3.1.3	Chapter Organization . . . . .	37
3.2	Preliminaries . . . . .	37
3.2.1	Notations . . . . .	37
3.2.2	Problem Statement . . . . .	39
3.2.3	Lagrangian Approach and Reformulation . . . . .	41
3.3	A Trajectory-based Policy Gradient Algorithm . . . . .	41
3.4	Actor-Critic Algorithms . . . . .	44
3.4.1	Gradient w.r.t. the Policy Parameters $\theta$ . . . . .	45
3.4.2	Gradient w.r.t. the Lagrangian Parameter $\lambda$ . . . . .	48
3.4.3	Sub-Gradient w.r.t. the VaR Parameter $\nu$ . . . . .	49
3.4.4	Convergence of Actor-Critic Methods . . . . .	50
3.5	Extension to Chance-Constrained Optimization of MDPs . . . . .	50
3.5.1	Policy Gradient Method . . . . .	51
3.5.2	Actor-Critic Method . . . . .	51
3.6	Experiments . . . . .	53
3.6.1	The Optimal Stopping Problem . . . . .	54
3.6.2	A Personalized Ad-Recommendation System . . . . .	56
3.7	Conclusion . . . . .	58
<b>4</b>	<b>Risk Sensitive Model Predictive Control</b>	<b>60</b>
4.1	Introduction . . . . .	60
4.1.1	Model Predictive Control . . . . .	60
4.1.2	MPC with Time Consistent Risk Measures . . . . .	60
4.1.3	Chapter Contribution . . . . .	61
4.1.4	Chapter Organization . . . . .	61
4.2	Model Description . . . . .	61



4.3	Markov Polytopic Risk Measures . . . . .	62
4.3.1	Polytopic Risk Measures . . . . .	62
4.3.2	Markov Dynamic Polytopic Risk Metrics . . . . .	64
4.3.3	Computational Aspects of Markov Dynamic Polytopic Risk Metrics . . . . .	65
4.4	Problem Formulation . . . . .	65
4.5	Risk-Sensitive Stability . . . . .	68
4.6	Model Predictive Control Problem . . . . .	68
4.6.1	The Unconstrained Case . . . . .	68
4.6.2	The Constrained Case . . . . .	69
4.7	Bounds on Optimal Cost . . . . .	72
4.7.1	Lower Bound . . . . .	72
4.7.2	Upper Bound . . . . .	73
4.8	Solution Algorithms . . . . .	73
4.8.1	Dynamic Programming Approach . . . . .	73
4.8.2	Convex Programming Approach . . . . .	75
4.9	Numerical Experiments . . . . .	76
4.9.1	Effects due to Risk Aversion . . . . .	77
4.9.2	A 2-state, 2-input Stochastic System . . . . .	78
4.9.3	Comparison with Bernadini and Bemporad's Algorithm [15] . . . . .	79
4.9.4	Safety Brake in Adaptive Cruise Control . . . . .	80
4.10	Conclusion . . . . .	81
<b>5</b>	<b>Stochastic Optimal Control with Dynamic Risk Constraints</b>	<b>83</b>
5.1	Introduction . . . . .	83
5.1.1	An Overview on Constrained Stochastic Optimal Control . . . . .	83
5.1.2	Chapter Contribution . . . . .	84
5.1.3	Chapter Organization . . . . .	84
5.2	Problem Formulation . . . . .	85
5.3	A Dynamic Programming Algorithm for Risk-Constrained Multi-Stage Decision-Making . .	86
5.3.1	Dynamic Programming Recursion . . . . .	86
5.3.2	Construction of optimal policies . . . . .	88
5.4	Discretization/Interpolation Algorithms for AMDP . . . . .	88
5.4.1	Discretization Algorithm . . . . .	88
5.4.2	Interpolation Algorithm . . . . .	90
5.5	Experiments . . . . .	92
5.5.1	Curse of Dimensionality with Discretization Approach . . . . .	92
5.6	Conclusion . . . . .	94

<b>6</b>	<b>Conclusion</b>	<b>95</b>
6.1	Inherent Uncertainty Versus Model Uncertainty . . . . .	96
6.2	Time Consistency in Risk-aware Planning . . . . .	97
6.3	Risk-shaping . . . . .	97
6.4	Future Work . . . . .	98
6.4.1	Exploration Versus Exploitation in Risk-sensitive Reinforcement Learning . . . . .	98
6.4.2	Risk Sensitive Importance Sampling . . . . .	98
6.4.3	Relationship to Safe Policy Improvement . . . . .	99
<b>7</b>	<b>Supplementary Materials</b>	<b>100</b>
7.1	Technical Results in Chapter 2 . . . . .	100
7.1.1	Proof of Proposition 2.2.1 . . . . .	100
7.1.2	Proof of Lemma 2.3.2 . . . . .	101
7.1.3	Proof of Theorem 2.3.3 . . . . .	103
7.1.4	Proof of Theorem 2.3.4 . . . . .	106
7.1.5	Proof of Lemma 2.4.3 . . . . .	107
7.1.6	Useful Intermediate Results . . . . .	110
7.1.7	Proof of Theorem 2.4.4 . . . . .	112
7.1.8	Proof of Theorem 2.5.1 . . . . .	116
7.1.9	Proof of Theorem 2.5.2 . . . . .	117
7.1.10	Proof of Theorem 2.6.1 . . . . .	119
7.2	Technical Results in Chapter 3: Policy Gradient Methods . . . . .	121
7.2.1	Computing the Gradients . . . . .	121
7.2.2	Proof of Convergence of the Policy Gradient Algorithm . . . . .	122
7.3	Technical Results in Chapter 3: Actor-Critic Algorithms . . . . .	136
7.3.1	Gradient with Respect to $\lambda$ (Proof of Lemma 3.4.4) . . . . .	136
7.3.2	Proof of Convergence of the Actor-Critic Algorithms . . . . .	137
7.4	Technical Results in Chapter 4 . . . . .	146
7.4.1	Proof of Lemma 4.5.1 . . . . .	146
7.4.2	Proof of Theorem 4.6.1 . . . . .	146
7.4.3	Proof of Lemma 4.6.2 . . . . .	148
7.4.4	Proof of Theorem 4.6.4 . . . . .	149
7.4.5	Proof of Theorem 4.6.6 . . . . .	150
7.4.6	Proof of Theorem 4.7.1 . . . . .	151
7.4.7	Proof of Theorem 4.7.2 . . . . .	152
7.4.8	Proof of Corollary 4.8.4 . . . . .	153
7.4.9	Proof of Theorem 4.8.1 and Corollary 4.8.2 . . . . .	153
7.4.10	Convex Programming Formulation of Problem $\mathcal{MPC}$ . . . . .	156

7.4.11	A Generalized Stability Condition . . . . .	158
7.4.12	Alternative Formulation of Problem $\mathcal{PE}$ and $\mathcal{MPC}$ . . . . .	161
7.4.13	Suboptimality Performance of $\pi^{\text{MPC}}$ . . . . .	162
7.5	Technical Results in Chapter 5 . . . . .	167
7.5.1	Proof of Theorem 5.3.2 . . . . .	167
7.5.2	Proof of Theorem 5.3.6 . . . . .	169
7.5.3	Proof of Lemma 5.4.3 . . . . .	170
7.5.4	Proof of Theorem 5.4.4 . . . . .	173
7.5.5	Proof of Theorem 5.4.7 . . . . .	177

<b>Bibliography</b>		<b>179</b>
---------------------	--	------------

# List of Tables

3.1	Performance comparison of the policies learned by the risk-constrained and risk-neutral algorithms. In this table $\sigma(\mathcal{C}^\theta(x^0))$ stands for the standard deviation of the total cost. . . . .	55
3.2	Performance comparison of the policies learned by the CVaR-constrained and risk-neutral algorithms. In this table $\sigma(\mathcal{R}^\theta(x^0))$ stands for the standard deviation of the total reward. . . . .	58
4.1	Statistics for Risk-Averse MPC. . . . .	78
4.2	Performance of Different Algorithms. . . . .	79
4.3	Statistics for Risk-Averse MPC. . . . .	80
4.4	Statistics for Risk-Sensitive ACC System (with Mean Absolute Semi-deviation Risk). . . . .	82
5.1	Computation Times with Different Discretization Step Sizes. . . . .	93

# List of Figures

1.1	Scenario Tree for Example 1.3.5. . . . .	10
1.2	Limitations of mean-variance optimization. Underlined numbers along the edges represent transition probabilities; non-underlined numbers represent stage-wise constraint and objective function costs (that are equal for this example). Terminal constraint costs are zero. Under policy $\pi_1$ , the costs per stage are given by $d(x_0, a_0) = 0.5 \cdot 0 + 0.5 \cdot 10 = 5$ , $d(x_1, a_1) = 10$ , and $d(x_2, a_1) = 10$ ; under policy $\pi_2$ , the costs per stage are given by $d(x_0, a_0) = 5$ , $d(x_1, a_1) = 20$ , and $d(x_2, a_1) = 10$ . One can verify that for policy $\pi_1$ one has $\text{var}\left(\sum_{k=0}^{N-1} d(x_k, a_k)\right) = 25$ , while for policy $\pi_2$ one has $\text{var}\left(\sum_{k=0}^{N-1} d(x_k, a_k)\right) = 0$ . Then, if the risk threshold is less than 25, the decision-maker would choose policy $\pi_2$ and would seek to incur losses in order to keep the variance small enough. . . . .	14
1.3	Limitations of chance-constrained optimization. The numbers along the edges represent transition probabilities, while the numbers below the terminal nodes represent the stage-wise constraint costs. The problem involves a single control policy (hence there is a unique transition graph). The constraint cost appears acceptable in states $x_1$ and $x_2$ , but unacceptable from the perspective of the first stage in state $x_0$ . . . . .	15
2.1	Grid-world simulation. First three plots show the value functions and corresponding paths for different CVaR confidence levels. The last plot shows a cost histogram (for 400 Monte Carlo trials) for a risk-neutral policy and a CVaR policy with confidence level $\alpha = 0.11$ . . .	34
3.1	Cost distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms. . . . .	56
3.2	Cost distributions for the policies learned by the chance-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms. . . . .	56
3.3	Reward distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms. . . . .	58

4.1	Effect of semi-deviation parameter $c$ . . . . .	77
5.1	Convergence of Approximated Value Functions using Different Discretization Step Sizes. . .	93

# Chapter 1

## Introduction

Decision-making is concerned with identifying the *optimal strategy* (a mapping from current system states to available actions) in which the performance is measured by an associated cost function. The cost function captures specific evaluation criteria that are deemed relevant to the decision makers. In general, decision-making is an interesting yet challenging problem. The challenges of decision-making are three-fold. First, the evaluation criteria usually contain multiple conflicting elements that make decision-making non-trivial. For example, in a Mars exploration mission the project manager (who serves as the decision maker) has to trade-off fuel-efficiency and mission safety during the design of Mars rover deployment strategies; or the business analyst of a manufacturing plant has to trade-off quality and cost in planning production schedules. Second, in most practical applications, decisions are often *temporally dependent*. Specifically, in sequential decision-making problems, the decision maker either interacts with the system *myopically* over multiple time periods, or the decision maker decides a strategy based on the *feedback* observations of the system. While the second approach is always preferable due to its full utilization of system information, it encounters a major computational difficulty. Unlike in the myopic optimization problem, where the optimizer is a point solution, this challenge arises from the fact that in sequential decision-making the problem is often cast as a *functional optimization* problem whose solution is a mapping from the history of states to actions. Third, in decision-making the system evolution and performance are often affected by *uncertain exogenous factors*, such as system noise or measurement errors. Notice that the number of possible strategies depends exponentially on the decision horizon and the realizations of uncertainties. Such vast amount of potential choices often makes the direct enumeration of solutions intractable.

The most widely-adopted optimization criterion for sequential decision-making is represented by the *risk-neutral* expectation of a cumulative cost. However despite its popularity, this criterion does not take *variability* of the cost and *sensitivity* to modeling errors into account. This may lead to potential modeling problems, where the downside risks incurred by outcome realizations are ignored. On the other hand, while risk-sensitive decision-making provides a promising approach to compute robust solution policies (with respect to cost variability and modeling errors), constructing a “good” risk criterion in a manner that is both

conceptually meaningful and computationally tractable still presents a nontrivial challenge to system designers.

## 1.1 Problem Description

In this thesis we investigate risk-aware planning and control in uncertain environments, namely, the problem of devising a provably-safe action strategy in the presence of inherent uncertainties and model uncertainties. Such a problem has recently been recognized as one of the main challenges in many areas such as robotic motion planning, personalized online marketing, portfolio optimization and intelligent transportation management. The issue of planning under uncertainty without the notion of risk-awareness, has been addressed extensively in the past; for example see [22] and references therein. In particular, in this planning problem where a stochastic sequential objective is involved, the solution often entails a decision-making strategy as opposed to an open-loop control sequence. Despite great strides in the theory of risk-modeling, the inclusion of risk-awareness in sequential planning has so far received limited attentions. Yet, the inclusion of risk awareness in stochastic optimal control is critical for several reasons. First, a guaranteed-feasible solution may not exist in stochastic planning problems, and the question becomes how to properly trade-off between planner's conservative-ness and the risk of infeasibility. Second, risk-awareness allows decision makers to increase policy robustness by including model uncertainties in the problem formulation. Third, by imposing various levels of risk to the inherent uncertainties presented in the environment, risk-aware planning can avoid rare undesirable events. Finally, in the reinforcement learning framework where the world model is not known accurately, a risk-aware planner can balance exploration versus exploitation for efficient policy learning and can guarantee safety by limiting the visiting frequency to states that lead to catastrophic failures.

Clearly, one would desire that sequential planning algorithms are able to take into the account of risks. Unfortunately most existing planning algorithms ignore risk-awareness and safety. From a technical standpoint, there are two main approaches for risk-sensitive decision-making: optimizing risk-sensitive objective functions, where a risk-neutral expectation operator is replaced by a risk function, or adding risk constraints to the optimization problem. The first approach is more suitable to portfolio optimization and online marketing problems for which the planning goals are to maximize expected revenue and to control variability. The second approach is more preferable to engineering problems such as motion planning due to their ability to enforce safety constraints.

The most common framework for planning under uncertainty is provided by Markov decision processes (MDPs), which represent a probabilistic sequential decision-making framework such that the set of transition probabilities to next states depend only on the current state and action of the system. This framework can be further generalized to reinforcement learning (RL), which combines the learning of transition probabilities and cost functions in MDPs with the computation of an optimal policy. The key research aspects that we will explore in this thesis are: risk models in the MDP framework that ensures rational decision-making, solution algorithms that are computationally efficient to solve real-world problems, and risk-constrained decision



making problems that optimize a risk-neutral objective function subjected to risk-sensitive constraints.

## 1.2 Markov Decision Processes (MDPs)

In this thesis, the underlying mathematical model of sequential decision making and reinforcement learning is the Markov decision process (MDP). An MDP is a tuple  $(\mathcal{X}, \mathcal{A}, C, P, \gamma, x_0)$ , where  $\mathcal{X}$  and  $\mathcal{A}$  are state and action spaces,  $C(x, a) \in [-C_{\max}, C_{\max}]$  is a bounded deterministic cost,  $P(\cdot|x, a)$  is the transition probability distribution,  $\gamma \in [0, 1]$  is the discounting factor<sup>1</sup>, and  $x_0$  is the initial state. Our results easily generalize to random initial states and random costs, but for simplicity we will focus on the case of deterministic initial state and immediate cost in this thesis. For each state  $x \in \mathcal{X}$ , we also denote by  $\mathcal{A}(x)$  the corresponding set of admissible control actions. In a more general setting when multi-stage constraints are taken into account, we define a constrained Markov decision process (CMDP) which extends the MDP model by introducing additional costs and associated constraints. A CMDP is defined by  $(\mathcal{X}, \mathcal{A}, C, D, P, \gamma, x_0, d_0)$  where the components  $\mathcal{X}, \mathcal{A}, C, P, x_0, \gamma$  are the same for the unconstrained MDP. Furthermore  $D(x, a) \in [-D_{\max}, D_{\max}]$  is a bounded deterministic constraint cost, and  $d_0 \in \mathbb{R}$  is an upper bound for the expected cumulative (through time)  $D$  cost. Intuitively, solving an MDP means determining a sequence of *policies*  $\pi$  (mappings from histories to control actions) which minimizes the risk-sensitive cumulated objective cost defined by  $C$ , while solving a CMDP means determining a sequence of policies  $\pi$  which minimizes the same objective function and at the same time ensures that the cumulated constraint cost defined by the functions  $D$  is (under specific risk metrics) bounded by  $d$ .

In order to formalize the optimization problems associated with MDPs or CMDPs, we define the feasible set of policies as follows. Let the space of admissible histories up to time  $t$  be  $h_t = H_{t-1} \times \mathcal{A} \times \mathcal{X}$ , for  $t \geq 1$ , and  $H_0 = \mathcal{X}$ . A generic element  $h_t \in h_t$  is of the form  $h_t = (x_0, a_0, \dots, x_{t-1}, a_{t-1}, x_t)$ . Let  $\Pi_{H,t}$  be the set of all history-dependent policies with the property that at each time  $t$  the policy is a function that maps  $h_t$  to the probability distribution over the action space  $\mathcal{A}$ . In other words,  $\Pi_{H,t} := \{\mu_0 : H_0 \rightarrow \mathbb{P}(\mathcal{A}), \mu_1 : H_1 \rightarrow \mathbb{P}(\mathcal{A}), \dots, \mu_t : H_t \rightarrow \mathbb{P}(\mathcal{A}) \mid \mu_j(\cdot|h_j) \in \mathbb{P}(\mathcal{A}) \text{ for all } h_j \in H_j, 1 \leq j \leq t\}$ . We also let  $\Pi_H = \lim_{t \rightarrow \infty} \Pi_{H,t}$  be the set of all history dependent policies.

While  $\Pi_H$  is the most generic class of policies in sequential decision making, oftentimes MDP or CMDP problems with history dependent policies are numerically intractable. Another commonly considered class of policies in literature is known as the class of Markovian policies  $\Pi_M$ , where at each time step  $t$  the policy is a function that maps states  $x_t$  to the probability distribution over the action space  $\mathcal{A}$ . Formally the class of Markovian policies is defined as  $\Pi_M = \lim_{t \rightarrow \infty} \Pi_{M,t}$  where  $\Pi_{M,t} := \{\mu_0 : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{A}), \mu_1 : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{A}), \dots, \mu_t : \mathcal{X} \rightarrow \mathbb{P}(\mathcal{A}) \mid \mu_j(\cdot|x_j) \in \mathbb{P}(\mathcal{A}) \text{ for all } x_j \in \mathcal{X}, 1 \leq j \leq t\}$ . In the special case when the policies are time-homogeneous, i.e.,  $\mu_j = \mu$  for all  $j \geq 0$ , then the class of policies is known as stationary Markovian and denoted by  $\Pi_{M,S}$ . When  $\pi$  is stationary and Markovian (i.e.,  $\pi \in \Pi_{M,S}$ ), it is merely a

<sup>1</sup> By introducing  $\gamma \in (0, 1)$  to the sum of multi-stage cost functions, we aim to solve the MDP problem with more focus on optimizing current costs over future costs. When  $\gamma = 1$ , the effect of discounting factor vanishes, and the corresponding MDP problem minimizes the total cost.

sequence of policies (denoted by  $\mu$ ). For notational convenience we use  $\mu$  and  $\pi$  interchangeably in this case. Compared to the structure of  $\Pi_H$ , the set of policies characterized by  $\Pi_{M,S}$  is more structured (i.e., the control actions only depend on current state information and its state-action mapping is time-independent). Computationally this makes the procedure of solving for an optimal policy under the class of stationary Markovian policies more tractable, and common solution techniques involve dynamic programming algorithms [17] such as Bellman iteration.

When the objective function for an MDP is given by the *risk-neutral* expectation of a cumulative cost, i.e.,

$$\min_{\pi \in \Pi_H} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t C(x_t, a_t) \mid x_0, a_t \sim \pi_t(\cdot | h_t) \right],$$

Bellman's principle of optimality [17] shows that the optimal policy lies in the class of stationary Markovian policies  $\Pi_{M,S}$ . On the other hand, for a CMDP whose objective function and constraints are modeled by the *risk-neutral* expectation of a cumulative cost and constraint cost, i.e.,

$$\begin{aligned} \min_{\pi \in \Pi_H} \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t C(x_t, a_t) \mid x_0, a_t \sim \pi_t(\cdot | h_t) \right], \\ \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t D(x_t, a_t) \mid x_0, a_t \sim \pi_t(\cdot | h_t) \right] \leq d_0, \end{aligned}$$

Altman (Theorem 3.8 in [4]) shows a similar result of optimality for which  $\Pi_{M,S}$  is called the “dominating class of policies”. While this nice property does not hold for arbitrary objective functions and constraints in a CMDP, we manage to show that by specifying an augmented state that keeps track of the risk evaluation in subsequent stages, the optimal policies of the corresponding CMDPs indeed belong to the class of stationary Markovian policies (with respect to the augmented states), for the risk-sensitive objective functions and constraints considered in this thesis (see Chapter 2, 3 and 5).

## 1.3 Overview of Risk Measures

### 1.3.1 Sources of Uncertainty in MDPs

Under the framework of MDPs, we hereby describe the two sources of uncertainty, i.e., inherent-uncertainty and model-uncertainty, incurred by the cumulated cost random variable. Inherent-uncertainty describes the uncertainty from stochastic transitions of a single, well-defined MDP. On the other hand, model-uncertainty characterizes the inaccuracy of transition probability and immediate cost of an MDP. In general, inherent-uncertainty accounts for the cost variability due to the stochasticity of an MDP, whereas model-uncertainty accounts for the errors in MDP representations.

The most widely-adopted optimization criterion for MDPs is represented by the *risk-neutral* expectation of a cumulative cost. This approach, while being popular and attractive from a computational standpoint,

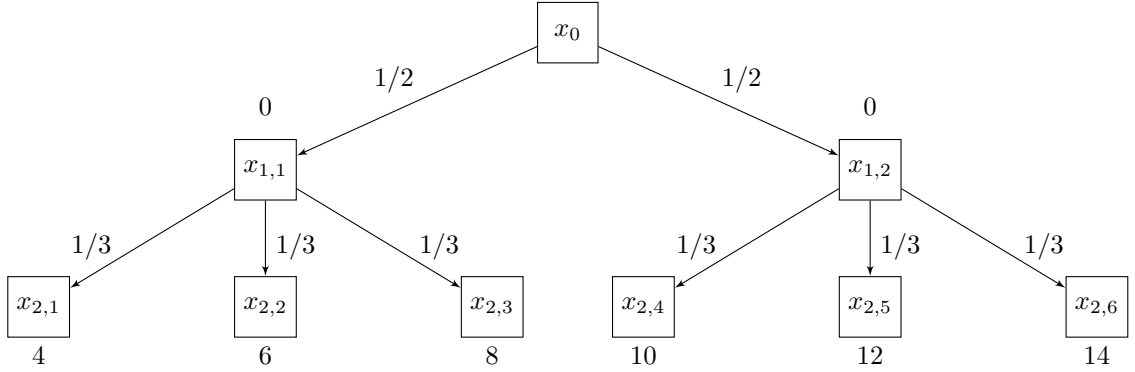
neither takes into account the *variability* of the cost (i.e., fluctuations around the mean) nor its *sensitivity* to modeling errors, and it may significantly affect overall performance [81]. Risk-sensitive MDPs [61] address the first aspect by replacing the risk-neutral expectation with a *risk-measure* of the total discounted cost, such as exponential utility, a variance-related measure and percentile risk measures (namely Value-at-Risk (VaR), or Conditional-VaR (CVaR)). Robust MDPs [92], on the other hand, address the second aspect by defining a set of plausible MDP parameters and optimize decision with respect to the expected cost under worst-case parameters. Indeed by using the representation theorem of coherent risk (Theorem 1.3.3), one can also show that Robust MDPs are equivalent to risk-sensitive MDPs with dynamic coherent risk metrics. Thus the problem of controlling cost variability and robustness in modeling errors of MDPs is equivalent to *risk shaping*, i.e., to construct a “good” risk criterion in a manner that is both conceptually meaningful and computationally tractable. While there are numerous off-the-shelf risk metrics available in the literature (for example, see the overview of risk metrics in Section 1.3.3 to Section 1.3.5), oftentimes risk shaping still presents a nontrivial challenge to system designers.

### 1.3.2 Entropic Risk and its Limitations

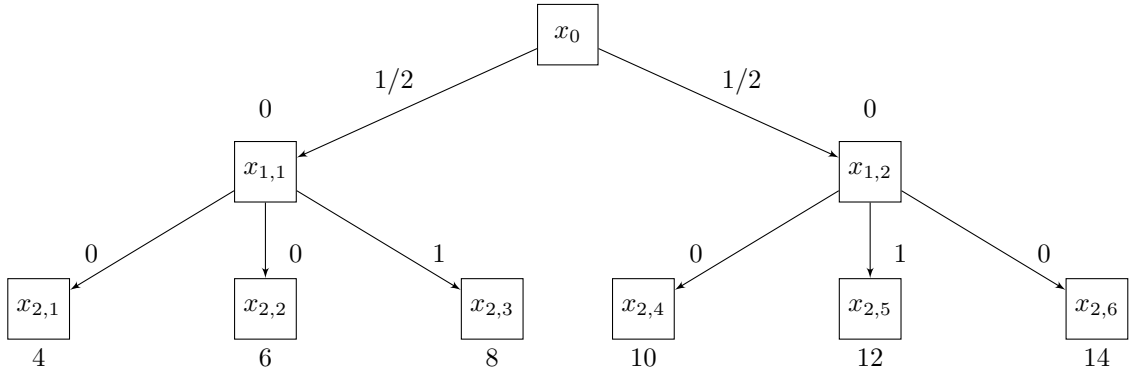
Although most disturbances are not normally distributed, the Markowitz mean-variance criterion [84], which relies on the first two moments of the distribution, has dominated risk management for over 50 years. However, solving a multi-stage stochastic optimal control problem using the mean-variance criterion is often computationally intractable [82]. This motivates the use of more computationally feasible metrics such as the entropic risk:  $\rho(X) = \log(\mathbb{E}[e^{\theta X}]) / \theta$ ,  $\theta \in (0, 1)$  [61]. Notice that the first two terms of the Taylor series expansion of  $\rho(X)$  form a weighted sum of mean and variance with regularizer  $\theta$ , i.e.  $\rho(X) \approx \mathbb{E}(X) + \theta \mathbb{E}(X - \mathbb{E}[X])^2$ .

Contrary to its popularity in literature, practical applications of the entropic risk metric have proven to be problematic [23]. The primary concerns are that the optimal control policies heavily weight a small number of risk averse decisions [56, 59] and are extremely sensitive to errors in the distribution models. Example 1.3.1 provides a counter-example that illustrates this issue in entropic risk. While the importance of risk aversion is clear from a financial standpoint [154, 153], a deficiency in the “exploration” characteristics of a policy is undesirable in many engineering applications. Consider a robotic terrain-mapping mission where the goal is to deploy a swarm of mini-drones for cost effective exploration and the actions represent the routes chosen for each robotic agent. If one optimizes the cost of routing using the mean-variance risk, the optimal policy would only consider a select few cost-effective routes, which defeats the purpose of a mapping mission. Therefore, to balance exploration (discover new terrain) and exploitation (find cost effective routes), one should consider alternative risk metrics, e.g. conditional value at risk (CVaR), which only prohibits exploration along dangerous terrains yet is not as conservative as the worst-case approach.

**Example 1.3.1.** Consider the following 3-stage, finite state example where one has the following state transitions if policy  $\pi_1$  is executed:



and one has the following state transitions if policy  $\pi_2$  is executed:



Now for  $\theta = 2$ , the entropic risk with respect to policy  $\pi_1$  and  $\pi_2$  is 13.113 and 11.654 respectively. On the other hand, consider the conditional value-at-risk  $\text{CVaR } \rho(Z) = \min_{\nu} \nu + \frac{1}{\alpha} E[Z - \nu]_+$ . With confidence interval  $\alpha = 0.6$ , the corresponding risk incurred by policy  $\pi_1$  and  $\pi_2$  is 10.833 and 12 respectively. Thus one obtains a less diversified policy  $\pi_2$  (which only has non-zero probabilities on  $x_{2,3}$  and  $x_{2,5}$ ) when entropic risk is optimized, in comparison to the uniform policy  $\pi_1$  (which is optimal to the minimization of CVaR risk).

Due to these potential issues arising in entropic risk measures, several approaches differing from the standard expectation or entropic risk, have been studied in sequential decision making. In [49], the authors considered the maximization of a strictly concave functional of the distribution of the terminal state. In [159, 31, 55], risk-sensitive MDPs are cast as the problem of maximizing percentile performance. Variance-related risk metrics are considered, e.g., in [137, 54]. Other mean, variance, and probabilistic criteria for risk-sensitive MDPs are discussed in the survey [157].

In the rest of this section, we briefly describe the theory of percentile, coherent and dynamic risk metrics, on which we will rely extensively in the later chapters. The material presented in this section summarizes several novel results in risk theory achieved in the past ten years. Our presentation strives to present this material in an intuitive fashion and with a notation tailored to control engineering, machine learning, and operations research applications.

### 1.3.3 Percentile Risk Metrics

Let  $Z$  be a bounded-mean random variable, i.e.,  $\mathbb{E}[|Z|] < \infty$ , on a probability space  $(\Omega, \mathcal{H}, \mathbb{P})$ , with cumulative distribution function  $F(z) = \mathbb{P}(Z \leq z)$ . In this paper we interpret  $Z$  as a cost. The *value-at-risk* (VaR) at confidence level  $\alpha \in (0, 1)$  is the  $1 - \alpha$  quantile of  $Z$ , i.e.,  $\text{VaR}_\alpha(Z) = \min \{z \mid F(z) \geq 1 - \alpha\}$ .

$$\text{VaR}_\alpha(Z) = \min \{z \mid F(z) \geq \alpha\}. \quad (1.1)$$

The minimum in (1.1) is attained because  $F$  is non-decreasing and right-continuous in  $z$ . When  $F$  is continuous and strictly increasing,  $\text{VaR}_\alpha(Z)$  is the unique  $z$  satisfying  $F(z) = \alpha$ ; otherwise, (1.1) can have no solution or a whole range of solutions. The VaR risk measure has many fundamental *engineering applications* such as motion planning, where a safety constraint is imposed to create an upper-bound of the probability of maneuvering into dangerous regimes.

Although VaR is a popular risk measure, it suffers from being unstable and difficult to work with numerically when  $Z$  is not normally distributed, which is often the case as loss distributions tend to exhibit fat tails or empirical discreteness. Moreover, VaR is not a *coherent* risk measure [7] and more importantly does not quantify the losses that might be suffered beyond its value at the  $\alpha$ -tail of the distribution [113].

In many *financial applications* such as portfolio optimization where the probability of undesirable events could be small but the cost incurred could still be significant, besides describing risk as the probability of incurring costs, it will be more informative to study the cost in the tail of the risk distribution. An alternative measure that addresses most of the VaR's shortcomings is *conditional value-at-risk*,  $\text{CVaR}_\alpha(Z)$ , which is the mean of the  $\alpha$ -tail distribution of  $Z$ . If there is no probability atom at  $\text{VaR}_\alpha(Z)$ ,  $\text{CVaR}_\alpha(Z)$  has a unique value that is defined as

$$\text{CVaR}_\alpha(Z) = \min_{w \in \mathbb{R}} \left\{ w + \frac{1}{\alpha} \mathbb{E}[(Z - w)^+] \right\}, \quad (1.2)$$

where  $(x)^+ = \max(x, 0)$  represents the positive part of  $x$ . If there is no probability atom at  $\text{VaR}_\alpha(Z)$ , it is well known from Theorem 6.2 in [132] that  $\text{CVaR}_\alpha(Z) = \mathbb{E}[Z \mid Z \geq \text{VaR}_\alpha(Z)]$ . Therefore,  $\text{CVaR}_\alpha(Z)$  may be interpreted as the worst-case expected value of  $Z$ , conditioned on the  $\alpha$ -portion of the tail distribution. It is well known that  $\text{CVaR}_\alpha(Z)$  is decreasing in  $\alpha$ ,  $\text{CVaR}_1(Z)$  equals to  $\mathbb{E}(Z)$ , and  $\text{CVaR}_\alpha(Z)$  tends to  $\max(Z)$  as  $\alpha \downarrow 0$ . CVaR is especially useful for controlling rare, but potentially disastrous events, which occur below the  $1 - \alpha$  quantile, and are neglected by VaR [127]. Furthermore, CVaR enjoys desirable axiomatic properties, such as coherence [7]. We refer to [112] for further motivation with respect to CVaR and a comparison with other risk measures such as VaR.

A useful property of CVaR, which we exploit in this paper, is its alternative dual representation [7]:

$$\text{CVaR}_\alpha(Z) = \max_{\xi \in \mathcal{U}_{\text{CVaR}}(\alpha, \mathbb{P})} \mathbb{E}_\xi[Z], \quad (1.3)$$

where  $\mathbb{E}_\xi[Z]$  denotes the  $\xi$ -weighted expectation of  $Z$ , and the *risk-envelope*  $\mathcal{U}_{\text{CVaR}}$  is given by

$$\mathcal{U}_{\text{CVaR}}(\alpha, \mathbb{P}) = \left\{ \xi : \xi(\omega) \in \left[0, \frac{1}{\alpha}\right], \int_{\omega \in \Omega} \xi(\omega) \mathbb{P}(\omega) d\omega = 1 \right\}.$$

Thus, the CVaR of a random variable  $Z$  may be interpreted as the worst-case expectation of  $Z$  under a perturbed distribution  $\xi\mathbb{P}$ .

Accordingly, in Chapter 2 and 3, we will focus on sequential decision making with *percentile risk measures* characterized by VaR and CVaR.

### 1.3.4 Static, Coherent Measures of Risk

Consider a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , where  $\Omega$  is the set of outcomes (sample space),  $\mathcal{F}$  is a  $\sigma$ -algebra over  $\Omega$  representing the set of events we are interested in, and  $\mathbb{P}$  is a probability measure over  $\mathcal{F}$ . In this paper we will focus on disturbance models characterized by probability *mass* functions, hence we restrict our attention to finite probability spaces (i.e.,  $\Omega$  has a finite number of elements or, equivalently,  $\mathcal{F}$  is a finitely generated algebra). Denote with  $\mathcal{Z}$  the space of random variables  $Z : \Omega \mapsto (-\infty, \infty)$  defined over the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . In this paper a random variable  $Z \in \mathcal{Z}$  is interpreted as a cost, i.e., the smaller the realization of  $Z$ , the better. For  $Z, W$ , we denote by  $Z \leq W$  the point-wise partial order, i.e.,  $Z(\omega) \leq W(\omega)$  for all  $\omega \in \Omega$ .

By a *risk measure* (or *risk metric*, we will use these terms interchangeably) we understand a function  $\rho(Z)$  that maps an uncertain outcome  $Z$  into the extended real line  $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ . In this paper we restrict our analysis to *coherent risk measures*, defined as follows:

**Definition 1.3.2** (Coherent Risk Measures). *A coherent risk measure is a mapping  $\rho : \mathcal{Z} \rightarrow \mathbb{R}$ , satisfying the following four axioms:*

- A1 Convexity:**  $\rho(\lambda Z + (1 - \lambda)W) \leq \lambda\rho(Z) + (1 - \lambda)\rho(W)$ , for all  $\lambda \in [0, 1]$  and  $Z, W \in \mathcal{Z}$ ;
- A2 Monotonicity:** if  $Z \leq W$  and  $Z, W \in \mathcal{Z}$ , then  $\rho(Z) \leq \rho(W)$ ;
- A3 Translation invariance:** if  $a \in \mathbb{R}$  and  $Z \in \mathcal{Z}$ , then  $\rho(Z + a) = \rho(Z) + a$ ;
- A4 Positive homogeneity:** if  $\lambda \geq 0$  and  $Z \in \mathcal{Z}$ , then  $\rho(\lambda Z) = \lambda\rho(Z)$ .

These axioms were originally conceived in [7] and ensure the “rationality” of single-period risk assessments (we refer the reader to [7] for a detailed motivation of these axioms). One of the main properties for coherent risk metrics is the universal representation theorem [132], which establishes the connection between coherent risk and distributionally robust expectation.

**Theorem 1.3.3.** *A risk measure  $\rho : \mathcal{Z} \rightarrow \mathbb{R}$  is coherent if and only if there exists a convex bounded and closed set  $\mathcal{U} \subset \mathcal{B}$  such that<sup>2</sup>*

$$\rho(Z) = \max_{\xi \in \mathcal{P} \in \mathcal{U}(P)} \mathbb{E}_\xi[Z]. \quad (1.4)$$

<sup>2</sup>When we study risk in MDPs, the risk-envelope  $\mathcal{U}(P)$  in Eq. 1.4 also depends on the state  $x$  and action  $a$ .

The result essentially states that any coherent risk measure is an expectation w.r.t. a worst-case density function  $\xi P$ , chosen adversarially from a suitable set of test density functions  $\mathcal{U}(P)$ , referred to as *risk envelop*. Moreover, it means that any coherent risk measure is *uniquely represented* by its risk envelop. Thus, in the sequel, we shall interchangeably refer to coherent risk-measures either by their explicit functional representation, or by their corresponding risk-envelop.

### 1.3.5 Dynamic, Time-Consistent Measures of Risk

Having motivated the need for risk-sensitive optimal control using metrics, we now address the challenges associated with appropriately quantifying risk in multi-period scenarios. Oftentimes, it appears to be difficult to model risk in multi-period settings in a way that matches intuition [87]. In particular, a common strategy to include risk-aversion in multi-period contexts is to apply a *static* risk metric, which assesses risk from the perspective of a single point in time, to the total cost of the future stream of random outcomes. However, due to the inability of risk re-evaluation in subsequent stages, using static risk metrics in multi-period decision problems can lead to an over- or under-estimation of the true dynamic risk, as well as potentially “inconsistent” behavior (see [62] and references therein).

This section provides a multi-period generalization of the concepts presented in Section 1.3.4 and follows closely the discussion in [119]. Consider a probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ , a filtration  $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \cdots \subset \mathcal{F}_N \subset \mathcal{F}$ , and an adapted sequence of real-valued random variables  $Z_k, k \in \{0, \dots, N\}$ . We assume that  $\mathcal{F}_0 = \{\Omega, \emptyset\}$ , i.e.,  $Z_0$  is deterministic. The variables  $Z_k$  can be interpreted as stage-wise costs. For each  $k \in \{0, \dots, N\}$ , denote with  $\mathcal{Z}_k$  the space of random variables defined over the probability space  $(\Omega, \mathcal{F}_k, \mathbb{P})$ ; also, let  $\mathcal{Z}_{k,N} := \mathcal{Z}_k \times \cdots \times \mathcal{Z}_N$ . Given sequences  $Z = \{Z_k, \dots, Z_N\} \in \mathcal{Z}_{k,N}$  and  $W = \{W_k, \dots, W_N\} \in \mathcal{Z}_{k,N}$ , we interpret  $Z \leq W$  component-wise, i.e.,  $Z_j \leq W_j$  for all  $j \in \{k, \dots, N\}$ .

The fundamental question in the theory of dynamic risk measures is the following: how do we evaluate the risk of the sequence  $\{Z_k, \dots, Z_N\}$  from the perspective of stage  $k$ ? The answer, within the modern theory of risk, relies on two key intuitive facts [119]. First, in dynamic settings, the specification of risk preferences should no longer entail constructing a single risk metric but rather a *sequence* of risk metrics  $\{\rho_{k,N}\}_{k=0}^N$ , each mapping a future stream of random costs into a risk metric/assessment at time  $k$ . This motivates the following definition.

**Definition 1.3.4** (Dynamic Risk Measure). *A dynamic risk measure is a sequence of mappings  $\rho_{k,N} : \mathcal{Z}_{k,N} \rightarrow \mathcal{Z}_k, k \in \{0, \dots, N\}$ , obeying the following monotonicity property:*

$$\rho_{k,N}(Z) \leq \rho_{k,N}(W) \text{ for all } Z, W \in \mathcal{Z}_{k,N} \text{ such that } Z \leq W.$$

The above monotonicity property (analogous to axiom A2 in Definition 1.3.2) is, arguably, a natural requirement for any meaningful dynamic risk measure.

The second intuitive fact is that the sequence of metrics  $\{\rho_{k,N}\}_{k=0}^N$  should be constructed so that the risk preference profile is *consistent* over time [43, 130, 62]. A widely accepted notion of time-consistency is as

follows [119]: if a certain outcome is considered less risky in all states of the world starting at stage  $k + 1$ , then it should also be considered less risky starting at stage  $k$ .

The following example (adapted from [115]) shows how dynamic risk measures as defined above might indeed result in *time-inconsistent*, and ultimately undesirable, behaviors.

**Example 1.3.5.** Consider the simple setting whereby there is a final cost  $Z$  and one seeks to evaluate such cost from the perspective of earlier stages. Consider the three-stage scenario tree in Figure 1.1, with the elementary events  $\Omega = \{UU, UD, DU, DD\}$ , and the filtration  $\mathcal{F}_0 = \{\emptyset, \Omega\}$ ,  $\mathcal{F}_1 = \{\emptyset, \{U\}, \{D\}, \Omega\}$ , and  $\mathcal{F}_2 = 2^\Omega$ . Consider the dynamic risk measure:

$$\rho_{k,N}(Z) := \max_{\xi \in \mathcal{U}} \mathbb{E}_\xi[Z | \mathcal{F}_k], \quad k = 0, 1, 2$$

where  $\mathcal{U}$  contains two probability measures, one corresponding to  $p = 0.4$ , and the other one to  $p = 0.6$ . Assume that the random cost is  $Z(UU) = Z(DD) = 0$ , and  $Z(UD) = Z(DU) = 100$ . Then, one has  $\rho_1(Z)(\omega) = 60$  for all  $\omega$ , and  $\rho_0(Z)(\omega) = 48$ . Therefore,  $Z$  is deemed strictly riskier than a deterministic cost  $W = 50$  in all states of nature at time  $k = 1$ , but nonetheless  $W$  is deemed riskier than  $Z$  at time  $k = 0$ , which is a paradox!

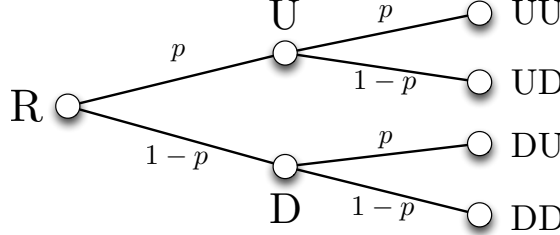


Figure 1.1: Scenario Tree for Example 1.3.5.

It is important to note that there is nothing special about the selection of this example, similar paradoxical results could be obtained with other risk metrics. We refer the reader to [119, 130, 62] for further insights into the notion of time consistency and its practical relevance. The issue then is what additional “structural” properties are required for a dynamic risk measure to be time consistent. We first provide a rigorous version of the previous definition of time-consistency.

**Definition 1.3.6** (Time Consistency ([119])). A dynamic risk measure  $\{\rho_{k,N}\}_{k=0}^N$  is time-consistent if, for all  $0 \leq l < k \leq N$  and all sequences  $Z, W \in \mathcal{Z}_{l,N}$ , the conditions

$$\begin{aligned} Z_i &= W_i, \quad i = l, \dots, k-1, \text{ and} \\ \rho_{k,N}(Z_k, \dots, Z_N) &\leq \rho_{k,N}(W_k, \dots, W_N), \end{aligned} \tag{1.5}$$



imply that

$$\rho_{l,N}(Z_l, \dots, Z_N) \leq \rho_{l,N}(W_l, \dots, W_N).$$

As we will see in Theorem 1.3.8, the notion of time-consistent risk measures is tightly linked to the notion of coherent risk measures, whose generalization to the multi-period setting is given below:

**Definition 1.3.7** (Coherent One-step Conditional Risk Measures ([119])). *A coherent one-step conditional risk measure is a mapping  $\rho_k : \mathcal{Z}_{k+1} \rightarrow \mathcal{Z}_k$ ,  $k \in \{0, \dots, N-1\}$  with the following four properties:*

- *Convexity:*  $\rho_k(\lambda Z + (1-\lambda)W) \leq \lambda \rho_k(Z) + (1-\lambda)\rho_k(W)$ ,  $\forall \lambda \in [0, 1]$  and  $Z, W \in \mathcal{Z}_{k+1}$ ;
- *Monotonicity:* if  $Z \leq W$  then  $\rho_k(Z) \leq \rho_k(W)$ ,  $\forall Z, W \in \mathcal{Z}_{k+1}$ ;
- *Translation invariance:*  $\rho_k(Z + W) = Z + \rho_k(W)$ ,  $\forall Z \in \mathcal{Z}_k$  and  $W \in \mathcal{Z}_{k+1}$ ;
- *Positive homogeneity:*  $\rho_k(\lambda Z) = \lambda \rho_k(Z)$ ,  $\forall Z \in \mathcal{Z}_{k+1}$  and  $\lambda \geq 0$ .

We are now in a position to state the main result of this section.

**Theorem 1.3.8** (Dynamic, Time-consistent Risk Measures [119]). *Consider, for each  $k \in \{0, \dots, N\}$ , the mappings  $\rho_{k,N} : \mathcal{Z}_{k,N} \rightarrow \mathcal{Z}_k$  defined as*

$$\rho_{k,N} = Z_k + \rho_k(Z_{k+1} + \rho_{k+1}(Z_{k+2} + \dots + \rho_{N-2}(Z_{N-1} + \rho_{N-1}(Z_N)) \dots)), \quad (1.6)$$

where the  $\rho_k$ 's are coherent one-step conditional risk measures. Then, the ensemble of these mappings is a dynamic, time-consistent risk measure.

Remarkably, Theorem 1 in [119] shows (under weak assumptions) that the “multi-stage composition” in Equation (1.6) is indeed *necessary for time consistency*. Accordingly, in Chapter 4 and 5, we will focus on sequential decision making with *dynamic, time-consistent risk measures* characterized in Theorem 1.3.8.

With dynamic, time-consistent risk measures, the value of  $\rho_k$  at stage  $k$  is  $\mathcal{F}_k$ -measurable. Therefore in general the evaluation of risk depends on the whole past (although in a time-consistent way). On the one hand, this appears to have little value in most practical applications, on the other hand, inclusion of this risk metric in an optimization problem potentially leads to intractability from a computational standpoint (and, in particular, this structure inhibits a dynamic programming solution). Therefore, in this chapter we consider a slight refinement of the concept of dynamic, time consistent risk measures, which involves a Markovian structure in risk evaluation [119].

**Definition 1.3.9** (Markov risk measures). *Consider the MDP  $(\mathcal{X}, \mathcal{A}, C, P, \gamma, x_0)$ . Let  $\mathcal{V} := L_p(\mathcal{X}, \mathcal{B}, P)$  be the space of random variables on  $\mathcal{X}$  with finite  $p^{\text{th}}$  moment. Given a controlled state process  $\{x_k\}$  generated by the MDP, a dynamic, time-consistent risk measure is a Markov risk measure if each coherent one-step risk measure  $\rho_k : \mathcal{Z}_{k+1} \rightarrow \mathcal{Z}_k$  in equation (1.6) can be written as*

$$\rho_k(V(x_{k+1})) = \sigma_k(V(x_{k+1}), x_k, a_k, P(x_{k+1}|x_k, a_k)), \quad (1.7)$$

for all  $V(x_{k+1}) \in \mathcal{V}$  and  $a_k \in \mathcal{A}(x_k)$ , where  $\sigma_k$  is a coherent one-step risk measure on  $\mathcal{V}$  —with the additional technical property that for every  $V(x_{k+1}) \in \mathcal{V}$  and  $a_k \in \mathcal{A}(x_k)$  the function

$$x_k \mapsto \sigma_k(V(x_{k+1}), x_k, a_k, P(x_{k+1}|x_k, a_k))$$

is an element of  $\mathcal{V}$ .

In other words, the evaluation of a Markov risk measure only depends on the current state of the MDP.

**Example 1.3.10.** An important example of a coherent one-step risk measure satisfying the requirements presented in the definition of Markov risk measures (Definition 1.3.9) is the mean-semideviation risk function:

$$\rho_k(V) = \mathbb{E}[V] + \lambda \left( \mathbb{E} \left[ [V - \mathbb{E}[V]]_+^p \right] \right)^{1/p}, \quad (1.8)$$

where  $p \in [1, \infty)$ ,  $[z]_+^p := (\max(z, 0))^p$ , and  $\lambda \in [0, 1]$ .

Other important examples include the conditional average value at risk and, of course, the risk-neutral expectation [119].

**Remark 1.3.11.** Notwithstanding the time dependency that occurs in the general definition of dynamic, time consistent risk measure, in this thesis we mainly focus on its stationary counterpart, where the dynamic, time consistent risk is a composition of homogenous and time-independent one-step coherent risk metrics, i.e.,  $\rho_{k,N} = \underbrace{\rho \circ \dots \circ \rho}_{N-k}$  for any  $k \leq N$  and  $N \in \mathbb{N}$ .

## 1.4 Existing Solution Approaches and Limitations

In the previous section, we reviewed several classes of risk metrics that are commonly used in risk-sensitive decision making problems. Nevertheless, inclusion of risk-awareness in MDPs is still difficult for several reasons. First, it appears to be difficult to model risk in multi-period settings in a way that matches our intuition of risk awareness; in particular widely adopted constraints such as variance or probability constraints lead to irrational behaviors. Second, MDPs involving risk metrics tend to be computationally intractable; for example, optimization under variance constraint and percentile optimization have been shown to be NP-hard in general [82]. Limitations of the state of the art to address such challenges are provided below.

### 1.4.1 Time Inconsistency in Risk-aware Planning

The most common strategy to model risk awareness in MDPs is to consider a static risk (i.e., a metric assessing risk from the perspective of a *single* point in time) applied to the entire stream of future costs. Typical examples include variance-constrained MDPs [104, 136, 145], or problems with probability constraints [104, 94, 2, 155, 29]. However, since static risks do not involve an incremental reassessment of uncertainties at subsequent decision stages, they generally lead to *irrational* behaviors. For example, an agent may

decide to incur losses or may avoid visiting states that are favorable under any uncertainty realizations. In this subsection, we will illustrate some irregular behaviors in risk sensitive multi-period planning by two examples.

**Example 1: Variance-constrained planning** — Given an MDP with time horizon  $T > 0$ , state space  $\mathcal{X}$ , action space  $\mathcal{A}$ , state transition  $x_{k+1} \sim P(\cdot|x_k, a_k)$  where  $x_k \in \mathcal{X}$  and  $a_k \in \mathcal{A}$  for  $T > k \geq 0$ , initial state  $x_0 \in \mathcal{X}$ , cost function  $c_k : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$  and constraint cost function  $d_k : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}$  for  $k \in \{0, \dots, T-1\}$ , solve

$$\begin{aligned} \min_{\pi \in \Pi_H} \quad & \mathbb{E} \left[ \sum_{k=0}^{T-1} c_k(x_k, a_k) \right] \\ \text{subject to} \quad & \text{var} \left( \sum_{k=0}^{T-1} d_k(x_k, a_k) \right) \leq r_0, \end{aligned}$$

where  $\Pi_H$  is the set of history-dependent policies and  $r_0 \in \mathbb{R}$  is a user-provided risk threshold.

Consider the example in Figure 1.2 in which all the cost functions are homogeneous in time. When the risk threshold  $r_0$  is below 25, policy  $\pi_1$  is infeasible and the optimal policy is  $\pi_2$ . According to policy  $\pi_2$ , if the decision maker does not incur a cost in the first stage it *seeks to incur losses* in subsequent stages to keep the variance small. This can be seen as a consequence of the fact that Bellman's principle of optimality does not hold for this class of problems.

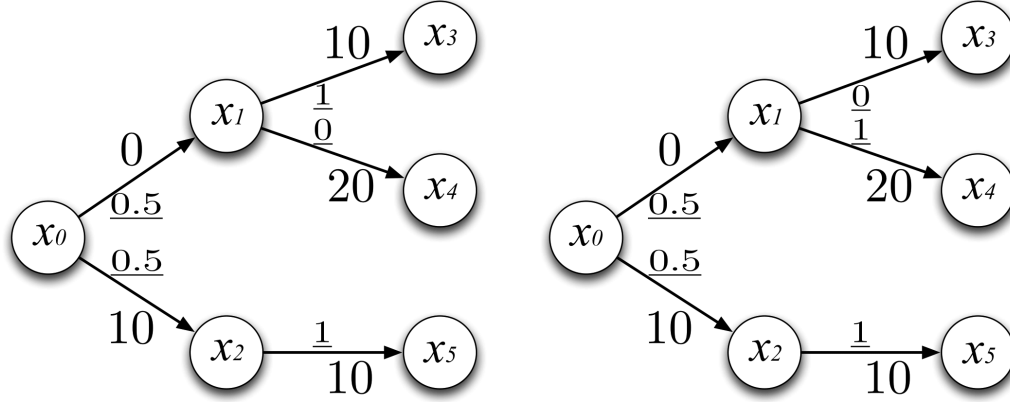
As a second example, we consider MDPs with probability (chance) constraints. Intuitively, this risk-constrained decision-making problem serves as the foundation of many safe-planning tasks in engineering and robotics.

**Example 2: Chance-constrained planning** — Given an MDP with time horizon  $T > 0$ , state space  $\mathcal{X}$ , action space  $\mathcal{A}$ , state transition  $x_{k+1} \sim P(\cdot|x_k, a_k)$  where  $x_k \in \mathcal{X}$  and  $a_k \in \mathcal{A}$  for  $T > k \geq 0$ , initial state  $x_0 \in \mathcal{X}$ , terminal cost function  $c : \mathcal{X} \rightarrow \mathbb{R}$  and constraint cost function  $d : \mathcal{X} \rightarrow \mathbb{R}$ , solve

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E} [c(x_T)] \\ \text{subject to} \quad & \mathbb{P} \left( d(x_T) \leq r_0 \right) \geq \alpha, \end{aligned}$$

where  $r_0 \in \mathbb{R}$  is a user-provided risk threshold.

Accordingly consider the example in Figure 1.3. Here we interpret the constraint costs  $d$  as acceptable if it is negative and unacceptable otherwise. Also let the constraint cost threshold  $r_0$  be 0 and confidence level  $\alpha$  be  $2/3$ . One can show that the problem (consisting of a single policy) is infeasible, since at the first stage  $\mathbb{P}(d(x_T) \leq 0) < 2/3$ . On the other hand, the constraint costs are acceptable in every state of the world from the perspective of the subsequent stage. In other words, the decision-maker would deem infeasible a problem that, at the second stage, appears feasible under any possible realization of the uncertainties.



(a) Stage-wise constraint and objective function costs and transition probabilities for policy  $\pi_1$ .

(b) Stage-wise constraint and objective function costs and transition probabilities for policy  $\pi_2$ .

Figure 1.2: Limitations of mean-variance optimization. Underlined numbers along the edges represent transition probabilities; non-underlined numbers represent stage-wise constraint and objective function costs (that are equal for this example). Terminal constraint costs are zero. Under policy  $\pi_1$ , the costs per stage are given by  $d(x_0, a_0) = 0.5 \cdot 0 + 0.5 \cdot 10 = 5$ ,  $d(x_1, a_1) = 10$ , and  $d(x_2, a_1) = 10$ ; under policy  $\pi_2$ , the costs per stage are given by  $d(x_0, a_0) = 5$ ,  $d(x_1, a_1) = 20$ , and  $d(x_2, a_1) = 10$ . One can verify that for policy  $\pi_1$  one has  $\text{var}\left(\sum_{k=0}^{N-1} d(x_k, a_k)\right) = 25$ , while for policy  $\pi_2$  one has  $\text{var}\left(\sum_{k=0}^{N-1} d(x_k, a_k)\right) = 0$ . Then, if the risk threshold is less than 25, the decision-maker would choose policy  $\pi_2$  and would seek to incur losses in order to keep the variance small enough.

Notice that similar there is nothing particularly special about the aforementioned risk-constrained problems. Similar paradoxical results could be obtained from other constrained planning problems with static risk constraints or from risk-sensitive control problems that relies on static risk metrics [119]. Subsequently, we will refer to the aforementioned irrational behaviors as “time-inconsistency”, since the corresponding solution policy is an inconsistent state-action mapping at different decision stages and risk levels.

### 1.4.2 Limitation 2: Complexity of solution approaches

Risk-aware planning falls under the category of stochastic decision-making, with the extra complexity that Bellman’s principle of optimality does not hold in the general setting, and Markovian (and stationary) policies are no longer optimal. Moreover, as mentioned before, optimization under static risk constraints such as variance constraints and probability constraints have been shown to be NP-hard [82]. Therefore current approaches to risk-constrained planning are mostly limited to global search methods such as mixed-integer linear-programming [16] and branch-and-bound techniques [140], whose applications are restricted to low-dimensional problems with relatively simple constraints (i.e., stochastic constraints that are induced by Gaussian distributions), dynamics (i.e., linear dynamics), and cost functions (i.e., quadratic cost). In contrast, by choosing risk metrics that results in rational decision-making and enjoys a compositionally efficient structure,

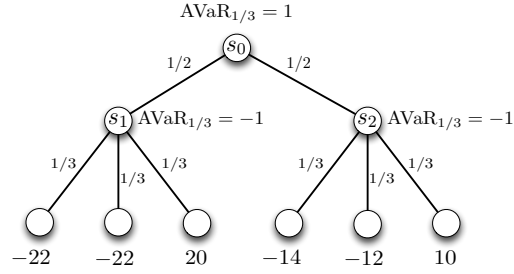


Figure 1.3: Limitations of chance-constrained optimization. The numbers along the edges represent transition probabilities, while the numbers below the terminal nodes represent the stage-wise constraint costs. The problem involves a single control policy (hence there is a unique transition graph). The constraint cost appears acceptable in states  $x_1$  and  $x_2$ , but unacceptable from the perspective of the first stage in state  $x_0$ .

we can address a more general class of risk-aware planning problems via the dynamic programming approach, which further leads to the development of reinforcement learning based solution algorithms that contain provable performance guarantees and are amenable to large-scale implementations.

## 1.5 Risk-sensitive Decision Making Versus Reward Shaping

To motivate our research in risk-sensitive decision making, we first compare risk-sensitive decision making with reward shaping, a standard technique that is used in literature of risk-neutral decision making, where one imposes safety guarantees by customizing reward functions of the MDP in particular regions of state and action spaces. While it has been shown in [61] that reward shaping is equivalent to risk-sensitive decision making with entropic risk measures, in several occasions where one also aims at maximizing the robustness of policies, the limitation of reward shaping can be clearly seen.

Recently it has been studied in [109] that, risk-sensitive reinforcement learning using conditional value-at-risk (CVaR) can be used to guarantee robustness and reduce sample complexity, when one solves for a policy using an ensemble of simulated source domains, and deploys this generalized policy to a broad range of possible target domains, including un-modeled effects. This result is important for applying reinforcement learning to tackle real-world tasks, especially when the policies are represented using rich function approximations like deep neural networks. Furthermore, it can be shown that risk-sensitive decision making with a CVaR objective function articulates a special form of adversarial training in policy learning, where one also takes the tail distribution of worst-case events into the account. Equipped with this feature, it allows additional guarantees in performance transfer in between domains, even when there are significant discrepancies between the simulated source domain and the target domain. Contrarily, it can be challenging to apply reward shaping in this case mostly because the uncertainties are originated from the modeling errors across domains. It has also been shown in [100] that, using reward-shaping to solve an optimal control problem with model uncertainties often leads to a larger sub-optimality gap.

Another advantage of adopting risk-sensitive decision making over reward shaping is its ability to capture sequential risk sensitivity. To illustrate this behavior, consider an example of a soccer game that is composed of complicated strategies, such as attack and defend, based on the status of the game (see [79] for detailed discussions). Consider a team that is one score behind, and there is only ten minutes remaining. In this case, the team needs to play aggressively, such as making long passes and shooting from distance, to maximize the likelihood to score before the game ends. In the opposite way, if the team is winning by one goal with ten minutes remaining, the team needs to play conservatively in order to prevent its opponent from scoring goals. While both scenarios share the same cumulative objective function, which is to score more and win the game, it is clear that by optimizing the strategy with risk-aware decision making techniques, one can generate policies under various risk attitudes, and allow the policy to adapt to real-time changes. However, to the best of our knowledge, it is unclear how one can design such a strategy with standard reward shaping techniques.

## 1.6 Thesis Contributions and Outline

To address the aforementioned limitations in risk-aware planning, at a high level this thesis investigates several important aspects in the aforementioned risk-sensitive sequential decision-making problem by taking into account the variability of stochastic costs and the robustness to modeling errors. In particular,

- we address decision-making problems in which the percentile (tail) risks are considered;
- we analyze a unifying planning framework with *coherent risk* measures, which is robust to inherent uncertainties and modeling errors, and the resulting risk-sensitive decision-making problem outputs a rational, *time-consistent* policy;
- we develop tractable algorithms for large scale risk-sensitive decision-making problems and extend these approaches to *data-driven* setups.

Based on the previously reviewed background knowledge on Markov decision processes (MDPs), constrained Markov decision processes (CMDPs), and theories on both static and dynamic risk measures in this chapter, we outline the content of the four chapters dedicated to each of these questions below, leaving the literature review and the precise statement of contributions to the introduction section of each chapter.

In Chapter 2, we investigate the well-known conditional value-at-risk (CVaR) MDP problem and propose a scalable approximate value-iteration algorithm on an augmented state space. In addition, we discover an interesting relationship between the CVaR risk of total cost and the worst-case expected cost under adversarial model perturbations, which leads to a robustness framework that is significantly less conservative than robust-MDP.

In Chapter 3, we study the CMDP problem formulation whose constraints are modeled via percentile risks, such as CVaR and tail event probability. We also propose novel policy gradient and actor-critic algorithms for CVaR-constrained and chance-constrained optimization in MDPs, and illustrate the effectiveness of our algorithms using an optimal stopping problem and a personalized ad-recommendation problem.

In Chapter 4, we propose a framework for risk-averse model predictive control (MPC), where the risk metric is chosen to be the time-consistent Markov risk. We also present a solution algorithm that can be implemented using semidefinite programming techniques, study its performance in terms of sub-optimality gap, and numerically illustrate its superiority over existing risk-neutral MPC methods.

In Chapter 5, we study a dynamic programming approach to stochastic optimal control problems with dynamic, time-consistent (in particular Markov) risk constraints. In particular we show that the optimal cost functions could be computed by value iteration and that the optimal control policies can be constructed recursively. Furthermore we propose both a uniform-grid discretization algorithm and an interpolation-based approximate dynamic programming algorithm for the solution of this stochastic optimal control problem.

## Chapter 2

# Risk-Sensitive Decision Making: A CVaR Optimization Approach

### 2.1 Introduction

#### 2.1.1 Risk Sensitive Decision Making with CVaR

In this work we consider risk-sensitive MDPs with a CVaR objective, referred to as CVaR MDPs. CVaR [7, 112] is a risk-measure that is rapidly gaining popularity in various engineering applications, for instance, finance, due to its favorable computational properties [7] and superior ability to safeguard a decision maker from the “outcomes that hurt the most” [127]. In this work, by *relating risk to robustness*, we derive a novel result that further motivates the usage of a CVaR objective in a Decision Making context. Specifically, we show that the CVaR of a discounted cost in an MDP is *equivalent* to the expected value of the same discounted cost in the presence of worst-case perturbations of the MDP parameters (specifically, transition probabilities), provided that such perturbations are within a certain error budget. This result suggests CVaR MDP as a method for decision making under *both* cost variability *and* model uncertainty, which in turn suggests its usefulness as a *unified framework* for planning under uncertainty.

Risk-sensitive MDPs have been studied for over four decades, with earlier efforts focusing on exponential utility [61], mean-variance [137], and percentile risk criteria [55]. Recently, for the reasons explained above, several authors have investigated CVaR MDPs [112]. Specifically, in [36], the authors propose a dynamic programming algorithm for finite-horizon risk-constrained MDPs where risk is measured according to CVaR. The algorithm is proven to asymptotically converge to an optimal risk-constrained policy. However, the algorithm involves computing integrals over continuous variables (Algorithm 1 in [36]) and, in general, its implementation appears extraordinarily difficult. In [10], the authors investigate the structure of CVaR optimal policies and show that a Markov policy is optimal on an augmented state space, where the additional (continuous) state variable is represented by the running cost. In [60], the authors leverage this result to



design an algorithm for CVaR MDPs that relies on discretizing occupation measures in the augmented-state MDP. This approach, however, involves solving a non-convex program via a sequence of linear-programming approximations, which can only be shown to converge asymptotically. A different approach is taken by [46], [105] and [146], which consider a finite dimensional parameterization of control policies, and show that a CVaR MDP can be optimized to a *local* optimum using stochastic gradient descent (policy gradient). A recent result by Pflug and Pichler [102] showed that by using a state-augmentation procedure different from the one in [10], CVaR MDPs admit a dynamic programming formulation. However in this formulation, the augmented state is also continuous, making the design of a solution algorithm challenging.

### 2.1.2 Chapter Contribution

The contribution of this chapter is twofold. First, as discussed above, we provide a novel interpretation for CVaR MDPs in terms of robustness to modeling errors. This result is of independent interest and further motivates the usage of CVaR MDPs for decision making under uncertainty. Second, we provide a new optimization algorithm for CVaR MDPs, which leverages the state augmentation procedure introduced by Pflug and Pichler [102]. We overcome the aforementioned computational challenges (due to the continuous augmented state) by designing an algorithm that merges approximate value iteration [17] with linear interpolation. Remarkably, we are able to provide explicit error bounds and convergence rates based on contraction-style arguments. In contrast to the algorithms in [36, 60, 46, 146], given the explicit MDP model our approach leads to finite-time error guarantees with respect to the *globally* optimal policy. In addition, our algorithm is significantly simpler than previous methods, and calculates the optimal policy *for all* CVaR confidence intervals and initial states simultaneously. The practicality of our approach is demonstrated in numerical experiments involving planning a path on a grid with thousands of states. To the best of our knowledge, this is the first algorithm to approximate globally-optimal policies for non-trivial CVaR MDPs whose error depends on the resolution of interpolation.

### 2.1.3 Chapter Organization

This chapter is structured as follows: In Section 2.2 we provide background on CVaR and MDPs. We then state the problem we wish to solve (i.e., CVaR MDPs), and motivate the CVaR MDP formulation by establishing a novel relation between CVaR and model perturbations. Section 2.3 provides the basis for our solution algorithm, based on a Bellman-style equation for the CVaR. Then, in Section 2.4 we present our algorithm and correctness analysis. In Section 2.7 we evaluate our approach via numerical experiments. Finally, complete proofs of the technical results can be found in Section 7.1.

## 2.2 Problem Formulation and Motivation

### 2.2.1 Problem Formulation

Let  $C(x_t, a_t)$  denote the stage-wise costs observed along a state/control trajectory in the MDP model. The risk-sensitive discounted-cost problem we wish to address is as follows:

$$\min_{\mu \in \Pi_H} \text{CVaR}_\alpha \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \middle| x_0, \mu \right), \quad (2.1)$$

where  $\mu = \{\mu_0, \mu_1, \dots\}$  is the policy sequence with actions  $a_t = \mu_t(h_t)$  for  $t \in \{0, 1, \dots\}$ . We refer to problem (2.1) as CVaR MDP.

The problem formulation in (2.1) directly addresses the aspect of risk sensitivity, as demonstrated by the numerous applications of CVaR optimization in finance (see, e.g., [114, 64, 52]) and the recent approaches for CVaR optimization in MDPs [36, 60, 46, 146]. In the following, we show a new result providing additional motivation for CVaR MDPs, from the point of view of *robustness to modeling errors*.

### 2.2.2 Motivation - Robustness to Modeling Errors

We show a new result relating the CVaR objective in (2.1) to the *expected* discounted-cost in the presence of worst-case perturbations of the MDP parameters, where the perturbations are budgeted according to the “number of things that can go wrong.” Thus, by minimizing CVaR, the decision maker also guarantees *robustness* of the policy.

Consider a trajectory

$$(x_0, a_0, \dots, x_T)$$

in a finite-horizon MDP problem with transition probability  $P_t(x_t|x_{t-1}, a_{t-1})$ . We explicitly denote the time index of the transition matrices for reasons that will become clear shortly. The total probability of the trajectory is

$$P(x_0, a_0, \dots, x_T) = P_0(x_0) \cdots P_T(x_T|x_{T-1}, a_{T-1}),$$

and  $\sum_{t=0}^T \gamma^t C(x_t, a_t)$  is its discounted cost.

We consider an adversarial setting, where an adversary is allowed to change the transition probabilities at each stage, under some budget constraints. We will show that, for a specific budget and perturbation structure, the expected cost under the worst-case perturbation is equivalent to the CVaR of the cost. Thus, we shall establish that, from this perspective, being risk averse is *equivalent* to being robust against model perturbations.

For each stage  $1 \leq t \leq T$ , consider a perturbed transition kernel  $\hat{P}_t = P_t \circ \delta_t$ , where  $\delta_t \in \mathbb{R}^{\mathcal{X} \times \mathcal{A} \times \mathcal{X}}$  is a *multiplicative probability perturbation* and  $\circ$  is the Hadamard product (a.k.a. elementwise product), under the condition that  $\hat{P}_t$  is a stochastic matrix, i.e.,  $\sum_{x'} \hat{P}_t(x'|x, a) = 1 \quad \forall x \in \mathcal{X}, a \in \mathcal{A}$ , and  $\hat{P}_t(x'|x, a) \geq 0$ .

0  $\forall x', x \in \mathcal{X}, a \in \mathcal{A}$ . Let  $\Delta_t$  denote the set of perturbation matrices that satisfy this condition, and let  $\Delta = \Delta_1 \times \cdots \times \Delta_T$  denote the set of all possible perturbations to the trajectory distribution.

We now impose a budget constraint on the perturbations as follows. For some budget  $\eta \geq 1$ , we consider the constraint

$$\delta_1(x_1|x_0, a_0)\delta_2(x_2|x_1, a_1) \cdots \delta_T(x_T|x_{T-1}, a_{T-1}) \leq \eta, \quad \forall x_0, \dots, x_T \in \mathcal{X}, \forall a_0, \dots, a_{T-1} \in \mathcal{A}. \quad (2.2)$$

Essentially, the product in Eq. (2.2) states that with small budget *the worst cannot happen at each time*. Instead, the perturbation budget has to be split (multiplicatively) along the trajectory. We note that Eq. (2.2) is in fact a constraint on the perturbation matrices, and we denote by  $\Delta_\eta \subset \Delta$  the set of perturbations that satisfy this constraint with budget  $\eta$ . The following result shows an equivalence between the CVaR and the worst-case expected loss.

**Proposition 2.2.1** (Interpretation of CVaR as Robustness Measure). *It holds*

$$\text{CVaR}_{\frac{1}{\eta}} \left( \sum_{t=0}^T \gamma^t C(x_t, a_t) \right) = \sup_{(\delta_1, \dots, \delta_T) \in \Delta_\eta} \mathbb{E}_{\hat{P}} \left[ \sum_{t=0}^T \gamma^t C(x_t, a_t) \right], \quad (2.3)$$

where  $\mathbb{E}_{\hat{P}}[\cdot]$  denotes expectation with respect to a Markov chain with transitions  $\hat{P}_t$ .

While the full proof of this proposition can be found in the Appendix, it is instructive to compare Proposition 2.2.1 with the dual representation of CVaR in (1.3) where both results convert the CVaR risk into a robustness measure. Note, in particular, that the perturbation budget in Proposition 2.2.1 has a *temporal* structure, which constrains the adversary from choosing the worst perturbation at each time step.

**Remark 2.2.2.** *An equivalence between robustness and risk-sensitivity was previously suggested by Osogami [95]. In that study, the iterated (dynamic) coherent risk was shown to be equivalent to a robust MDP [63] with a rectangular uncertainty set. The iterated risk (and, correspondingly, the rectangular uncertainty set) is very conservative [160], in the sense that the worst can happen at each time step. In contrast, the perturbations considered here are much less conservative. In general, solving robust MDPs without the rectangularity assumption is NP-hard. Nevertheless, Mannor et. al. [80] showed that, for cases where the number of perturbations to the parameters along a trajectory is upper bounded (budget-constrained perturbation), the corresponding robust MDP problem is tractable. Analogous to the constraint set (1) in [80], the perturbation set in Proposition 2.2.1 limits the total number of log-perturbations along a trajectory. Accordingly, we shall later see that optimizing problem (2.1) with perturbation structure (2.2) is indeed also tractable.*

The next section provides the fundamental theoretical ideas behind our approach to the solution of (2.1).

## 2.3 Bellman Equation for CVaR

In this section, by leveraging a recent result from [102], we present a dynamic programming (DP) formulation for the CVaR MDP problem in (2.1). As we shall see, the value function in this formulation depends on both the state and the CVaR confidence level  $\alpha$ . We then establish important properties of such a dynamic programming formulation, which will later enable us to derive an efficient dynamic programming-based approximate solution algorithm and provide correctness guarantees on the approximation error. As mentioned in Section 4.1, all proofs are presented in the Appendix.

Our starting point is a recursive decomposition of CVaR, whose proof is detailed in Theorem 10 of [102].

**Theorem 2.3.1** (CVaR Decomposition, Theorem 21 in [102]). *For any  $t \geq 0$ , denote by  $Z = (Z_{t+1}, Z_{t+2}, \dots)$  the cost sequence from time  $t + 1$  onwards. The conditional CVaR under policy  $\mu$ , i.e.,  $\text{CVaR}_\alpha(Z \mid h_t, \mu)$ , obeys the following decomposition:*

$$\text{CVaR}_\alpha(Z \mid h_t, \mu) = \max_{\xi \in \mathcal{U}_{\text{CVaR}}(\alpha, P(\cdot \mid x_t, a_t))} \mathbb{E} \left[ \xi(x_{t+1}) \cdot \text{CVaR}_{\alpha \xi(x_{t+1})}(Z \mid h_{t+1}, \mu) \mid h_t, \mu \right],$$

where  $a_t$  is the action induced by policy  $\mu_t(h_t)$ , and the expectation is with respect to  $x_{t+1}$ .

Theorem 2.3.1 concerns a fixed policy  $\mu$ ; we now extend it to a general dynamic programming formulation. Specific to our problem setup, we replace the supremum operator in Theorem 21 of [102] with the maximum operator because the feasibility set  $\mathcal{U}_{\text{CVaR}}(\alpha, P(\cdot \mid x_t, a_t))$  is a convex and compact subset of real vectors, and by Theorem 10 of [113], the objective function  $\mathbb{E} \left[ \xi(x_{t+1}) \cdot \text{CVaR}_{\alpha \xi(x_{t+1})}(Z \mid h_{t+1}, \mu) \mid h_t, \mu \right]$  is a continuous function of the real vector  $\xi$ . Note that in the recursive decomposition in Theorem 2.3.1 the right-hand side involves CVaR terms with different confidence levels than the one in the left-hand side. Accordingly, we augment the state space  $\mathcal{X}$  with an additional continuous state  $\mathcal{Y} = (0, 1]$ , which corresponds to the confidence level. For any  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ , the *value-function*  $V(x, y)$  for the augmented state  $(x, y)$  is defined as:

$$V(x, y) = \min_{\mu \in \Pi_H} \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0 = x, \mu \right).$$

Similar to standard dynamic programming, it is convenient to work with operators defined on the space of value functions [17]. In our case, Theorem 2.3.1 leads to the following definition of CVaR Bellman operator  $\mathbf{T} : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{X} \times \mathcal{Y}$ :

$$\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))} \sum_{x' \in \mathcal{X}} \xi(x') V(x', y \xi(x')) P(x' \mid x, a) \right]. \quad (2.4)$$

We now establish several useful properties for the Bellman operator  $\mathbf{T}[V]$ .

**Lemma 2.3.2** (Properties of CVaR Bellman Operator). *The Bellman operator  $\mathbf{T}[V]$  has the following properties:*

1. (Contraction.)  $\|\mathbf{T}[V_1] - \mathbf{T}[V_2]\|_\infty \leq \gamma \|V_1 - V_2\|_\infty$ , where  $\|f\|_\infty = \sup_{x \in \mathcal{X}, y \in \mathcal{Y}} |f(x, y)|$ .
2. (Concavity preserving in  $y$ .) For any  $x \in \mathcal{X}$ , suppose  $yV(x, y)$  is concave in  $y \in \mathcal{Y}$ . Then the maximization problem in (2.4) is concave. Furthermore,  $y\mathbf{T}[V](x, y)$  is concave in  $y$ .

The first property in Lemma 2.3.2 is similar to standard dynamic programming [17], and is instrumental to the design of a converging value-iteration approach. The second property is nonstandard and specific to our approach. It will be used to show that the computation of value-iteration updates involves concave, and therefore *tractable* optimization problems. Furthermore, it will be used to show that a linear-interpolation of  $V(x, y)$  in the augmented state  $y$  has a bounded error.

Equipped with the results in Theorem 2.3.1 and Lemma 2.3.2, we can now show that the fixed-point solution of  $\mathbf{T}[V](x, y) = V(x, y)$  is unique, and equals to the solution of the CVaR MDP problem (2.1) with  $x_0 = x$  and  $\alpha = y$ .

**Theorem 2.3.3** (Optimality Condition). *For any  $x \in \mathcal{X}$  and  $y \in (0, 1]$ , the solution to  $\mathbf{T}[V](x, y) = V(x, y)$  is unique, and equals to*

$$V^*(x, y) = \min_{\mu \in \Pi_H} \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0 = x, \mu \right).$$

Next, we show that the optimal value of the CVaR MDP problem (2.1) can be attained by a stationary Markov policy, defined as a greedy policy with respect to the value function  $V^*(x, y)$ . Thus, while the original problem is defined over the intractable space of history-dependent policies, a stationary Markov policy (over the augmented state space) is optimal, and can be readily derived from  $V^*(x, y)$ . Furthermore, an optimal history-dependent policy can be readily obtained from an (augmented) optimal Markov policy according to the following theorem.

**Theorem 2.3.4** (Optimal Policies). *Let  $\pi_H^* = \{\mu_0, \mu_1, \dots\} \in \Pi_H$  be a history-dependent policy recursively defined as:*

$$\mu_k(h_k) = u^*(x_k, y_k), \quad \forall k \geq 0, \quad (2.5)$$

*with initial conditions  $x_0$  and  $y_0 = \alpha$ , and state transitions*

$$x_k \sim P(\cdot \mid x_{k-1}, u^*(x_{k-1}, y_{k-1})), \quad y_k = y_{k-1} \xi_{x_{k-1}, y_{k-1}, u^*}^*(x_k), \quad \forall k \geq 1, \quad (2.6)$$

*where the stationary Markovian policy  $u^*(x, y)$  and risk factor  $\xi_{x, y, u^*}^*(\cdot)$  are solution to the min-max optimization problem in the CVaR Bellman operator  $\mathbf{T}[V^*](x, y)$ . Then,  $\pi_H^*$  is an optimal policy for problem (2.1) with initial state  $x_0$  and CVaR confidence level  $\alpha$ .*

Theorems 2.3.3 and 2.3.4 suggest that a value-iteration method [17] can be used to solve the CVaR MDP problem (2.1). Let an initial value-function guess  $V_0 : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  be chosen arbitrarily. Value iteration

proceeds recursively as follows:

$$V_{k+1}(x, y) = \mathbf{T}[V_k](x, y), \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, k \in \{0, 1, \dots\}. \quad (2.7)$$

Specifically, by combining the contraction property in Lemma 2.3.2 and uniqueness result of fixed point solutions from Theorem 2.3.3, one concludes that  $\lim_{k \rightarrow \infty} V_k(x, y) = V^*(x, y)$ . By selecting  $x = x_0$  and  $y = \alpha$ , one immediately obtains

$$V^*(x_0, \alpha) = \min_{\mu \in \Pi_H} \text{CVaR}_\alpha \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0, \mu \right).$$

Furthermore, an optimal policy may be derived from  $V^*(x, y)$  according to the policy construction procedure given in Theorem 2.3.4.

Unfortunately, while value iteration is conceptually appealing, its direct implementation in our setting is generally impractical since, chiefly, the state variable  $y$  is continuous. In the following, we pursue an *approximation* to the value iteration algorithm (2.7), based on a linear interpolation scheme for  $y$ .

## 2.4 Value Iteration with Linear Interpolation

In this section we present an approximate dynamic programming algorithm for solving CVaR MDPs, based on the theoretical results of Section 2.3. The value iteration algorithm in Eq. (2.7) presents two main implementation challenges. The first is due to the fact that the augmented state  $y$  is continuous. We handle this challenge by using interpolation, and exploit the concavity of  $yV(x, y)$  to bound the error introduced by this procedure. The second challenge stems from the fact that applying  $\mathbf{T}$  involves maximizing over  $\xi$ . Our strategy is to exploit the concavity of the maximization problem to guarantee that such optimization can indeed be performed effectively.

As discussed, our approach relies on the fact that the Bellman operator  $\mathbf{T}$  preserves concavity as established in Lemma 2.3.2. Accordingly, we require the following assumption for the initial guess  $V_0(x, y)$ ,

**Assumption 2.4.1.** *The guess for the initial value function  $V_0(x, y)$  satisfies the following properties: 1)  $yV_0(x, y)$  is concave in  $y \in \mathcal{Y}$  and 2)  $V_0(x, y)$  is continuous and bounded in  $y \in \mathcal{Y}$  for any  $x \in \mathcal{X}$ .*

Assumption 2.4.1 may easily be satisfied, for example, by choosing  $V_0(x, y) = \text{CVaR}_y(Z \mid x_0 = x)$ , where  $Z$  is any arbitrary bounded random variable. As stated earlier, a key difficulty in applying value iteration (2.7) is that, for each state  $x \in \mathcal{X}$ , the Bellman operator has to be calculated for each  $y \in \mathcal{Y}$ , and  $\mathcal{Y}$  is continuous. As an approximation, we propose to calculate the Bellman operator only for a finite set of values  $y$ , and interpolate the value function in between such interpolation points.

Formally, let  $N(x)$  denote the number of interpolation points. For every  $x \in \mathcal{X}$ , denote by  $\mathbf{Y}(x) = \{y_1, \dots, y_{N(x)}\} \in [0, 1]^{N(x)}$  the set of interpolation points. We denote by  $\mathcal{I}_x[V](y)$  the linear interpolation

**Algorithm 1** CVaR Value Iteration with Linear Interpolation1: **Given:**

- $N(x)$  interpolation points  $\mathbf{Y}(x) = \{y_1, \dots, y_{N(x)}\} \in [0, 1]^{N(x)}$  for every  $x \in \mathcal{X}$  with  $y_i < y_{i+1}$ ,  $y_1 = 0$  and  $y_{N(x)} = 1$ .
- Initial value function  $V_0(x, y)$  that satisfies Assumption 2.4.1.

2: For  $t = 1, 2, \dots$ 

- For each  $x \in \mathcal{X}$  and each  $y_i \in \mathbf{Y}(x)$ , update the value function estimate as follows:

$$V_t(x, y_i) = \mathbf{T}_{\mathcal{I}}[V_{t-1}](x, y_i),$$

3: Set the converged value iteration estimate as  $\widehat{V}^*(x, y_i)$ , for any  $x \in \mathcal{X}$ , and  $y_i \in \mathbf{Y}(x)$ .

of the function  $yV(x, y)$  on these points, i.e.,

$$\mathcal{I}_x[V](y) = y_i V(x, y_i) + \frac{y_{i+1} V(x, y_{i+1}) - y_i V(x, y_i)}{y_{i+1} - y_i} (y - y_i),$$

where  $y_i = \max\{y' \in \mathbf{Y}(x) : y' \leq y\}$  and  $y_{i+1}$  is the closest interpolation point such that  $y \in [y_i, y_{i+1}]$ . The interpolation of  $yV(x, y)$  instead of  $V(x, y)$  is key to our approach. The motivation is twofold: first, it can be shown [112] that for a discrete random variable  $Z$ ,  $y \text{CVaR}_y(Z)$  is piecewise linear in  $y$ . Second, one can show that the Lipschitzness of  $yV(x, y)$  is preserved during value iteration, and exploit this fact to bound the linear interpolation error.

We now define the *interpolated* Bellman operator  $\mathbf{T}_{\mathcal{I}}$  as follows:

$$\mathbf{T}_{\mathcal{I}}[V](x, y) = \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x'|x, a) \right]. \quad (2.8)$$

**Remark 2.4.2.** Notice that by L'Hospital's rule one has  $\lim_{y \rightarrow 0} \mathcal{I}_x[V](y\xi(x))/y = V(x, 0)\xi(x)$ . This implies that at  $y = 0$  the interpolated Bellman operator is equivalent to the original Bellman operator, i.e.,  $\mathbf{T}[V](x, 0) = \min_{a \in \mathcal{A}} \{C(x, a) + \gamma \max_{x' \in \mathcal{X}: P(x'|x, a) > 0} V(x', 0)\} = \mathbf{T}_{\mathcal{I}}[V](x, 0)$ .

Algorithm 1 presents CVaR value iteration with linear interpolation. The only difference between this algorithm and standard value iteration (2.7) is the linear interpolation procedure described above. In the following, we show that Algorithm 1 converges, and bound the error due to interpolation. We begin by showing that the useful properties established in Lemma 2.3.2 for the Bellman operator  $\mathbf{T}$  extend to the interpolated Bellman operator  $\mathbf{T}_{\mathcal{I}}$ .

**Lemma 2.4.3** (Properties of interpolated Bellman operator).  $\mathbf{T}_{\mathcal{I}}[V]$  has the same properties of  $\mathbf{T}[V]$  as in Lemma 2.3.2, namely 1) contraction and 2) concavity preservation.

Lemma 2.4.3 implies several important consequences for Algorithm 1. The first one is that the maximization problem in (2.8) is concave, and thus may be solved efficiently at each step. This guarantees that the algorithm is *tractable*. Second, the contraction property in Lemma 2.4.3 guarantees that Algorithm 1 converges, i.e., there exists a value function  $\widehat{V}^* \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$  such that  $\lim_{n \rightarrow \infty} \mathbf{T}_{\mathcal{I}}^n[V_0](x, y_i) = \widehat{V}^*(x, y_i)$ . In addition, the convergence rate is geometric and equals  $\gamma$ .

The following theorem provides an error bound between approximate value iteration and exact value iteration (2.1) in terms of the interpolation resolution.

**Theorem 2.4.4** (Convergence and Error Bound). *Suppose the initial value function  $V_0(x, y)$  satisfies Assumption 2.4.1 and let  $\epsilon > 0$  be an error tolerance parameter. For any state  $x \in \mathcal{X}$  and step  $t \geq 0$ , choose  $y_2 > 0$  such that  $V_t(x, y_2) - V_t(x, 0) \geq -\epsilon$  and update the interpolation points according to the logarithmic rule:  $y_{i+1} = \theta y_i$ ,  $\forall i \geq 2$ , with uniform constant  $\theta \geq 1$ . Then, Algorithm 1 has the following error bound:*

$$0 \geq \widehat{V}^*(x_0, \alpha) - \min_{\mu \in \Pi_H} \text{CVaR}_{\alpha} \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0, \mu \right) \geq \frac{-\gamma}{1-\gamma} \mathbf{O}((\theta - 1) + \epsilon),$$

and the following finite time convergence error bound:

$$\left| \mathbf{T}_{\mathcal{I}}^n[V_0](x_0, \alpha) - \min_{\mu \in \Pi_H} \text{CVaR}_{\alpha} \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0, \mu \right) \right| \leq \frac{\mathbf{O}((\theta - 1) + \epsilon)}{1-\gamma} + \mathbf{O}(\gamma^n).$$

Theorem 2.4.4 shows that 1) the interpolation-based value function is a *conservative estimate* for the optimal solution to problem (2.1); 2) the interpolation procedure is *consistent*, i.e., when the number of interpolation points is arbitrarily large (specifically,  $\epsilon \rightarrow 0$  and  $y_{i+1}/y_i \rightarrow 1$ ), the approximation error tends to zero; and 3) the approximation error bound is  $\mathbf{O}((\theta - 1) + \epsilon)$ , where  $\log \theta$  is the *log-difference* of the interpolation points, i.e.,  $\log \theta = \log y_{i+1} - \log y_i$ ,  $\forall i$ . In the above theorem, the big-O notation implies that there exists a real number  $M > 0$  such that the error bound  $\widehat{V}^*(x_0, \alpha) - \min_{\mu \in \Pi_H} \text{CVaR}_{\alpha} \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0, \mu \right)$  is lower-bounded by  $-\gamma M((\theta - 1) + \epsilon)/(1 - \gamma)$ .

For a pre-specified  $\epsilon$ , the condition  $V_t(x, y_2) - V_t(x, 0) \geq -\epsilon$  may be satisfied by a simple *adaptive procedure* for selecting the interpolation points  $\mathbf{Y}(x)$ . At each iteration  $t > 0$ , after calculating  $V_t(x, y_i)$  in Algorithm 1, at each state  $x$  in which the condition does not hold, add a new interpolation point  $y'_2 = \epsilon y_2 / |V_t(x, y_2) - V_t(x, 0)|$ , and additional points between  $y'_2$  and  $y_2$  such that the condition  $\log \theta \geq \log y_{i+1} - \log y_i$  is maintained. Since all the additional points belong to the segment  $[0, y_2]$ , the linearly interpolated  $V_t(x, y_i)$  remains unchanged, and Algorithm 1 proceeds as is. For bounded costs and  $\epsilon > 0$ , the number of additional points required is bounded.

The full proof of Theorem 2.4.4 is detailed in the appendix; we highlight here the main ideas and challenges involved. In the first part of the proof we bound, for all  $t > 0$ , the Lipschitz constant of  $yV_t(x, y)$  in  $y$ . The key to this result is to show that the Bellman operator  $\mathbf{T}$  preserves the Lipschitz property for  $yV_t(x, y)$ . Using the Lipschitz bound and the concavity of  $yV_t(x, y)$ , we then bound the error  $\frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y)$  for all  $y$ . The condition on  $y_2$  is required for this bound to hold when  $y \rightarrow 0$ . Finally, we use this result to bound  $\|\mathbf{T}_{\mathcal{I}}[V_t](x, y) - \mathbf{T}[V_t](x, y)\|_{\infty}$ . The results of Theorem 2.4.4 follow from contraction arguments, similar to approximate dynamic programming [17].



## 2.5 CVaR $Q$ -learning with Linear Interpolation

The value iteration algorithm in Section 2.4 assumes that the transition probabilities for the underlying MDP are known, which is oftentimes not the case. Accordingly, in this section we present a sampling-based  $Q$ -learning counterpart for the value-iteration algorithm in Section 2.4, which approximates the solution to the CVaR MDP problem. Before getting into the main results, we begin by introducing the state-action value function ( $Q$ -function) for CVaR MDP,

$$Q^*(x, y, a) = \min_{\mu \in \Pi_H} \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0 = x, a_0 = a, \mu \right),$$

which can be interpreted as the CVaR cost of starting at state  $x \in \mathcal{X}$ , using control action  $a \in \mathcal{A}$  in the first stage, and using an optimal policy thereafter. The  $Q$  function is the unique fixed point of the state-action Bellman operator  $\mathbf{F}$ , defined as

$$\mathbf{F}[Q](x, y, a) = C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \xi(x') V(y\xi(x')) P(x'|x, a),$$

where

$$V(x, y) = \min_{a \in \mathcal{A}} Q(x, y, a), \quad y \in \mathbf{Y}(x), \quad x \in \mathcal{X}.$$

We now define the state-action *interpolated* Bellman operator

$$\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x'|x, a),$$

and the corresponding interpolated value iteration update:

$$Q(x, y, a) := C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x'|x, a).$$

Let  $\hat{Q}^*(x, y, a)$  denote the fixed point of  $\mathbf{F}_{\mathcal{I}}$ , i.e., the unique solution of  $\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$ ,  $\forall x \in \mathcal{X}, y \in \mathbf{Y}(x), a \in \mathcal{A}$ , where the existence and uniqueness of the solution follows from contraction arguments similar to the ones for the state value function  $\hat{V}^*$ . The value  $\hat{Q}^*(x, y, a)$  can be interpreted as the approximate CVaR cost of starting at state  $x \in \mathcal{X}$ , using control action  $a \in \mathcal{A}$  in the first stage, and using a near-optimal policy (modulo the CVaR value function interpolation error) thereafter. Without loss of generality, in this section we assume that the set of CVaR-level interpolation points  $\mathbf{Y}(x)$  is uniform at any state  $x \in \mathcal{X}$ , i.e.,  $\mathbf{Y}(x) = \mathbf{Y}$ . Notice that  $\mathbf{Y}$  is a user-defined finite set of discretized interpolation points.

We consider both synchronous and asynchronous versions of  $Q$ -learning for CVaR MDP. In the synchronous version, the  $Q$ -function estimates of all state-action pairs are updated at each step. In contrast, in the asynchronous version, only the  $Q$ -function estimate of a sampled state-action pair is updated. Under mild assumptions, we show that both algorithms are asymptotically optimal. While the convergence rate of

synchronous  $Q$ -learning is higher [67], asynchronous  $Q$ -learning is more computationally efficient.

### 2.5.1 Synchronous CVaR $Q$ -learning

Similar to the CVaR approximate value-iteration algorithm in Algorithm 1, in CVaR  $Q$ -value iteration (or, generally, in CVaR  $Q$ -learning), one must repeatedly solve the following inner optimization problem:

$$\max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_k](y\xi(x'))}{y} P(x'|x, a).$$

When the transition probability  $P$  is unknown, one cannot simply apply the solution methods in Section 2.4 to solve such optimization problem. To tackle this issue, following the insights in Chapter 5 of [132] (or in Section 3.4 of [147]), we adopt a sample average approximation (SAA) approach. Specifically, at each state-action pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$ , given that we've seen  $N_k$  transitions  $\{x'^1, \dots, x'^{N_k}\} \sim P(x'|x, a)$ , we calculate the empirical transition probability  $P_{N_k}(x'|x, a)$  via the following equation:

$$P_{N_k}(x'|x, a) = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbf{1}\{x'^i = x' \mid x, a\}, \quad \forall x, x' \in \mathcal{X}, a \in \mathcal{A},$$

and replace the aforementioned inner optimization problem with the following SAA inner optimization problem:

$$\max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P_{N_k}(\cdot|x, a))} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\mathcal{I}_{x'^i}[V_k](y\xi(x'^i))}{y}.$$

As shown in [12], the solution to this optimization problem is *consistent*, i.e., it converges to the solution of the original (unsampled) inner optimization problem as  $N_k \rightarrow \infty$ . Details on the consistency property can be found in Chapter 5 of [132].

Equipped with the above SAA analysis, we now turn to the main algorithm for synchronous CVaR  $Q$ -learning. Suppose  $Q_0(x, y, a)$  is an initial  $Q$ -function estimate such that  $Q_0(x, y, a) = 0$  for any  $x \in \mathcal{X}$ ,  $y \in \mathbf{Y}$ ,  $a \in \mathcal{A}$ . At iteration  $k \in \{0, 1, \dots\}$ , the synchronous  $Q$ -learning algorithm samples  $N_k \geq 1$  states  $(x'^1, \dots, x'^{N_k})$  from each state  $x$  and action  $a$ , and updates the  $Q$ -function estimates for each state-action pair  $(x, y, a) \in \mathcal{X} \times \mathbf{Y} \times \mathcal{A}$  as follows:

$$Q_{k+1}(x, y, a) = Q_k(x, y, a) + \zeta_k(x, y, a) \cdot \left( -Q_k(x, y, a) + C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P_{N_k}(\cdot|x, a))} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\mathcal{I}_{x'^i}[V_k](y\xi(x'^i))}{y} \right), \quad (2.9)$$

where the value function is  $V_k(x, y) = \min_{a \in \mathcal{A}} Q_k(x, y, a)$ , and the step size  $\zeta_k(x, y, a)$  satisfies

$$\sum_k \zeta_k(x, y, a) = \infty, \quad \sum_k \zeta_k^2(x, y, a) < \infty. \quad (2.10)$$

Asymptotic convergence for synchronous CVaR  $Q$ -learning is provided by the following theorem.

**Theorem 2.5.1** (Convergence of Synchronous  $Q$ -learning). *Suppose the step-size  $\zeta_k(x, y, a)$  follows the update rule in (2.10) and the sample size  $N_k$  tends to infinity at  $k \rightarrow \infty$ . Then the sequence of estimates  $\{Q_k(x, y, a)\}_{k \in \mathbb{N}}$  computed via synchronous  $Q$ -learning (by iterative scheme (2.9)) converges to the fixed point solution  $\hat{Q}^*(x, y, a)$  component-wise with probability 1.*

After the  $Q$ -function converges, a near-optimal policy can be computed as

$$\tilde{\mu}^*(x, y) \in \arg \min_{a \in \mathcal{A}} Q_{\bar{k}}(x, y, a), \quad \forall x \in \mathcal{X}, \quad \forall y \in \mathbf{Y}, \quad (2.11)$$

where  $\bar{k}$  is the iteration index when the learning is stopped.

## 2.5.2 Asynchronous CVaR $Q$ -learning

Suppose  $Q_0(x, y, a)$  is an initial  $Q$ -function estimate such that  $Q_0(x, y, a) = 0$  for any  $x \in \mathcal{X}$ ,  $y \in \mathbf{Y}$ ,  $a \in \mathcal{A}$ . At iteration  $k \in \{0, 1, \dots\}$ , let  $x_k \in \mathcal{X}$ ,  $a_k \in \mathcal{A}$  denote the current state and action to be updated. The asynchronous  $Q$ -learning algorithm proceeds as follows:

1. Sample  $N_k \geq 1$  states  $(x'^1, \dots, x'^{N_k})$  from state  $x$  and action  $a$ ;
2. For every  $y \in \mathbf{Y}$ , update the  $Q$ -function estimate as follows:
  - for  $x = x_k$  and  $a = a_k$ , the  $Q$ -function estimate is updated according to Equation (2.9),
  - for all other states and actions the  $Q$ -function estimate is set equal to its previous value, i.e.,  $Q_{k+1}(x, y, a) = Q_k(x, y, a)$ .
3. Select state and action  $x_{k+1} \in \mathcal{X}$ ,  $a_{k+1} \in \mathcal{A}$  to update in next iteration.

For step (1), note that when  $N_k \geq 1$  one requires a generative simulator to obtain additional transition state samples. In order to enhance sampling efficiency, for each individual state-action pair one may implement a buffer that stores transition state samples generated from previous  $Q$ -learning iterations.

**Theorem 2.5.2** (Convergence of Asynchronous  $Q$ -learning). *Suppose the step-size  $\zeta_k(x, y, a)$  follows the update rule in (2.10) and the sample size  $N_k$  tends to infinity at  $k \rightarrow \infty$ . Also, suppose each state action pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$  is visited infinitely often. Then, the sequence of estimates  $\{Q_k(x, y, a)\}_{k \in \mathbb{N}}$  computed via asynchronous  $Q$ -learning converges to the fixed point solution  $Q^*(x, y, a)$  component-wise with probability 1.*

Note that the convergence result relies on the assumption that each state-action pair  $(x, a) \in \mathcal{X} \times \mathcal{A}$  is visited infinitely often. This is a standard assumption in the  $Q$ -learning literature [19], and can be satisfied by an  $\epsilon$ -greedy exploration policy that selects next states according to the following scheme:

- Select  $y_k \in \mathbf{Y}$  uniformly;
- For state  $x_k$ , select action  $a_k$  according to

$$a_k \in \begin{cases} \operatorname{argmin}_{a \in \mathcal{A}} Q_k(x_k, y_k, a) & \text{w.p. } 1 - \epsilon \\ \text{uniformly drawn from } \mathcal{A} & \text{otherwise} \end{cases}, \quad (2.12)$$

where  $\epsilon \in (0, 1)$  controls the degree of exploration;

- Select  $x_{k+1}$  by sampling  $x_{k+1} \sim P(\cdot | x_k, a_k)$ .

We remark that by following analogous arguments as in [139], the above result can be proven under milder assumptions by using PAC analysis. We refer the interested reader to the aforementioned references for more details. Similarly to synchronous  $Q$ -learning, a near optimal policy can be computed using (2.11) after the  $Q$ -functions converge.

## 2.6 Extension to Mean-CVaR MDP

In many cases besides minimizing worst case cost, one also aims to balance between cost and risk. In particular for many applications in financial engineering, [3, 64], the objective is to minimize the expected cost while controlling the CVaR. Analogous to the well-known mean-variance formulation proposed by Markowitz [84] in portfolio optimization, we hereby propose the mean-CVaR MDP problem, where the users can specify the level of risk aversion via tuning the regularization constant. While mean-CVaR MDP resembles similar trade-off between expected cost and variability as the mean-variance MDP, unlike mean-variance MDP problem that is NP-hard, we hereby show that similar techniques derived for the CVaR MDP problem can be extended to this formulation, thus it allows us to derive tractable solution algorithms for the mean-CVaR MDP problem.

Recall that  $C(x_t, a_t)$  is the stage-wise costs observed along a state/control trajectory in the MDP model, and  $\sum_{t=0}^T \gamma^t C(x_t, a_t)$  is the total discounted cost up to time  $T$ . For the set of all history dependent policies  $\Pi_H$ , the risk-sensitive discounted-cost problem we wish to address is as follows:

$$\min_{\mu \in \Pi_H} (1 - \lambda) \mathbb{E} \left[ \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \middle| x_0, \mu \right] + \lambda \text{CVaR}_\alpha \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \middle| x_0, \mu \right), \quad (2.13)$$

where  $\mu = \{\mu_0, \mu_1, \dots\}$  is the policy sequence with actions  $a_t = \mu_t(h_t)$  for  $t \in \{0, 1, \dots\}$  and  $\lambda \in [0, 1]$  is the regularizer that specifies the degree of risk aversion. We refer to formulation in (2.13) as the mean-CVaR MDP problem. When  $\lambda = 0$ , problem (2.13) coincides with the conventional risk-neutral MDP problem, and when  $\lambda = 1$ , it becomes the CVaR MDP problem depicted in (2.1). Now instead of solving the mean-CVaR

MDP problem directly, consider the following general mixed risk MDP problem:

$$\min_{\mu \in \Pi_H} \rho_{\delta, \beta} \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \middle| x_0, \mu \right), \quad (2.14)$$

where  $\rho_{\delta, \beta}$  is a coherent risk with envelop

$$\mathcal{U}_{\text{Mix}}(\delta, \beta, \mathbb{P}) = \left\{ \xi : \xi(\omega) \in [\delta, 1/\beta], \int_{\omega \in \Omega} \xi(\omega) \mathbb{P}(\omega) d\omega = 1 \right\}$$

and  $\delta \in [0, 1], \beta \in [0, 1]$ . According to Example 6.16 in [132], the mixed risk MDP problem can be reduced to the mean-CVaR MDP problem by setting the constants as

$$\delta = 1 - \lambda, \quad \beta = \left( \frac{\lambda}{\alpha} + 1 - \lambda \right)^{-1}.$$

Thus in order to solve for an optimal policy of the mean-CVaR MDP problem, in the rest of this section we aim to derive a dynamic programming algorithm for the mixed risk MDP problem by extending the techniques derived in Section 2.3 and set the confidence levels as  $(\delta, \beta) = (1 - \lambda, (\lambda/\alpha + 1 - \lambda)^{-1})$ .

### 2.6.1 Bellman Equation

By leveraging the result from Section 2.3, in this section we present a dynamic programming (DP) formulation for the mixed risk MDP problem in (2.14). As we shall see, the value function in this formulation depends on both the state  $x_t$  and the risk confidence levels  $(\delta, \beta)$ . Here the first risk confidence level keeps track of the penalty constant  $1 - \lambda$ , and the second risk confidence level keeps track of the CVaR level. To start with, consider the following recursive decomposition result for the mixed risk, which is an extension to the CVaR recursive decomposition result in Theorem 2.3.1. The proof of this theorem is given in the appendix.

**Theorem 2.6.1.** *For any  $t \geq 0$ , denote by  $Z = (Z_{t+1}, Z_{t+2}, \dots)$  the cost sequence from time  $t + 1$  onwards. The conditional mixed risk under policy  $\mu$ , i.e.,  $\rho_{\delta, \beta}(Z \mid h_t, \mu)$ , obeys the following decomposition:*

$$\rho_{\delta, \beta}(Z \mid h_t, \mu) = \max_{\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot \mid x_t, a_t))} \mathbb{E}[\xi(x_{t+1}) \cdot \rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \mu) \mid h_t, \mu],$$

where  $a_t$  is the action induced by policy  $\mu_t(h_t)$ , and the expectation is with respect to  $x_{t+1}$ .

Note that in the recursive decomposition in Theorem 2.6.1 the right-hand side involves mixed risk terms with different confidence levels than that in the left-hand side. Similar to the value function for CVaR MDPs, in order to account for the updates of risk confidence levels, we augment the state space  $\mathcal{X}$  with additional continuous two dimensional state space  $\mathcal{Y} \times \mathcal{Z} = [0, 1]^2$ , which corresponds to the confidence levels. Thus for any  $x \in \mathcal{X}$  and  $(y, z) \in \mathcal{Y} \times \mathcal{Z}$ , the *value-function*  $V(x, y, z)$  for the augmented state  $(x, y, z)$  is defined

as:

$$V(x, y, z) = \min_{\mu \in \Pi_H} \rho_{y,z} \left( \lim_{T \rightarrow \infty} \sum_{t=0}^T \gamma^t C(x_t, a_t) \mid x_0 = x, \mu \right).$$

This further leads to the following definition of mixed risk Bellman operator  $\mathbf{T} : \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \rightarrow \mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$ :

$$\mathbf{T}[V](x, y, z) = \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{Mix}}(y, z, P(\cdot | x, a))} \sum_{x' \in \mathcal{X}} \xi(x') V(x', y/\xi(x'), z\xi(x')) P(x' | x, a) \right].$$

Similar to the Bellman operator for CVaR MDP, this Bellman operator  $\mathbf{T}[V]$  is a contraction mapping in the infinity norm  $\|\cdot\|_\infty$ , and for any given non-negative constant  $M$ ,  $z\mathbf{T}[V](x, M/z, z)$  satisfies the concave preserving property in  $z$  for every  $x \in \mathcal{X}$ . Thus enumerating this Bellman operator only requires solving a concave inner maximization problem, which can be computed effectively using the interior point algorithm [37].

Equipped with these results, we can also show that the fixed point solution of  $\mathbf{T}[V](x, y, z) = V(x, y, z)$  is unique, and equals to the solution of the mixed risk MDP problem (2.14) with initial state  $x_0 = x$  and confidence levels  $(y_0, z_0) = (y, z)$ . Based on this convergence result, a history-dependent policy  $\pi_H^* = \{\mu_0, \mu_1, \dots\} \in \Pi_H$  can be readily obtained from the optimal Markov policy according to the recursive scheme:

$$\mu_k(h_k) = u^*(x_k, y_k, z_k), \quad \forall k \geq 0, \quad (2.15)$$

with initial conditions  $x_0$  and confidence levels  $(y_0, z_0) = (\delta, \beta)$ , and state transitions

$$\begin{aligned} x_k &\sim P(\cdot \mid x_{k-1}, u^*(x_{k-1}, y_{k-1}, z_{k-1})), \\ y_k &= y_{k-1} / \xi_{x_{k-1}, y_{k-1}, z_{k-1}, u^*}^*(x_k), \\ z_k &= z_{k-1} \xi_{x_{k-1}, y_{k-1}, z_{k-1}, u^*}^*(x_k), \quad \forall k \geq 1, \end{aligned} \quad (2.16)$$

where the stationary Markovian policy  $u^*(x, y, z)$  and risk factor  $\xi_{x,y,z,u^*}^*(\cdot)$  are solution to the min-max optimization problem in the mixed risk Bellman operator  $\mathbf{T}[V^*](x, y, z)$ .

The above analysis suggests that a value-iteration DP method [17] can be used to solve the mixed risk MDP problem (2.14). Unfortunately, similar to the case for CVaR MDP, its direct implementation is impractical due to the uncountable state space  $\mathcal{Y}$ . To alleviate this issue, one resorts to the approximate dynamic programming approach where the Bellman operator  $\mathbf{T}[V]$  is approximated by the 2D interpolated Bellman operator  $\mathbf{T}_{\mathcal{I}}[V]$ . A similar approach can be found in Section 2.4 for CVaR interpolated value iteration. Furthermore, similar to CVaR  $Q$ -learning with linear interpolation in Section 2.5, when the size of state space  $\mathcal{X}$  is large, one can extend the interpolated value iteration approach to  $Q$ -learning. Notice that such extensions follow immediately from the same arguments from CVaR MDP, thus the details are omitted for the sake of brevity.

## 2.7 Experiments

In this section we illustrate the performance of the CVaR MDP algorithms by studying the following 2D motion planning experiment, where states represent grid points on a 2D terrain map. An agent (e.g., a robotic vehicle) starts in a safe region and its objective is to travel to a given destination. At each time step the agent can move to any of its four neighboring states. Due to sensing and control noise, however, with probability  $\delta$  a move to a random neighboring state occurs. The stage-wise cost of each move until reaching the destination is 1, to account for fuel usage. In between the starting point and the destination there are a number of obstacles that the agent should avoid. Hitting an obstacle costs  $M \gg 1$  and terminates the mission. The objective is to compute a *safe* (i.e., obstacle-free) path that is *fuel efficient*.

For the experimental details, we choose a  $64 \times 53$  grid-world (see Figure 2.1), with a total of 3,312 states. The destination is at position  $(60, 2)$ , and there are 80 obstacles plotted in yellow. By leveraging Theorem 2.4.4, we use 21 log-spaced interpolation points for Algorithm 1 in order to achieve a small value function error. We choose  $\delta = 0.05$ , and a discount factor  $\gamma = 0.95$  for an effective horizon of 20 steps. Furthermore, we set the penalty cost equal to  $M = 2/(1 - \gamma)$ —such choice trades off high penalty for collisions and computational complexity (that increases as  $M$  increases).

In Figure 2.1 we plot the value function  $V(x, y)$  for three different values of the CVaR confidence parameter  $\alpha$ , and the corresponding paths starting from the initial position  $(60, 50)$ . The first three figures in Figure 2.1 show how by decreasing the confidence parameter  $\alpha$  the average travel distance (and hence fuel consumption) slightly increases but the collision probability decreases, as expected. We next discuss robustness to modeling errors. We conducted simulations in which with probability 0.5 each obstacle position is perturbed in a random direction to one of the neighboring grid cells. This emulates, for example, measurement errors in the terrain map. We then trained both the risk-averse ( $\alpha = 0.11$ ) and risk-neutral ( $\alpha = 1$ ) policies on the nominal (i.e., unperturbed) terrain map, and evaluated them on 400 perturbed scenarios (20 perturbed maps with 20 Monte Carlo evaluations each). While the risk-neutral policy finds a shorter route (with average cost equal to 18.137 on successful runs), it is vulnerable to perturbations and fails more often (with over 120 failed runs). In contrast, the risk-averse policy chooses slightly longer routes (with average cost equal to 18.878 on successful runs), but is much more robust to model perturbations (with only 5 failed runs).

For the computation of Algorithm 1 we represented the concave piecewise linear maximization problem in (2.8) as a linear program, and concatenated several problems to reduce repeated overhead stemming from the initialization of the CPLEX linear programming solver. This resulted in a computation time on the order of two hours. We believe there is ample room for improvement, for example by leveraging parallelization and sampling-based methods. Overall, we believe our proposed approach is currently the most practical method available for solving CVaR MDPs (as a comparison, the recently proposed method in [60] involves infinite dimensional optimization).

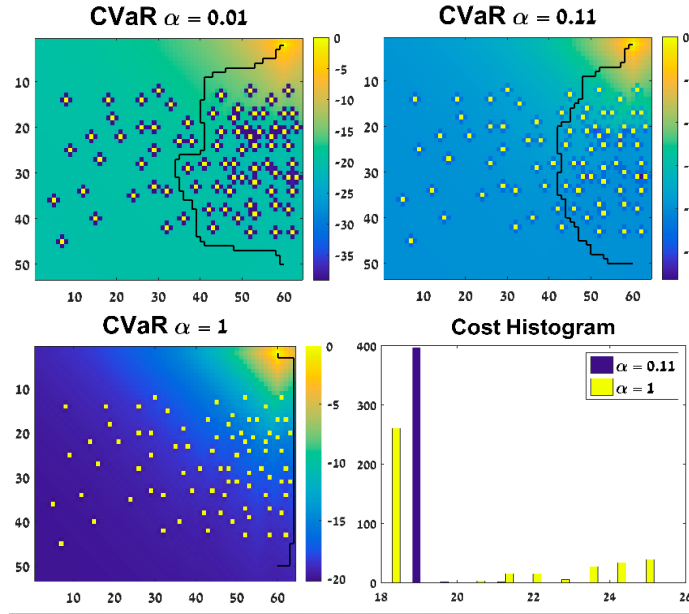


Figure 2.1: Grid-world simulation. First three plots show the value functions and corresponding paths for different CVaR confidence levels. The last plot shows a cost histogram (for 400 Monte Carlo trials) for a risk-neutral policy and a CVaR policy with confidence level  $\alpha = 0.11$ .

## 2.8 Conclusion

In this chapter we presented an algorithm for CVaR MDPs, based on approximate value-iteration on an augmented state space. We established convergence of our algorithm, and derived finite-time error bounds. These bounds are useful to stop the algorithm at a desired error threshold. In addition, we uncovered an interesting relationship between the CVaR of the total cost and the worst-case expected cost under adversarial model perturbations. In this formulation, the perturbations are correlated in time, and lead to a robustness framework significantly less conservative than the popular robust-MDP framework, where the uncertainty is temporally independent. Collectively, our work suggests CVaR MDPs as a unifying and practical framework for computing control policies that are robust with respect to both stochasticity and model perturbations.

In order to extend the aforementioned techniques to other engineering applications such as robotics, in the next chapter we will investigate another class of risk-sensitive planning problems for which the objective function is given by an expected cumulative cost, and the associated constraint function is modeled by a percentile risk.



## Chapter 3

# Risk-Constrained Reinforcement Learning with Percentile Risk

### 3.1 Introduction

#### 3.1.1 Risk Sensitive Reinforcement Learning

In many applications one is interested in taking into account risk, i.e., increased awareness of events of small probability and high consequences. Accordingly, in *risk-sensitive* MDPs the objective is to minimize a risk-sensitive criterion. In order to quantify costs that might be encountered in the tail of a cost distribution, one often considers *Value-at-risk* (VaR) and *conditional value-at-risk* (CVaR). Specifically, for continuous cost distributions,  $\text{VaR}_\alpha$  measures risk as the maximum cost that might be incurred with respect to a given confidence level  $\alpha$ , and is appealing for its intuitive meaning and its relationship to chance-constraints. This risk metric is particularly useful when there is a well-defined failure state, e.g., a state that leads a robot to collide with an obstacle. A  $\text{VaR}_\alpha$  constraint is often referred to as a chance (probability) constraint, especially in the engineering literature, and we will use this terminology in the remainder of the chapter. In contrast,  $\text{CVaR}_\alpha$  measures risk as the expected cost given that such cost is greater than or equal to  $\text{VaR}_\alpha$ , and provides a number of theoretical and computational advantages. CVaR optimization was first developed by Rockafellar and Uryasev [112] and its numerical effectiveness has been demonstrated in several portfolio optimization and option hedging problems. Risk-sensitive MDPs with a conditional value at risk metric were considered in [30, 96, 10], and a mean-average-value-at-risk problem has been solved in [9] for minimizing risk in financial markets.

The aforementioned works focus on the derivation of exact solutions, and the resulting algorithms are only applicable to relatively small problems. This has recently motivated the application of reinforcement learning (RL) methods to risk-sensitive MDPs to address large-scale problems. We will refer to such problems

as risk-sensitive RL. Reinforcement learning [19, 142] can be viewed as a class of sampling-based methods for solving MDPs. Popular reinforcement learning techniques include policy gradient [158, 83, 11] and actor-critic methods [143, 69, 99, 34, 27, 25], whereby policies are parameterized in terms of a parameter vector and policy search is performed via gradient flow approaches. One effective way to estimate gradients in RL problems is by simultaneous perturbation stochastic approximation (SPSA) [138]. Risk-sensitive RL with expected exponential utility has been considered in [32, 33]. More recently, the works in [145, 106] present RL algorithms for several variance-related risk measures, while the works in [88, 146, 101] consider CVaR-based formulations, and the works in [144, 133] consider nested CVaR-based formulations.

### 3.1.2 Chapter Contribution

Despite the rather large literature on risk-sensitive MDPs and RL, *risk-constrained* formulations have largely gone unaddressed, with only a few exceptions, e.g., [48, 36]. Yet constrained formulations naturally arise in several domains, including engineering, finance, and logistics, and provide a principled approach to address multi-objective problems. The objective of this chapter is to fill this gap, by devising policy gradient and actor-critic algorithms for risk-constrained MDPs where risk is represented via a constraint on the conditional value-at-risk (CVaR) of the cumulative cost, or as a chance-constraint. Specifically, the contribution is fourfold.

1. We formulate two risk-constrained MDP problems. The first one involves a CVaR constraint and the second one involves a chance (probability) constraint. For the CVaR-constrained optimization problem, we consider both discrete and continuous cost distributions. By re-writing the problems using a Lagrangian formulation, we derive for both problems a Bellman optimality condition with respect to an augmented MDP.
2. We devise a trajectory-based policy gradient algorithm for both CVaR-constrained and chance-constrained MDPs. The key novelty of this algorithm lies in an unbiased gradient estimation procedure under Monte Carlo sampling. Using an ordinary differential equation (ODE) approach, we establish convergence of the algorithm to locally optimal policies.
3. Using the aforementioned Bellman optimality condition, we derive several actor-critic algorithms to optimize policy and value function approximation parameters in an online fashion. As for the trajectory-based policy gradient algorithm, we show that the proposed actor-critic algorithms converge to locally optimal solutions.
4. We demonstrate the effectiveness of our algorithms in an optimal stopping problem as well as in a realistic personalized ad recommendation problem (see [51] for more details). For the latter problem, we empirically show that our CVaR-constrained RL algorithms successfully guarantee that the worst-case revenue is lower-bounded by the pre-specified company yearly target.

### 3.1.3 Chapter Organization

The rest of the chapter is structured as follows. In Section 3.2 we introduce our notation and rigorously state the problem we wish to address, namely risk-constrained RL. The next two sections provide various RL methods to approximately compute (locally) optimal policies for CVaR constrained MDPs. A trajectory-based policy gradient algorithm is presented in Section 3.3 and its convergence analysis is provided in Appendix 7.2 (Appendix 7.2.1 provides the gradient estimates of the CVaR parameter, the policy parameter, and the Lagrange multiplier, and Appendix 7.2.2 gives their convergence proofs). Actor-critic algorithms are presented in Section 3.4 and their convergence analysis is provided in Appendix 7.3 (Appendix 7.3.1 derives the gradient of the Lagrange multiplier as a function of the state-action value function, Appendix 7.3.2.1 analyzes the convergence of the critic, and Appendix 7.3.2.2 provides the multi-timescale convergence results of the CVaR parameter, the policy parameter, and the Lagrange multiplier). Section 3.5 generalizes the above policy gradient and actor-critic methods to the chance-constrained case. Empirical evaluation of our algorithms is the subject of Section 3.6.

## 3.2 Preliminaries

We begin by defining some notation that is used throughout this chapter, as well as defining the problem addressed herein and stating some basic assumptions.

### 3.2.1 Notations

We consider decision-making problems modeled as a finite MDP (an MDP with finite state and action spaces). A finite MDP is a tuple  $(\mathcal{X}, \mathcal{A}, C, D, P, P_0)$  where  $\mathcal{X} = \{1, \dots, n, x_{\text{Tar}}\}$  and  $\mathcal{A} = \{1, \dots, m\}$  are the state and action spaces,  $x_{\text{Tar}}$  is a recurrent target state, and for a state  $x$  and an action  $a$ ,  $C(x, a)$  is a cost function with  $|C(x, a)| \leq C_{\max}$ ,  $D(x, a)$  is a constraint cost function with  $|D(x, a)| \leq D_{\max}$ <sup>1</sup>,  $P(\cdot|x, a)$  is the transition probability distribution, and  $P_0(\cdot)$  is the initial state distribution. For simplicity, in this paper we assume  $P_0 = \mathbf{1}\{x = x^0\}$  for some given initial state  $x^0 \in \{1, \dots, n\}$ . Generalizations to non-atomic initial state distributions are straightforward, for which the details are omitted for the sake of brevity. A *stationary policy*  $\mu(\cdot|x)$  for an MDP is a probability distribution over actions, conditioned on the current state. In policy gradient methods, such policies are parameterized by a  $\kappa$ -dimensional vector  $\theta$ , so the space of policies can be written as  $\{\mu(\cdot|x; \theta), x \in \mathcal{X}, \theta \in \Theta \subseteq R^\kappa\}$ . Since in this setting a policy  $\mu$  is uniquely defined by its parameter vector  $\theta$ , policy-dependent functions can be written as a function of  $\mu$  or  $\theta$ , and we use  $\mu$  and  $\theta$  interchangeably in the paper.

Given a fixed  $\gamma \in (0, 1)$ , we denote by  $d_\gamma^\mu(x|x^0) = (1 - \gamma) \sum_{k=0}^{\infty} \gamma^k \mathbb{P}(x_k = x | x_0 = x^0; \mu)$  and  $\pi_\gamma^\mu(x, a|x^0) = d_\gamma^\mu(x|x^0) \mu(a|x)$ , the  $\gamma$ -discounted occupation measure of state  $x$  and state-action pair  $(x, a)$  under policy  $\mu$ , respectively. This occupation measure is a  $\gamma$ -discounted probability distribution for visiting

<sup>1</sup>Without loss of generality, we set the cost function  $C(x, a)$  and constraint cost function  $D(x, a)$  to zero when  $x = x_{\text{Tar}}$ .

each state and action pair, and it plays an important role in sampling states and actions from the real system in policy gradient and actor-critic algorithms, and in guaranteeing their convergence. Because the state and action spaces are finite, Theorem 3.1 in [4] shows that the occupation measure  $d_\gamma^\mu(x|x^0)$  is a well-defined probability distribution. On the other hand, when  $\gamma = 1$  the occupation measure of state  $x$  and state-action pair  $(x, a)$  under policy  $\mu$  are respectively defined by  $d^\mu(x|x^0) = \sum_{t=0}^{\infty} \mathbb{P}(x_t = x|x^0; \mu)$  and  $\pi^\mu(x, a|x^0) = d^\mu(x|x^0)\mu(a|x)$ . In this case the occupation measures characterize the total sums of visiting probabilities (although they are not in general probability distributions themselves) of state  $x$  and state-action pair  $(x, a)$ . To study the well-posedness of the occupation measure, we define the following notion of a transient MDP.

**Definition 3.2.1.** Define  $\mathcal{X}' = \mathcal{X} \setminus \{x_{Tar}\} = \{1, \dots, n\}$  as a state space of transient states. An MDP is said to be transient if,

1.  $\sum_{k=0}^{\infty} \mathbb{P}(x_k = x|x^0, \mu) < \infty$  for every  $x \in \mathcal{X}'$  and every stationary policy  $\mu$ ,
2.  $P(x_{Tar}|x_{Tar}, a) = 1$  for every admissible control action  $a \in \mathcal{A}$ .

Furthermore let  $T_{\mu,x}$  be the first-hitting time of the target state  $x_{Tar}$  from an arbitrary initial state  $x \in \mathcal{X}$  in the Markov chain induced by transition probability  $P(\cdot|x, a)$  and policy  $\mu$ . Although transience implies the first-hitting time is square integrable and finite almost surely, we will make the stronger assumption (which implies transience) on the uniform boundedness of the first-hitting time.

**Assumption 3.2.2.** The first-hitting time  $T_{\mu,x}$  is bounded almost surely over all stationary policies  $\mu$  and all initial states  $x \in \mathcal{X}$ . We will refer to this upper bound as  $T$ , i.e.,  $T_{\mu,x} \leq T$  almost surely.

The above assumption can be justified by the fact that sample trajectories collected in most reinforcement learning algorithms (including policy gradient and actor-critic methods) consist of a finite stopping time (also known as a time-out). If nothing else, such a bound ensures that the computation time is not unbounded. Note that although a bounded stopping time would seem to conflict with the time-stationarity of the transition probabilities, this can be resolved by augmenting the state space with a time-counter state, analogous to the arguments given in Section 4.7 in [17].

Finally, we define the constraint and cost functions. Let  $Z$  be a finite-mean ( $\mathbb{E}[|Z|] < \infty$ ) random variable representing cost, with the cumulative distribution function  $F_Z(z) = \mathbb{P}(Z \leq z)$  (e.g., one may think of  $Z$  as the total cost of an investment strategy  $\mu$ ). We define the *value-at-risk* at confidence level  $\alpha \in (0, 1)$  as

$$\text{VaR}_\alpha(Z) = \min \{z \mid F_Z(z) \geq \alpha\}.$$

Here the minimum is attained because  $F_Z$  is non-decreasing and right-continuous in  $z$ . When  $F_Z$  is continuous and strictly increasing,  $\text{VaR}_\alpha(Z)$  is the unique  $z$  satisfying  $F_Z(z) = \alpha$ . As mentioned, we refer to a constraint on the VaR as a chance constraint.

Although VaR is a popular risk measure, it is not a *coherent* risk measure [7] and does not quantify the costs that might be suffered beyond its value in the  $\alpha$ -tail of the distribution [112], [113]. In many *financial*

*applications* such as portfolio optimization where the probability of undesirable events could be small but the cost incurred could still be significant, besides describing risk as the probability of incurring costs, it will be more interesting to study the cost in the tail of the risk distribution. In this case, an alternative measure that addresses most of the VaR's shortcomings is the *conditional value-at-risk*, defined as [112]

$$\text{CVaR}_\alpha(Z) := \min_{\nu \in \mathbb{R}} \left\{ \nu + \frac{1}{1-\alpha} \mathbb{E}[(Z - \nu)^+] \right\}, \quad (3.1)$$

where  $(x)^+ = \max(x, 0)$  represents the positive part of  $x$ . Although this definition is somewhat opaque, CVaR can be thought of as the average of the worst-case  $\alpha$ -fraction of losses. Define  $H_\alpha(Z, \nu) := \nu + \frac{1}{1-\alpha} \mathbb{E}[(Z - \nu)^+]$ , so that  $\text{CVaR}_\alpha(Z) = \min_{\nu \in \mathbb{R}} H_\alpha(Z, \nu)$ .

We define the parameter  $\gamma \in (0, 1]$  as the *discounting factor* for the cost and constraint cost functions. When  $\gamma < 1$ , we are aiming to solve the MDP problem with more focus on optimizing current costs over future costs. For a policy  $\mu$ , we define the cost of a state  $x$  (state-action pair  $(x, a)$ ) as the sum of (discounted) costs encountered by the decision-maker when it starts at state  $x$  (state-action pair  $(x, a)$ ) and then follows policy  $\mu$ , i.e.,

$$\mathcal{C}^\theta(x) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k) \mid x_0 = x, \mu(\cdot|\cdot, \theta), \quad \mathcal{D}^\theta(x) = \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k) \mid x_0 = x, \mu(\cdot|\cdot, \theta),$$

and

$$\begin{aligned} \mathcal{C}^\theta(x, a) &= \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k) \mid x_0 = x, a_0 = a, \mu(\cdot|\cdot, \theta), \\ \mathcal{D}^\theta(x, a) &= \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k) \mid x_0 = x, a_0 = a, \mu(\cdot|\cdot, \theta). \end{aligned}$$

The expected values of the random variables  $\mathcal{C}^\theta(x)$  and  $\mathcal{C}^\theta(x, a)$  are known as the value and action-value functions of policy  $\mu$ , and are denoted by

$$V^\theta(x) = \mathbb{E}[\mathcal{C}^\theta(x)], \quad Q^\theta(x, a) = \mathbb{E}[\mathcal{C}^\theta(x, a)].$$

### 3.2.2 Problem Statement

The goal for standard discounted MDPs is to find an optimal policy that solves

$$\theta^* = \underset{\theta}{\operatorname{argmin}} V^\theta(x^0).$$

For *CVaR-constrained* optimization in MDPs, we consider the discounted cost optimization problem with

$\gamma \in (0, 1)$ , i.e., for a given confidence level  $\alpha \in (0, 1)$  and cost tolerance  $\beta \in \mathbb{R}$ ,

$$\min_{\theta} V^{\theta}(x^0) \quad \text{subject to} \quad \text{CVaR}_{\alpha}(\mathcal{D}^{\theta}(x^0)) \leq \beta. \quad (3.2)$$

Using the definition of  $H_{\alpha}(Z, \nu)$ , one can reformulate (3.2) as:

$$\min_{\theta, \nu} V^{\theta}(x^0) \quad \text{subject to} \quad H_{\alpha}(\mathcal{D}^{\theta}(x^0), \nu) \leq \beta. \quad (3.3)$$

It is shown in [112] and [113] that the optimal  $\nu$  actually equals  $\text{VaR}_{\alpha}$ , so we refer to this parameter as the VaR parameter. Here we choose to analyze the discounted-cost CVaR-constrained optimization problem, i.e., with  $\gamma \in (0, 1)$ , as in many financial and marketing applications where CVaR constraints are used, it is more intuitive to put more emphasis on current costs rather than on future costs. The analysis can be easily generalized for the case where  $\gamma = 1$ .

For *chance-constrained* optimization in MDPs, we consider the stopping cost optimization problem with  $\gamma = 1$ , i.e., for a given confidence level  $\beta \in (0, 1)$  and cost tolerance  $\alpha \in \mathbb{R}$ ,

$$\min_{\theta} V^{\theta}(x^0) \quad \text{subject to} \quad \mathbb{P}(\mathcal{D}^{\theta}(x^0) \geq \alpha) \leq \beta. \quad (3.4)$$

Here we choose  $\gamma = 1$  because in many engineering applications, where chance constraints are used to ensure overall safety, there is no notion of discounting since future threats are often as important as the current one. Similarly, the analysis can be easily extended to the case where  $\gamma \in (0, 1)$ .

There are a number of mild technical and notational assumptions which we will make throughout the paper, so we state them here:

**Assumption 3.2.3** (Differentiability). *For any state-action pair  $(x, a)$ ,  $\mu(a|x; \theta)$  is continuously differentiable in  $\theta$  and  $\nabla_{\theta} \mu(a|x; \theta)$  is a Lipschitz function in  $\theta$  for every  $a \in \mathcal{A}$  and  $x \in \mathcal{X}$ .<sup>2</sup>*

**Assumption 3.2.4** (Strict Feasibility). *There exists a transient policy  $\mu(\cdot|x; \theta)$  such that  $H_{\alpha}(\mathcal{D}^{\theta}(x^0), \nu) < \beta$  in the CVaR-constrained optimization problem, and  $\mathbb{P}(\mathcal{D}^{\theta}(x^0) \geq \alpha) < \beta$  in the chance-constrained problem.*

Note that Assumption 3.2.3 imposes smoothness on the optimal policy. Assumption 3.2.4 guarantees the existence of a locally optimal policy for the CVaR-constrained optimization problem via the Lagrangian analysis introduced in the next subsection.

In the remainder of the paper we first focus on studying stochastic approximation algorithms for the CVaR-constrained optimization problem (Sections 3.3 and 3.4) and then adapt the results to the chance-constrained optimization problem in Section 3.5. Our solution approach relies on a Lagrangian relaxation procedure, which is discussed next.

---

<sup>2</sup>In actor-critic algorithms, the assumption on continuous differentiability holds for the augmented state Markovian policies  $\mu(a|x, s; \theta)$ .

### 3.2.3 Lagrangian Approach and Reformulation

To solve (3.3), we employ a Lagrangian relaxation procedure [16], which leads to the unconstrained problem:

$$\max_{\lambda \geq 0} \min_{\theta, \nu} \left( L(\nu, \theta, \lambda) := V^\theta(x^0) + \lambda \left( H_\alpha(\mathcal{D}^\theta(x^0), \nu) - \beta \right) \right), \quad (3.5)$$

where  $\lambda$  is the Lagrange multiplier. Notice that  $L(\nu, \theta, \lambda)$  is a linear function in  $\lambda$  and  $H_\alpha(\mathcal{D}^\theta(x^0), \nu)$  is a continuous function in  $\nu$ . Corollary 4 in [152] implies the existence of a locally optimal policy  $\theta^*$  for the CVaR-constrained optimization problem, which corresponds to the existence of the local saddle point  $(\nu^*, \theta^*, \lambda^*)$  for the minimax optimization problem  $\max_{\lambda \geq 0} \min_{\theta, \nu} L(\nu, \theta, \lambda)$ , defined as follows.

**Definition 3.2.5.** A local saddle point of  $L(\nu, \theta, \lambda)$  is a point  $(\nu^*, \theta^*, \lambda^*)$  such that for some  $r > 0$ ,  $\forall (\theta, \nu) \in \Theta \times \left[ -\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma} \right] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r)$  and  $\forall \lambda \geq 0$ , we have

$$L(\nu, \theta, \lambda^*) \geq L(\nu^*, \theta^*, \lambda^*) \geq L(\nu^*, \theta^*, \lambda), \quad (3.6)$$

where  $\mathcal{B}_{(\theta^*, \nu^*)}(r)$  is a hyper-dimensional ball centered at  $(\theta^*, \nu^*)$  with radius  $r > 0$ .

In [96, 10] it is shown that there exists a *deterministic history-dependent* optimal policy for CVaR-constrained optimization. The important point is that this policy does not depend on the complete history, but only on the current time step  $k$ , current state of the system  $x_k$ , and accumulated discounted constraint cost  $\sum_{i=0}^k \gamma^i D(x_i, a_i)$ .

In the following two sections, we present a policy gradient (PG) algorithm (Section 3.3) and several actor-critic (AC) algorithms (Section 3.4) to optimize (3.5) (and hence find a locally optimal solution to problem (3.3)). While the PG algorithm updates its parameters after observing several trajectories, the AC algorithms are incremental and update their parameters at each time-step.

## 3.3 A Trajectory-based Policy Gradient Algorithm

In this section, we present a policy gradient algorithm to solve the optimization problem (3.5). The idea of the algorithm is to descend in  $(\theta, \nu)$  and ascend in  $\lambda$  using the gradients of  $L(\nu, \theta, \lambda)$  w.r.t.  $\theta$ ,  $\nu$ , and  $\lambda$ , i.e.,<sup>3</sup>

$$\nabla_\theta L(\nu, \theta, \lambda) = \nabla_\theta V^\theta(x^0) + \frac{\lambda}{(1-\alpha)} \nabla_\theta \mathbb{E} \left[ (\mathcal{D}^\theta(x^0) - \nu)^+ \right], \quad (3.7)$$

$$\partial_\nu L(\nu, \theta, \lambda) = \lambda \left( 1 + \frac{1}{(1-\alpha)} \partial_\nu \mathbb{E} \left[ (\mathcal{D}^\theta(x^0) - \nu)^+ \right] \right) \ni \lambda \left( 1 - \frac{1}{(1-\alpha)} \mathbb{P}(\mathcal{D}^\theta(x^0) \geq \nu) \right), \quad (3.8)$$

$$\nabla_\lambda L(\nu, \theta, \lambda) = \nu + \frac{1}{(1-\alpha)} \mathbb{E} \left[ (\mathcal{D}^\theta(x^0) - \nu)^+ \right] - \beta. \quad (3.9)$$

<sup>3</sup>The notation  $\ni$  in (3.8) means that the right-most term is a member of the sub-gradient set  $\partial_\nu L(\nu, \theta, \lambda)$ .

The unit of observation in this algorithm is a system trajectory generated by following the current policy. At each iteration, the algorithm generates  $N$  trajectories by following the current policy, uses them to estimate the gradients in (3.7)–(3.9), and then uses these estimates to update the parameters  $\nu, \theta, \lambda$ .

Let  $\xi = \{x_0, a_0, c_0, x_1, a_1, c_1, \dots, x_{T-1}, a_{T-1}, c_{T-1}, x_T\}$  be a trajectory generated by following the policy  $\theta$ , where  $x_T = x_{\text{Tar}}$  is the target state of the system. The cost, constraint cost, and probability of  $\xi$  are defined as

$$\mathcal{C}(\xi) = \sum_{k=0}^{T-1} \gamma^k C(x_k, a_k), \quad \mathcal{D}(\xi) = \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k),$$

and

$$\mathbb{P}_\theta(\xi) = P_0(x_0) \prod_{k=0}^{T-1} \mu(a_k | x_k; \theta) P(x_{k+1} | x_k, a_k),$$

respectively. Based on the definition of  $\mathbb{P}_\theta(\xi)$ , one obtains  $\nabla_\theta \log \mathbb{P}_\theta(\xi) = \sum_{k=0}^{T-1} \nabla_\theta \log \mu(a_k | x_k; \theta)$ .

Algorithm 2 contains the pseudo-code of our proposed policy gradient algorithm. What appears inside the parentheses on the right-hand-side of the update equations are the estimates of the gradients of  $L(\nu, \theta, \lambda)$  w.r.t.  $\theta, \nu, \lambda$  (estimates of (3.7)–(3.9)). Gradient estimates of the Lagrangian function can be found in Appendix 7.2.1. In the algorithm,  $\Gamma_\Theta$  is an operator that projects a vector  $\theta \in \mathbb{R}^\kappa$  to the closest point in a compact and convex set  $\Theta \subset \mathbb{R}^\kappa$ , i.e.,  $\Gamma_\Theta(\theta) = \arg \min_{\hat{\theta} \in \Theta} \|\theta - \hat{\theta}\|_2^2$ ,  $\Gamma_N$  is a projection operator to  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ , i.e.,  $\Gamma_N(\nu) = \arg \min_{\hat{\nu} \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]} \|\nu - \hat{\nu}\|_2^2$ , and  $\Gamma_\Lambda$  is a projection operator to  $[0, \lambda_{\max}]$ , i.e.,  $\Gamma_\Lambda(\lambda) = \arg \min_{\hat{\lambda} \in [0, \lambda_{\max}]} \|\lambda - \hat{\lambda}\|_2^2$ . These projection operators are necessary to ensure the convergence of the algorithm. Next we introduce the following assumptions for the step-sizes of the policy gradient method in Algorithm 2.

**Assumption 3.3.1** (Step Sizes for Policy Gradient). *The step size schedules  $\{\zeta_1(k)\}$ ,  $\{\zeta_2(k)\}$ , and  $\{\zeta_3(k)\}$  satisfy*

$$\sum_k \zeta_1(k) = \sum_k \zeta_2(k) = \sum_k \zeta_3(k) = \infty, \quad (3.10)$$

$$\sum_k \zeta_1(k)^2, \quad \sum_k \zeta_2(k)^2, \quad \sum_k \zeta_3(k)^2 < \infty, \quad (3.11)$$

$$\zeta_1(k) = o(\zeta_2(k)), \quad \zeta_2(k) = o(\zeta_3(k)). \quad (3.12)$$

These step-size schedules satisfy the standard conditions for stochastic approximation algorithms, and ensure that the  $\nu$  update is on the fastest time-scale  $\{\zeta_3(k)\}$ , the policy  $\theta$  update is on the intermediate time-scale  $\{\zeta_2(k)\}$ , and the Lagrange multiplier  $\lambda$  update is on the slowest time-scale  $\{\zeta_1(k)\}$ . This results in a three time-scale stochastic approximation algorithm.

In the following theorem, we prove that our policy gradient algorithm converges to a locally optimal policy for the CVaR-constrained optimization problem.

**Theorem 3.3.2.** *Under Assumptions 3.2.2–3.3.1, the sequence of policy updates in Algorithm 2 converges*



almost surely to a locally optimal policy  $\theta^*$  for the CVaR-constrained optimization problem.

While we refer the reader to Appendix 7.2.2 for the technical details of this proof, a high level overview of the proof technique is given as follows.

1. First we show that each update of the multi-time scale discrete stochastic approximation algorithm  $(\nu_i, \theta_i, \lambda_i)$  converges almost surely, but at different speeds, to the stationary point  $(\nu^*, \theta^*, \lambda^*)$  of the corresponding continuous time system.
2. Then by using Lyapunov analysis, we show that the continuous time system is locally asymptotically stable at the stationary point  $(\nu^*, \theta^*, \lambda^*)$ .
3. Since the Lyapunov function used in the above analysis is the Lagrangian function  $L(\nu, \theta, \lambda)$ , we finally conclude that the stationary point  $(\nu^*, \theta^*, \lambda^*)$  is also a local saddle point, which implies  $\theta^*$  is the locally optimal policy for the CVaR-constrained optimization problem.

This convergence proof procedure is rather standard for stochastic approximation algorithms, see [27, 25, 106] for more details, and represents the structural backbone for the convergence analysis of the other policy gradient and actor-critic methods provided in this paper.

Notice that the difference in convergence speeds between  $\theta_i$ ,  $\nu_i$ , and  $\lambda_i$  is due to the step-size schedules. Here  $\nu$  converges faster than  $\theta$  and  $\theta$  converges faster than  $\lambda$ . This multi-time scale convergence property allows us to simplify the convergence analysis by assuming that  $\theta$  and  $\lambda$  are fixed in  $\nu$ 's convergence analysis, assuming that  $\nu$  converges to  $\nu^*(\theta)$  and  $\lambda$  is fixed in  $\theta$ 's convergence analysis, and finally assuming that  $\nu$  and  $\theta$  have already converged to  $\nu^*(\lambda)$  and  $\theta^*(\lambda)$  in  $\lambda$ 's convergence analysis. To illustrate this idea, consider the following two-time scale stochastic approximation algorithm for updating  $(x_i, y_i) \in \mathbf{X} \times \mathbf{Y}$ :

$$x_{i+1} = x_i + \zeta_1(i)(f(x_i, y_i) + M_{i+1}), \quad (3.13)$$

$$y_{i+1} = y_i + \zeta_2(i)(g(x_i, y_i) + N_{i+1}), \quad (3.14)$$

where  $f(x_i, y_i)$  and  $g(x_i, y_i)$  are Lipschitz continuous functions,  $M_{i+1}$ ,  $N_{i+1}$  are square integrable Martingale differences w.r.t. the  $\sigma$ -fields  $\sigma(x_k, y_k, M_k, k \leq i)$  and  $\sigma(x_k, y_k, N_k, k \leq i)$ , and  $\zeta_1(i)$  and  $\zeta_2(i)$  are non-summable, square summable step sizes. If  $\zeta_2(i)$  converges to zero faster than  $\zeta_1(i)$ , then (3.13) is a faster recursion than (3.14) after some iteration  $i_0$  (i.e., for  $i \geq i_0$ ), which means (3.13) has uniformly larger increments than (3.14). Since (3.14) can be written as

$$y_{i+1} = y_i + \zeta_1(i) \left( \frac{\zeta_2(i)}{\zeta_1(i)} (g(x_i, y_i) + N_{i+1}) \right),$$

and by the fact that  $\zeta_2(i)$  converges to zero faster than  $\zeta_1(i)$ , (3.13) and (3.14) can be viewed as noisy Euler discretizations of the ODEs  $\dot{x} = f(x, y)$  and  $\dot{y} = 0$ . Note that one can consider the ODE  $\dot{x} = f(x, y_0)$  in place of  $\dot{x} = f(x, y)$ , where  $y_0$  is constant, because  $\dot{y} = 0$ . One can then show (see e.g., Theorem 6.2 of Borkar 2008) the main two-timescale convergence result, i.e., under the above assumptions associated with

**Algorithm 2** Trajectory-based Policy Gradient Algorithm for CVaR Optimization

---

**Input:** parameterized policy  $\mu(\cdot|\cdot; \theta)$ , confidence level  $\alpha$ , and cost tolerance  $\beta$   
**Initialization:** policy  $\theta = \theta_0$ , VaR parameter  $\nu = \nu_0$ , and the Lagrangian parameter  $\lambda = \lambda_0$   
**while** TRUE **do**  
  **for**  $i = 0, 1, 2, \dots$  **do**  
    **for**  $j = 1, 2, \dots$  **do**  
      Generate  $N$  trajectories  $\{\xi_{j,i}\}_{j=1}^N$  by starting at  $x_0 = x^0$  and following the current policy  $\theta_i$ .  
    **end for**  
     **$\nu$  Update:**  $\nu_{i+1} = \Gamma_N \left[ \nu_i - \zeta_3(i) \left( \lambda_i - \frac{\lambda_i}{(1-\alpha)N} \sum_{j=1}^N \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$   
     **$\theta$  Update:**  $\theta_{i+1} = \Gamma_\Theta \left[ \theta_i - \zeta_2(i) \left( \frac{1}{N} \sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} \mathcal{C}(\xi_{j,i}) \right. \right.$   
       $\left. \left. + \frac{\lambda_i}{(1-\alpha)N} \sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i})|_{\theta=\theta_i} (\mathcal{D}(\xi_{j,i}) - \nu_i) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$   
     **$\lambda$  Update:**  $\lambda_{i+1} = \Gamma_\Lambda \left[ \lambda_i + \zeta_1(i) \left( \nu_i - \beta + \frac{1}{(1-\alpha)N} \sum_{j=1}^N (\mathcal{D}(\xi_{j,i}) - \nu_i) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]$   
  **end for**  
  **if**  $\{\lambda_i\}$  converges to  $\lambda_{\max}$ , i.e.,  $|\lambda_i - \lambda_{\max}| \leq \epsilon$  for some tolerance parameter  $\epsilon > 0$  **then**  
    Set  $\lambda_{\max} \leftarrow 2\lambda_{\max}$ .  
  **else**  
    **return** parameters  $\nu, \theta, \lambda$  and **break**  
  **end if**  
**end while**

---

(3.14), the sequence  $(x_i, y_i)$  converges to  $(\mu(y^*), y^*)$  as  $i \rightarrow \infty$ , with probability one, where  $\mu(y_0)$  is a globally asymptotically stable equilibrium of the ODE  $\dot{x} = f(x, y_0)$ ,  $\mu$  is a Lipschitz continuous function, and  $y^*$  is a globally asymptotically stable equilibrium of the ODE  $\dot{y} = g(\mu(y), y)$ .

### 3.4 Actor-Critic Algorithms

As mentioned in Section 3.3, the unit of observation in our policy gradient algorithm (Algorithm 2) is a system trajectory. This may result in high variance for the gradient estimates, especially when the length of the trajectories is long. To address this issue, in this section, we propose two actor-critic algorithms that approximate some quantities in the gradient estimates by linear combinations of basis functions and update the parameters (linear coefficients) incrementally (after each state-action transition). We present two actor-critic algorithms for optimizing (3.5). These algorithms are based on the gradient estimates of Sections 3.4.1-3.4.3. While the first algorithm (SPSA-based) is fully incremental and updates all the parameters  $\theta, \nu, \lambda$  at each time-step, the second one updates  $\theta$  at each time-step and updates  $\nu$  and  $\lambda$  only at the end of each

trajectory, thus is regarded as a semi-trajectory-based method. Algorithm 3 contains the pseudo-code of these algorithms. The projection operators  $\Gamma_\Theta$ ,  $\Gamma_N$ , and  $\Gamma_\Lambda$  are defined as in Section 3.3 and are necessary to ensure the convergence of the algorithms. Next, we introduce the following assumptions for the step-sizes of the actor-critic method in Algorithm 3.

**Assumption 3.4.1** (Step Sizes). *The step size schedules  $\{\zeta_1(k)\}$ ,  $\{\zeta_2(k)\}$ ,  $\{\zeta_3(k)\}$ , and  $\{\zeta_4(k)\}$  satisfy*

$$\sum_k \zeta_1(k) = \sum_k \zeta_2(k) = \sum_k \zeta_3(k) = \sum_k \zeta_4(k) = \infty, \quad (3.15)$$

$$\sum_k \zeta_1(k)^2, \sum_k \zeta_2(k)^2, \sum_k \zeta_3(k)^2, \sum_k \zeta_4(k)^2 < \infty, \quad (3.16)$$

$$\zeta_1(k) = o(\zeta_2(k)), \quad \zeta_2(k) = o(\zeta_3(k)), \quad \zeta_3(k) = o(\zeta_4(k)). \quad (3.17)$$

Furthermore, the SPSA step size  $\{\Delta_k\}$  in the actor-critic algorithm satisfies  $\Delta_k \rightarrow 0$  as  $k \rightarrow \infty$  and  $\sum_k (\zeta_2(k)/\Delta_k)^2 < \infty$ .

These step-size schedules satisfy the standard conditions for stochastic approximation algorithms, and ensure that the critic update is on the fastest time-scale  $\{\zeta_4(k)\}$ , the policy and VaR parameter updates are on the intermediate time-scale, with the  $\nu$ -update  $\{\zeta_3(k)\}$  being faster than the  $\theta$ -update  $\{\zeta_2(k)\}$ , and finally the Lagrange multiplier update is on the slowest time-scale  $\{\zeta_1(k)\}$ . This results in four time-scale stochastic approximation algorithms.

### 3.4.1 Gradient w.r.t. the Policy Parameters $\theta$

The gradient of the objective function w.r.t. the policy  $\theta$  in (3.7) may be rewritten as

$$\nabla_\theta L(\nu, \theta, \lambda) = \nabla_\theta \left( \mathbb{E}[\mathcal{C}^\theta(x^0)] + \frac{\lambda}{(1-\alpha)} \mathbb{E}[(\mathcal{D}^\theta(x^0) - \nu)^+] \right). \quad (3.24)$$

Given the original MDP  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, D, P, P_0)$  and the parameter  $\lambda$ , we define the augmented MDP  $\bar{\mathcal{M}} = (\bar{\mathcal{X}}, \bar{\mathcal{A}}, \bar{C}_\lambda, \bar{P}, \bar{P}_0)$  as  $\bar{\mathcal{X}} = \mathcal{X} \times \mathcal{S}$ ,  $\bar{\mathcal{A}} = \mathcal{A}$ ,  $\bar{P}_0(x, s) = P_0(x) \mathbf{1}\{s_0 = s\}$ , and

$$\bar{C}_\lambda(x, s, a) = \begin{cases} \lambda(-s)^+/(1-\alpha) & \text{if } x = x_{\text{Tar}}, \\ C(x, a) & \text{otherwise,} \end{cases}$$

$$\bar{P}(x', s'|x, s, a) = \begin{cases} P(x'|x, a) \mathbf{1}\{s' = (s - D(x, a))/\gamma\} & \text{if } x \in \mathcal{X}', \\ \mathbf{1}\{x' = x_{\text{Tar}}, s' = 0\} & \text{if } x = x_{\text{Tar}}, \end{cases}$$

where  $\mathcal{S}$  is the finite state space of the augmented state  $s$ ,  $s_0$  is the initial state of the augmented MDP,  $x_{\text{Tar}}$  is the target state of the original MDP  $\mathcal{M}$  and  $s_{\text{Tar}}$  is the  $s$  part of the state when a policy  $\theta$  reaches a target state  $x_{\text{Tar}}$ , which we assume occurs before an upper-bound  $T$  number of steps, i.e.,  $s_{\text{Tar}} = \frac{1}{\gamma^T} \left( \nu - \sum_{k=0}^{T-1} \gamma^k D(x_k, a_k) \right)$ , such that the initial state is given by  $s_0 = \nu$ . We will now use  $n$  to indicate the size of the *augmented* state

**Algorithm 3** Actor-Critic Algorithms for CVaR Optimization

**Input:** Parameterized policy  $\mu(\cdot|\cdot; \theta)$  and value function feature vector  $\phi(\cdot)$  (both over the augmented MDP  $\bar{\mathcal{M}}$ ), confidence level  $\alpha$ , and cost tolerance  $\beta$

**Initialization:** policy  $\theta = \theta_0$ ; VaR parameter  $\nu = \nu_0$ ; Lagrangian parameter  $\lambda = \lambda_0$ ; value function weight vector  $v = v_0$ ; initial condition  $(x_0, s_0) = (x^0, \nu)$

**while** TRUE **do**

**// (1) SPSA-based Algorithm:**

**for**  $k = 0, 1, 2, \dots$  **do**

    Draw action  $a_k \sim \mu(\cdot|x_k, s_k; \theta_k)$ ;

    Observe cost  $\bar{C}_{\lambda_k}(x_k, s_k, a_k)$ ;

    Observe next state  $(x_{k+1}, s_{k+1}) \sim \bar{P}(\cdot|x_k, s_k, a_k)$ ; *// note that  $s_{k+1} = (s_k - D(x_k, a_k))/\gamma$*

**// AC Algorithm:**

$$\text{TD Error: } \delta_k(v_k) = \bar{C}_{\lambda_k}(x_k, s_k, a_k) + \gamma v_k^\top \phi(x_{k+1}, s_{k+1}) - v_k^\top \phi(x_k, s_k) \quad (3.18)$$

$$\text{Critic Update: } v_{k+1} = v_k + \zeta_4(k) \delta_k(v_k) \phi(x_k, s_k) \quad (3.19)$$

$$\nu \text{ Update: } \nu_{k+1} = \Gamma_N \left( \nu_k - \zeta_3(k) \left( \lambda_k + \frac{v_k^\top [\phi(x^0, \nu_k + \Delta_k) - \phi(x^0, \nu_k - \Delta_k)]}{2\Delta_k} \right) \right) \quad (3.20)$$

$$\theta \text{ Update: } \theta_{k+1} = \Gamma_\Theta \left( \theta_k - \frac{\zeta_2(k)}{1-\gamma} \nabla_\theta \log \mu(a_k|x_k, s_k; \theta) \cdot \delta_k(v_k) \right) \quad (3.21)$$

$$\lambda \text{ Update: } \lambda_{k+1} = \Gamma_\Lambda \left( \lambda_k + \zeta_1(k) (\nu_k - \beta + \frac{1}{(1-\alpha)(1-\gamma)} \mathbf{1}\{x_k = x_{\text{Tar}}\} (-s_k)^+) \right) \quad (3.22)$$

**if**  $x_k = x_{\text{Tar}}$  (reach a target state), **then** set  $(x_{k+1}, s_{k+1}) = (x^0, \nu_{k+1})$

**end for**

**// (2) Semi Trajectory-based Algorithm:**

  Initialize  $t = 0$

**for**  $k = 0, 1, 2, \dots$  **do**

    Draw action  $a_k \sim \mu(\cdot|x_k, s_k; \theta_k)$ , observe cost  $\bar{C}_{\lambda_k}(x_k, s_k, a_k)$ , and next state  $(x_{k+1}, s_{k+1}) \sim \bar{P}(\cdot|x_k, s_k, a_k)$ ; Update  $(\delta_k, v_k, \theta_k, \lambda_k)$  using Eqs. 3.18, 3.19, 3.21, and 3.22

**if**  $x_k = x_{\text{Tar}}$  **then**

    Update  $\nu$  as

$$\nu \text{ Update: } \nu_{k+1} = \Gamma_N \left( \nu_k - \zeta_3(k) \left( \lambda_k - \frac{\lambda_k}{1-\alpha} \mathbf{1}\{x_k = x_{\text{Tar}}, s_k \leq 0\} \right) \right) \quad (3.23)$$

    Set  $(x_{k+1}, s_{k+1}) = (x^0, \nu_{k+1})$  and  $t = 0$

**else**

$t \leftarrow t + 1$

**end if**

**end for**

**if**  $\{\lambda_i\}$  converges to  $\lambda_{\max}$ , i.e.,  $|\lambda_{i^*} - \lambda_{\max}| \leq \epsilon$  for some tolerance parameter  $\epsilon > 0$  **then**

    Set  $\lambda_{\max} \leftarrow 2\lambda_{\max}$ .

**else**

**return** parameters  $v, w, \nu, \theta, \lambda$ , and **break**

**end if**

**end while**

space  $\bar{\mathcal{X}}$  instead of the size of the original state space  $\mathcal{X}$ . It can be later seen that the augmented state  $s$  in the MDP  $\bar{\mathcal{M}}$  keeps track of the cumulative CVaR constraint cost, and allows one to reformulate the CVaR Lagrangian problem as an MDP (with respect to  $\bar{\mathcal{M}}$ ).

We define a class of parameterized stochastic policies  $\{\mu(\cdot|x, s; \theta), (x, s) \in \bar{\mathcal{X}}, \theta \in \Theta \subseteq \mathbb{R}^{\kappa_1}\}$  for this augmented MDP. Recall that  $\mathcal{C}^\theta(x)$  is the discounted cumulative cost and  $\mathcal{D}^\theta(x)$  is the discounted cumulative constraint cost. Therefore, the total (discounted) cost of a trajectory can be written as

$$\sum_{k=0}^T \gamma^k \bar{C}_\lambda(x_k, s_k, a_k) \mid x_0 = x, s_0 = s, \mu = \mathcal{C}^\theta(x) + \frac{\lambda}{(1-\alpha)} (\mathcal{D}^\theta(x) - s)^+. \quad (3.25)$$

From (3.25), it is clear that the quantity in the parenthesis of (3.24) is the value function of the policy  $\theta$  at state  $(x^0, \nu)$  in the augmented MDP  $\bar{\mathcal{M}}$ , i.e.,  $V^\theta(x^0, \nu)$ . Thus, it is easy to show that<sup>4</sup>

$$\nabla_\theta L(\nu, \theta, \lambda) = \nabla_\theta V^\theta(x^0, \nu) = \frac{1}{1-\gamma} \sum_{x, s, a} \pi_\gamma^\theta(x, s, a \mid x^0, \nu) \nabla \log \mu(a \mid x, s; \theta) Q^\theta(x, s, a),^5 \quad (3.26)$$

where  $\pi_\gamma^\theta$  is the discounted occupation measure (defined in Section 3.2) and  $Q^\theta$  is the action-value function of policy  $\theta$  in the augmented MDP  $\bar{\mathcal{M}}$ . We can show that  $\frac{1}{1-\gamma} \nabla \log \mu(a_k \mid x_k, s_k; \theta) \cdot \delta_k$  is an unbiased estimate of  $\nabla_\theta L(\nu, \theta, \lambda)$ , where

$$\delta_k = \bar{C}_\lambda(x_k, s_k, a_k) + \gamma \hat{V}(x_{k+1}, s_{k+1}) - \hat{V}(x_k, s_k)$$

is the temporal-difference (TD) error in the MDP  $\bar{\mathcal{M}}$  from (3.18), and  $\hat{V}$  is an unbiased estimator of  $V^\theta$  (see e.g., [27]). In our actor-critic algorithms, the critic uses linear approximation for the value function  $V^\theta(x, s) \approx v^\top \phi(x, s) = \tilde{V}^{\theta, v}(x, s)$ , where the feature vector  $\phi(\cdot)$  belongs to a low-dimensional space  $\mathbb{R}^{\kappa_1}$  with dimension  $\kappa_1$ . The linear approximation  $\tilde{V}^{\theta, v}$  belongs to a low-dimensional subspace  $S_V = \{\Phi v \mid v \in \mathbb{R}^{\kappa_1}\}$ , where  $\Phi$  is the  $n \times \kappa_1$  matrix whose rows are the transposed feature vectors  $\phi^\top(\cdot)$ . To ensure that the set of feature vectors forms a well-posed linear approximation to the value function, we impose the following assumption to the basis functions.

**Assumption 3.4.2** (Independent Basis Functions). *The basis functions  $\{\phi^{(i)}\}_{i=1}^{\kappa_1}$  are linearly independent. In particular,  $\kappa_1 \leq n$  and  $\Phi$  is full column rank. Moreover, for every  $v \in \mathbb{R}^{\kappa_1}$ ,  $\Phi v \neq e$ , where  $e$  is the  $n$ -dimensional vector with all entries equal to one.*

The following theorem shows that the critic update  $v_k$  converges almost surely to  $v^*$ , the minimizer of the Bellman residual. Details of the proof can be found in Appendix 7.3.2.

**Theorem 3.4.3.** *Define  $v^* \in \arg \min_v \|B_\theta[\Phi v] - \Phi v\|_{d_\theta}^2$  as the minimizer to the Bellman residual, where*

<sup>4</sup>Note that the second equality in Equation (3.26) is the result of the policy gradient theorem [143, 99].

<sup>5</sup>Notice that the state and action spaces of the original MDP are finite, and there is only a finite number of outcomes in the transition of  $s$  (due to the assumption of a bounded first hitting time). Therefore the augmented state  $s$  belongs to a finite state space as well.

the Bellman operator is given by

$$B_\theta[V](x, s) = \sum_a \mu(a|x, s; \theta) \left\{ \bar{C}_\lambda(x, s, a) + \sum_{x', s'} \gamma \bar{P}(x', s'|x, s, a) V(x', s') \right\}$$

and  $\tilde{V}^*(x, s) = (v^*)^\top \phi(x, s)$  is the projected Bellman fixed point of  $V^\theta(x, s)$ , i.e.,  $\tilde{V}^*(x, s) = \Pi B_\theta[\tilde{V}^*](x, s)$ . Suppose the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$  is used to generate samples of  $(x_k, s_k, a_k)$  for any  $k \in \{0, 1, \dots\}$ . Then under Assumptions 3.4.1–3.4.2, the  $v$ -update in the actor-critic algorithm converges to  $v^*$  almost surely.

### 3.4.2 Gradient w.r.t. the Lagrangian Parameter $\lambda$

We may rewrite the gradient of the objective function w.r.t. the Lagrangian parameters  $\lambda$  in (3.9) as

$$\nabla_\lambda L(\nu, \theta, \lambda) = \nu - \beta + \nabla_\lambda \left( \mathbb{E}[C^\theta(x^0)] + \frac{\lambda}{(1-\alpha)} \mathbb{E}[(\mathcal{D}^\theta(x^0) - \nu)^+] \right) \stackrel{(a)}{=} \nu - \beta + \nabla_\lambda V^\theta(x^0, \nu). \quad (3.27)$$

Similar to Section 3.4.1, equality (a) comes from the fact that the quantity in parenthesis in (3.27) is  $V^\theta(x^0, \nu)$ , the value function of the policy  $\theta$  at state  $(x^0, \nu)$  in the augmented MDP  $\bar{\mathcal{M}}$ . Note that the dependence of  $V^\theta(x^0, \nu)$  on  $\lambda$  comes from the definition of the cost function  $\bar{C}_\lambda$  in  $\bar{\mathcal{M}}$ . We now derive an expression for  $\nabla_\lambda V^\theta(x^0, \nu)$ , which in turn will give us an expression for  $\nabla_\lambda L(\nu, \theta, \lambda)$ .

**Lemma 3.4.4.** *The gradient of  $V^\theta(x^0, \nu)$  w.r.t. the Lagrangian parameter  $\lambda$  may be written as*

$$\nabla_\lambda V^\theta(x^0, \nu) = \frac{1}{1-\gamma} \sum_{x, s, a} \pi_\gamma^\theta(x, s, a|x^0, \nu) \frac{1}{(1-\alpha)} \mathbf{1}\{x = x_{\text{Tar}}\} (-s)^+. \quad (3.28)$$

*Proof.* See Appendix 7.3.1. □

From Lemma 3.4.4 and (3.27), it is easy to see that  $\nu - \beta + \frac{1}{(1-\gamma)(1-\alpha)} \mathbf{1}\{x = x_{\text{Tar}}\} (-s)^+$  is an unbiased estimate of  $\nabla_\lambda L(\nu, \theta, \lambda)$ . An issue with this estimator is that its value is fixed to  $\nu_k - \beta$  all along a system trajectory, and only changes at the end to  $\nu_k - \beta + \frac{1}{(1-\gamma)(1-\alpha)} (-s_{\text{Tar}})^+$ . This may affect the incremental nature of our actor-critic algorithm. To address this issue, [46] previously proposed a different approach to estimate the gradients w.r.t.  $\theta$  and  $\lambda$  which involves another value function approximation to the constraint. However this approach is less desirable in many practical applications as it increases the approximation error and impedes the speed of convergence.

Another important issue is that the above estimator is unbiased only if the samples are generated from the distribution  $\pi_\gamma^\theta(\cdot|x^0, \nu)$ . If we just follow the policy  $\theta$ , then we may use  $\nu_k - \beta + \frac{\gamma^k}{(1-\alpha)} \mathbf{1}\{x_k = x_{\text{Tar}}\} (-s_k)^+$  as an estimate for  $\nabla_\lambda L(\nu, \theta, \lambda)$ . Note that this is an issue for all discounted actor-critic algorithms: their (likelihood ratio based) estimate for the gradient is unbiased only if the samples are generated from  $\pi_\gamma^\theta$ , and not when we simply follow the policy. This might also be the reason why, to the best of our knowledge, no

rigorous convergence analysis can be found in the literature for (likelihood ratio based) discounted actor-critic algorithms under the sampling distribution.<sup>6</sup>

### 3.4.3 Sub-Gradient w.r.t. the VaR Parameter $\nu$

We may rewrite the sub-gradient of our objective function w.r.t. the VaR parameter  $\nu$  in (3.8) as

$$\partial_\nu L(\nu, \theta, \lambda) \ni \lambda \left( 1 - \frac{1}{(1-\alpha)} \mathbb{P} \left( \sum_{k=0}^{\infty} \gamma^k D(x_k, a_k) \geq \nu \mid x_0 = x^0; \theta \right) \right). \quad (3.29)$$

From the definition of the augmented MDP  $\bar{\mathcal{M}}$ , the probability in (3.29) may be written as  $\mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu; \theta)$ , where  $s_{\text{Tar}}$  is the  $s$  part of the state in  $\bar{\mathcal{M}}$  when we reach a target state, i.e.,  $x = x_{\text{Tar}}$  (see Section 3.4.1). Thus, we may rewrite (3.29) as

$$\partial_\nu L(\nu, \theta, \lambda) \ni \lambda \left( 1 - \frac{1}{(1-\alpha)} \mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu; \theta) \right). \quad (3.30)$$

From (3.30), it is easy to see that  $\lambda - \lambda \mathbf{1}\{s_{\text{Tar}} \leq 0\}/(1-\alpha)$  is an unbiased estimate of the sub-gradient of  $L(\nu, \theta, \lambda)$  w.r.t.  $\nu$ . An issue with this (unbiased) estimator is that it can only be applied at the end of a system trajectory (i.e., when we reach the target state  $x_{\text{Tar}}$ ), and thus, using it prevents us from having a fully incremental algorithm. In fact, this is the estimator that we use in our *semi-trajectory-based* actor-critic algorithm.

One approach to estimate this sub-gradient incrementally is to use the *simultaneous perturbation stochastic approximation* (SPSA) method [26]. The idea of SPSA is to estimate the sub-gradient  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$  using two values of  $g$  at  $\nu^- = \nu - \Delta$  and  $\nu^+ = \nu + \Delta$ , where  $\Delta > 0$  is a positive perturbation (see [26, 106] for the detailed description of  $\Delta$ ).<sup>7</sup> In order to see how SPSA can help us to estimate our sub-gradient incrementally, note that

$$\partial_\nu L(\nu, \theta, \lambda) = \lambda + \partial_\nu \left( \mathbb{E}[D^\theta(x^0)] + \frac{\lambda}{(1-\alpha)} \mathbb{E}[(D^\theta(x^0) - \nu)^+] \right) \stackrel{(a)}{=} \lambda + \partial_\nu V^\theta(x^0, \nu). \quad (3.31)$$

Similar to Sections 3.4.1–3.4.3, equality (a) comes from the fact that the quantity in parenthesis in (3.31) is  $V^\theta(x^0, \nu)$ , the value function of the policy  $\theta$  at state  $(x^0, \nu)$  in the augmented MDP  $\bar{\mathcal{M}}$ . Since the critic uses a linear approximation for the value function, i.e.,  $V^\theta(x, s) \approx v^\top \phi(x, s)$ , in our actor-critic algorithms (see Section 3.4.1 and Algorithm 3), the SPSA estimate of the sub-gradient would be of the form  $g(\nu) \approx \lambda + v^\top [\phi(x^0, \nu^+) - \phi(x^0, \nu^-)]/2\Delta$ .

<sup>6</sup>Note that the discounted actor-critic algorithm with convergence proof in [24] is based on SPSA.

<sup>7</sup>SPSA-based gradient estimate was first proposed in [138] and has been widely used in various settings, especially those involving a high-dimensional parameter. The SPSA estimate described above is two-sided. It can also be implemented single-sided, where we use the values of the function at  $\nu$  and  $\nu^+$ . We refer the readers to [26] for more details on SPSA and to [106] for its application to learning in mean-variance risk-sensitive MDPs.

### 3.4.4 Convergence of Actor-Critic Methods

In this section, we will prove that the actor-critic algorithms converge to a locally optimal policy for the CVaR-constrained optimization problem. Define

$$\epsilon_\theta(v_k) = \|B_\theta[\Phi v_k] - \Phi v_k\|_\infty$$

as the residual of the value function approximation at step  $k$ , induced by policy  $\mu(\cdot|\cdot, \cdot; \theta)$ . By the triangle inequality and fixed point theorem  $B_\theta[V^*] = V^*$ , it can be easily seen that  $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k) + \|B_\theta[\Phi v_k] - B_\theta[V^*]\|_\infty \leq \epsilon_\theta(v_k) + \gamma\|\Phi v_k - V^*\|_\infty$ . The last inequality follows from the contraction property of the Bellman operator. Thus, one concludes that  $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k)/(1 - \gamma)$ . Now, we state the main theorem for the convergence of actor-critic methods.

**Theorem 3.4.5.** *Suppose  $\epsilon_{\theta_k}(v_k) \rightarrow 0$  and the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$  is used to generate samples of  $(x_k, s_k, a_k)$  for any  $k \in \{0, 1, \dots\}$ . For the SPSA-based algorithms, suppose the feature vector satisfies the technical Assumption 7.3.2 (provided in Appendix 7.3.2.2) and suppose the SPSA step-size satisfies the condition  $\epsilon_{\theta_k}(v_k) = o(\Delta_k)$ , i.e.,  $\epsilon_{\theta_k}(v_k)/\Delta_k \rightarrow 0$ . Then under Assumptions 3.2.2–3.2.4 and 3.4.1–3.4.2, the sequence of policy updates in Algorithm 3 converges almost surely to a locally optimal policy for the CVaR-constrained optimization problem.*

Details of the proof can be found in Appendix 7.3.2.

## 3.5 Extension to Chance-Constrained Optimization of MDPs

In many applications, in particular in engineering (see, for example, [94]), *chance constraints* are imposed to ensure mission success with high probability. Accordingly, in this section we extend the analysis of CVaR-constrained MDPs to chance-constrained MDPs (i.e., (3.4)). As for CVaR-constrained MDPs, we employ a Lagrangian relaxation procedure [16] to convert a chance-constrained optimization problem into the following unconstrained problem:

$$\max_{\lambda} \min_{\theta, \alpha} \left( L(\theta, \lambda) := \mathcal{C}^\theta(x^0) + \lambda \left( \mathbb{P}(\mathcal{D}^\theta(x^0) \geq \alpha) - \beta \right) \right), \quad (3.32)$$

where  $\lambda$  is the Lagrange multiplier. Recall Assumption 3.2.4 which assumed strict feasibility, i.e., there exists a transient policy  $\mu(\cdot|x; \theta)$  such that  $\mathbb{P}(\mathcal{D}^\theta(x^0) \geq \alpha) < \beta$ . This is needed to guarantee the existence of a local saddle point.



### 3.5.1 Policy Gradient Method

In this section we propose a policy gradient method for chance-constrained MDPs (similar to Algorithm 2). Since we do not need to estimate the  $\nu$ -parameter in chance-constrained optimization, the corresponding policy gradient algorithm can be simplified and at each inner loop of Algorithm 2 we only perform the following updates at the end of each trajectory:

$$\begin{aligned} \theta \text{ Update: } \theta_{i+1} &= \Gamma_{\Theta} \left[ \theta_i - \frac{\zeta_2(i)}{N} \left( \sum_{j=1}^N \nabla_{\theta} \log \mathbb{P}(\xi_{j,i}) \mathcal{C}(\xi_{j,i}) + \lambda_i \nabla_{\theta} \log \mathbb{P}(\xi_{j,i}) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \alpha\} \right) \right] \\ \lambda \text{ Update: } \lambda_{i+1} &= \Gamma_{\Lambda} \left[ \lambda_i + \zeta_1(i) \left( -\beta + \frac{1}{N} \sum_{j=1}^N \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \alpha\} \right) \right] \end{aligned}$$

Considering the multi-time-scale step-size rules in Assumption 3.3.1, the  $\theta$  update is on the fast time-scale  $\{\zeta_2(i)\}$  and the Lagrange multiplier  $\lambda$  update is on the slow time-scale  $\{\zeta_1(i)\}$ . This results in a two time-scale stochastic approximation algorithm. In the following theorem, we prove that our policy gradient algorithm converges to a locally optimal policy for the chance-constrained problem.

**Theorem 3.5.1.** *Under Assumptions 3.2.2–3.3.1, the sequence of policy updates in Algorithm 2 converges to a locally optimal policy  $\theta^*$  for the chance-constrained optimization problem almost surely.*

*Sketch.* By taking the gradient of  $L(\theta, \lambda)$  w.r.t.  $\theta$ , we have

$$\nabla_{\theta} L(\theta, \lambda) = \nabla_{\theta} \mathcal{C}^{\theta}(x^0) + \lambda \nabla_{\theta} \mathbb{P}(\mathcal{D}^{\theta}(x^0) \geq \alpha) = \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) \mathcal{C}(\xi) + \lambda \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) \mathbf{1}\{\mathcal{D}(\xi) \geq \alpha\}.$$

On the other hand, the gradient of  $L(\theta, \lambda)$  w.r.t.  $\lambda$  is given by

$$\nabla_{\lambda} L(\theta, \lambda) = \mathbb{P}(\mathcal{D}^{\theta}(x^0) \geq \alpha) - \beta.$$

One can easily verify that the  $\theta$  and  $\lambda$  updates are therefore unbiased estimates of  $\nabla_{\theta} L(\theta, \lambda)$  and  $\nabla_{\lambda} L(\theta, \lambda)$ , respectively. Then the rest of the proof follows analogously from the convergence proof of Algorithm 2 in steps 2 and 3 of Theorem 3.3.2.  $\square$

### 3.5.2 Actor-Critic Method

In this section, we present an actor-critic algorithm for the chance-constrained optimization. Given the original MDP  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, C, D, P, P_0)$  and parameter  $\lambda$ , we define the augmented MDP  $\bar{\mathcal{M}} = (\bar{\mathcal{X}}, \bar{\mathcal{A}}, \bar{C}_{\lambda}, \bar{P}, \bar{P}_0)$  as in the CVaR counterpart, except that  $\bar{P}_0(x, s) = P_0(x) \mathbf{1}\{s = \alpha\}$  and

$$\bar{C}_{\lambda}(x, s, a) = \begin{cases} \lambda \mathbf{1}\{s \leq 0\} & \text{if } x = x_{\text{Tar}}, \\ C(x, a) & \text{otherwise.} \end{cases}$$

Thus, the total cost of a trajectory can be written as

$$\sum_{k=0}^T \bar{C}_\lambda(x_k, s_k, a_k) \mid x_0 = x, s_0 = \beta, \mu = \mathcal{C}^\theta(x) + \lambda \mathbb{P}(\mathcal{D}^\theta(x) \geq \beta). \quad (3.33)$$

Unlike the actor-critic algorithms for CVaR-constrained optimization, here the value function approximation parameter  $v$ , policy  $\theta$ , and Lagrange multiplier estimate  $\lambda$  are updated episodically, i.e., after each episode ends by time  $T$  when  $(x_k, s_k) = (x_{\text{Tar}}, s_{\text{Tar}})$ <sup>8</sup>, as follows:

$$\textbf{Critic Update: } v_{k+1} = v_k + \zeta_3(k) \sum_{h=0}^T \phi(x_h, s_h) \delta_h(v_k) \quad (3.34)$$

$$\textbf{Actor Updates: } \theta_{k+1} = \Gamma_\Theta \left( \theta_k - \zeta_2(k) \sum_{h=0}^T \nabla_\theta \log \mu(a_h | x_h, s_h; \theta) |_{\theta=\theta_k} \cdot \delta_h(v_k) \right) \quad (3.35)$$

$$\lambda_{k+1} = \Gamma_\Lambda \left( \lambda_k + \zeta_1(k) (-\beta + \mathbf{1}\{s_{\text{Tar}} \leq 0\}) \right) \quad (3.36)$$

From analogous analysis as for the CVaR actor-critic method, the following theorem shows that the critic update  $v_k$  converges almost surely to  $v^*$ .

**Theorem 3.5.2.** *Let  $v^* \in \arg \min_v \|B_\theta[\Phi v] - \Phi v\|_{d^\theta}^2$  be a minimizer of the Bellman residual, where the undiscounted Bellman operator at every  $(x, s) \in \bar{\mathcal{X}}'$  is given by*

$$B_\theta[V](x, s) = \sum_{a \in \mathcal{A}} \mu(a | x, s; \theta) \left\{ \bar{C}_\lambda(x, s, a) + \sum_{(x', s') \in \bar{\mathcal{X}}'} \bar{P}(x', s' | x, s, a) V(x', s') \right\}$$

and  $\tilde{V}^*(x, s) = \phi^\top(x, s) v^*$  is the projected Bellman fixed point of  $V^\theta(x, s)$ , i.e.,  $\tilde{V}^*(x, s) = \Pi B_\theta[\tilde{V}^*](x, s)$  for  $(x, s) \in \bar{\mathcal{X}}'$ . Then under Assumptions 3.4.1–3.4.2, the  $v$ -update in the actor-critic algorithm converges to  $v^*$  almost surely.

*Sketch.* The proof of this theorem follows the same steps as those in the proof of Theorem 3.4.3, except replacing the  $\gamma$ -occupation measure  $d_\gamma^\theta$  with the occupation measure  $d^\theta$  (the total visiting probability). Similar analysis can also be found in the proof of Theorem 10 in [145]. Under Assumption 3.2.2, the occupation measure of any transient states  $x \in \mathcal{X}'$  (starting at an arbitrary initial transient state  $x_0 \in \mathcal{X}'$ ) can be written as  $d^\mu(x | x^0) = \sum_{t=0}^{T_{\mu,x}} \mathbb{P}(x_t = x | x^0; \mu)$  when  $\gamma = 1$ . This further implies the total visiting probabilities are bounded as follows:  $d^\mu(x | x^0) \leq T_{\mu,x}$  and  $\pi^\mu(x, a | x^0) \leq T_{\mu,x}$  for any  $x, x_0 \in \mathcal{X}'$ . Therefore, when the sequence of states  $\{(x_h, s_h)\}_{h=0}^T$  is sampled by the  $h$ -step transition distribution  $\mathbb{P}(x_h, s_h \mid x^0, s^0, \theta)$ ,  $\forall h \leq T$ , the unbiased estimators of

$$A := \sum_{(y, s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a' | x, s) \phi(y, s') \left( \phi^\top(y, s') - \sum_{(z, s'') \in \bar{\mathcal{X}}'} \bar{P}(z, s'' | y, s', a) \phi^\top(z, s'') \right)$$

<sup>8</sup>Note that  $s_{\text{Tar}}$  is the state of  $s_t$  when  $x_t$  hits the (recurrent) target state  $x_{\text{Tar}}$ .

and

$$b := \sum_{(y,s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a' | x, s) \phi(y, s') \bar{C}_\lambda(y, s', a')$$

are given by  $\sum_{h=0}^T \phi(x_h, s_h) (\phi^\top(x_h, s_h) - \phi^\top(x_{h+1}, s_{h+1}))$  and  $\sum_{h=0}^T \phi(x_h, s_h) \bar{C}_\lambda(x_h, s_h, a_h)$ , respectively. Note that in this theorem, we directly use the results from Theorem 7.1 in [17] to show that every eigenvalue of matrix  $A$  has positive real part, instead of using the technical result in Lemma 7.3.1.  $\square$

Recall that  $\epsilon_\theta(v_k) = \|B_\theta[\Phi v_k] - \Phi v_k\|_\infty$  is the residual of the value function approximation at step  $k$  induced by policy  $\mu(\cdot|\cdot, \cdot; \theta)$ . By the triangle inequality and fixed-point theorem of stochastic stopping problems, i.e.,  $B_\theta[V^*] = V^*$  from Theorem 3.1 in [17], it can be easily seen that  $\|V^* - \Phi v_k\|_\infty \leq \epsilon_\theta(v_k) + \|B_\theta[\Phi v_k] - B_\theta[V^*]\|_\infty \leq \epsilon_\theta(v_k) + \kappa \|\Phi v_k - V^*\|_\infty$  for some  $\kappa \in (0, 1)$ . Similar to the actor-critic algorithm for CVaR-constrained optimization, the last inequality also follows from the contraction mapping property of  $B_\theta$  from Theorem 3.2 in [17]. Now, we state the main theorem for the convergence of the actor-critic method.

**Theorem 3.5.3.** *Under Assumptions 3.2.2–3.4.2, if  $\epsilon_{\theta_k}(v_k) \rightarrow 0$ , then the sequence of policy updates converges almost surely to a locally optimal policy  $\theta^*$  for the chance-constrained optimization problem.*

*Sketch .* From Theorem 3.5.2, the critic update converges to the minimizer of the Bellman residual. Since the critic update converges on the fastest scale, as in the proof of Theorem 3.4.5, one can replace  $v_k$  by  $v^*(\theta_k)$  in the convergence proof of the actor update. Furthermore, by sampling the sequence of states  $\{(x_h, s_h)\}_{h=0}^T$  with the  $h$ -step transition distribution  $\mathbb{P}(x_h, s_h | x^0, s^0, \theta)$ ,  $\forall h \leq T$ , the unbiased estimator of the gradient of the linear approximation to the Lagrangian function is given by

$$\nabla_\theta \tilde{L}^v(\theta, \lambda) := \sum_{(x,s) \in \bar{\mathcal{X}}', a \in \mathcal{A}} \pi^\theta(x, s, a | x_0 = x^0, s_0 = v) \nabla_\theta \log \mu(a | x, s; \theta) \tilde{A}^{\theta, v}(x, s, a),$$

where  $\tilde{Q}^{\theta, v}(x, s, a) - v^\top \phi(x, s)$  is given by  $\sum_{h=0}^T \nabla_\theta \log \mu(a_h | x_h, s_h; \theta)|_{\theta=\theta_k} \cdot \delta_h(v^*)$  and the unbiased estimator of  $\nabla_\lambda L(\theta, \lambda) = -\beta + \mathbb{P}(s_{\text{Tar}} \leq 0)$  is given by  $-\beta + \mathbf{1}\{s_{\text{Tar}} \leq 0\}$ . Analogous to equation (7.52) in the proof of Theorem 7.3.5, by convexity of quadratic functions, we have for any value function approximation  $v$ ,

$$\sum_{(y,s') \in \bar{\mathcal{X}}', a' \in \mathcal{A}} \pi^\theta(y, s', a' | x, s) (A_\theta(y, s', a') - \tilde{A}_\theta^v(y, s', a')) \leq 2T \frac{\epsilon_\theta(v)}{1 - \kappa},$$

which further implies that  $\nabla_\theta L(\theta, \lambda) - \nabla_\theta \tilde{L}^v(\theta, \lambda) \rightarrow 0$  when  $\epsilon_\theta(v) \rightarrow 0$  at  $v = v^*(\theta_k)$ . The rest of the proof follows identical arguments as in steps 3 to 5 of the proof of Theorem 3.4.5.  $\square$

## 3.6 Experiments

In this section we illustrate the effectiveness of our risk-constrained policy gradient and actor-critic algorithms by testing them on an American option stopping problem and on a long-term personalized advertisement-recommendation (ad-recommendation) problem.

### 3.6.1 The Optimal Stopping Problem

We consider an optimal stopping problem in which the state at each time step  $k \leq T$  consists of the cost  $c_k$  and time  $k$ , i.e.,  $x = (c_k, k)$ , where  $T$  is the stopping time. The agent (buyer) should decide either to accept the present cost ( $u_k = 1$ ) or wait ( $u_k = 0$ ). If he/she accepts or when  $k = T$ , the system reaches a terminal state and the cost  $\max(K, c_k)$  is received ( $K$  is the maximum cost threshold), otherwise, she receives a holding cost  $p_h$  and the new state is  $(c_{k+1}, k+1)$ , where  $c_{k+1}$  is  $f_u c_k$  w.p.  $p$  and  $f_d c_k$  w.p.  $1-p$  ( $f_u > 1$  and  $f_d < 1$  are constants). Moreover, there is a discount factor  $\gamma \in (0, 1)$  to account for the increase in the buyer's affordability. Note that if we change cost to reward and minimization to maximization, this is exactly the American option pricing problem, a standard testbed to evaluate risk-sensitive algorithms (e.g., see [145]). Since the state space size  $n$  is exponential in  $T$ , finding an exact solution via dynamic programming (DP) quickly becomes infeasible, and thus the problem requires approximation and sampling techniques.

The optimal stopping problem can be reformulated as follows

$$\min_{\theta} \mathbb{E} [\mathcal{C}^{\theta}(x^0)] \quad \text{subject to} \quad \text{CVaR}_{\alpha}(\mathcal{C}^{\theta}(x^0)) \leq \beta \quad \text{or} \quad \mathbb{P}(\mathcal{C}^{\theta}(x^0) \geq \alpha) \leq \beta, \quad (3.37)$$

where the discounted cost and constraint cost functions are identical ( $\mathcal{C}^{\theta}(x) = \mathcal{D}^{\theta}(x)$ ) and are both given by  $\mathcal{C}^{\theta}(x) = \sum_{k=0}^T \gamma^k (\mathbf{1}\{u_k = 1\} \max(K, c_k) + \mathbf{1}\{u_k = 0\} p_h) \mid x_0 = x, \mu$ . We set the parameters of the MDP as follows:  $x_0 = [1; 0]$ ,  $p_h = 0.1$ ,  $T = 20$ ,  $K = 5$ ,  $\gamma = 0.95$ ,  $f_u = 2$ ,  $f_d = 0.5$ , and  $p = 0.65$ . The confidence interval and constraint threshold are given by  $\alpha = 0.95$  and  $\beta = 3$ . The number of sample trajectories  $N$  is set to 500,000 and the parameter bounds are  $\lambda_{\max} = 5,000$  and  $\Theta = [-20, 20]^{\kappa_1}$ , where the dimension of the basis functions is  $\kappa_1 = 1024$ . We implement radial basis functions (RBFs) as feature functions and search over the class of Boltzmann policies  $\left\{ \theta : \theta = \{\theta_{x,a}\}_{x \in \mathcal{X}, a \in \mathcal{A}}, \mu_{\theta}(a|x) = \frac{\exp(\theta_{x,a}^{\top} x)}{\sum_{a \in \mathcal{A}} \exp(\theta_{x,a}^{\top} x)} \right\}$ .

We consider the following trajectory-based algorithms:

1. **PG:** This is a policy gradient algorithm that minimizes the expected discounted cost function without considering any risk criteria.
2. **PG-CVaR/PG-CC:** These are the CVaR/chance-constrained simulated trajectory-based policy gradient algorithms given in Section 3.3.

The experiments for each algorithm comprise the following two phases:

1. **Tuning phase:** We run the algorithm and update the policy until  $(\nu, \theta, \lambda)$  converges.
2. **Converged run:** Having obtained a converged policy  $\theta^*$  in the tuning phase, in the converged run phase, we perform a Monte Carlo simulation of 10,000 trajectories and report the results as averages over these trials.

We also consider the following incremental algorithms:

1. **AC:** This is an actor-critic algorithm that minimizes the expected discounted cost function without considering any risk criteria. This is similar to Algorithm 1 in [24].
2. **AC-CVaR/AC-VaR:** These are the CVaR/chance-constrained semi-trajectory actor-critic algorithms given in Section 3.4.
3. **AC-CVaR-SPSA:** This is the CVaR-constrained SPSA actor-critic algorithm given in Section 3.4.

Similar to the trajectory-based algorithms, we use RBF features for  $[x; s]$  and consider the family of augmented state Boltzmann policies. Similarly, the experiments comprise two phases: 1) the tuning phase, where the set of parameters  $(v, \nu, \theta, \lambda)$  is obtained after the algorithm converges, and 2) the converged run, where the policy is simulated with 10,000 trajectories.

We compare the performance of PG-CVaR and PG-CC (given in Algorithm 2), and AC-CVaR-SPSA, AC-CVaR, and AC-VaR (given in Algorithm 3), with PG and AC, their risk-neutral counterparts. Figures 3.1 and 3.2 show the distribution of the discounted cumulative cost  $\mathcal{C}^\theta(x^0)$  for the policy  $\theta$  learned by each of these algorithms. The results indicate that the risk-constrained algorithms yield a higher expected cost, but less worst-case variability, compared to the risk-neutral methods. More precisely, the cost distributions of the risk-constrained algorithms have lower right-tail (worst-case) distribution than their risk-neutral counterparts. Table 3.1 summarizes the performance of these algorithms. The numbers reiterate what we concluded from Figures 3.1 and 3.2.

Notice that while the risk averse policy satisfies the CVaR constraint, it is not tight (i.e., the constraint is not matched). In fact this is a problem of local optimality, and other experiments in the literature (for example see the numerical results in [106] and in [25]) have the same problem of producing solutions which obey the constraints but not tightly. However, since both the expectation and CVaR risk metrics are sub-additive and convex, one can always construct a policy that is a linear combination of the risk neutral optimal policy and the risk averse policy, such that it matches the constraint threshold and has a lower cost compared to the risk averse policy.

	$\mathbb{E}(\mathcal{C}^\theta(x^0))$	$\sigma(\mathcal{C}^\theta(x^0))$	$\text{CVaR}(\mathcal{C}^\theta(x^0))$	$\text{VaR}(\mathcal{C}^\theta(x^0))$
PG	1.177	1.065	4.464	4.005
PG-CVaR	1.997	0.060	2.000	2.000
PG-CC	1.994	0.121	2.058	2.000
AC	1.113	0.607	3.331	3.220
AC-CVaR-SPSA	1.326	0.322	2.145	1.283
AC-CVaR	1.343	0.346	2.208	1.290
AC-VaR	1.817	0.753	4.006	2.300

Table 3.1: Performance comparison of the policies learned by the risk-constrained and risk-neutral algorithms. In this table  $\sigma(\mathcal{C}^\theta(x^0))$  stands for the standard deviation of the total cost.

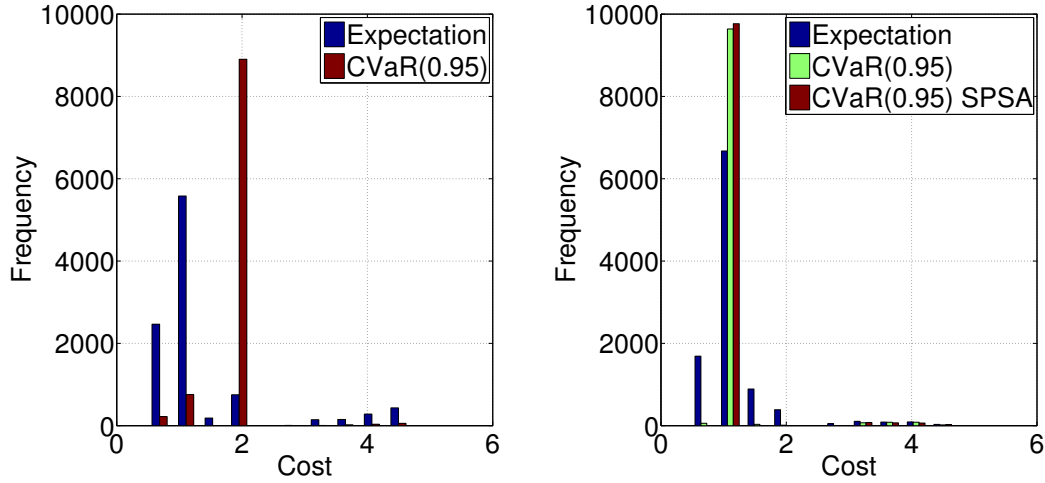


Figure 3.1: Cost distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

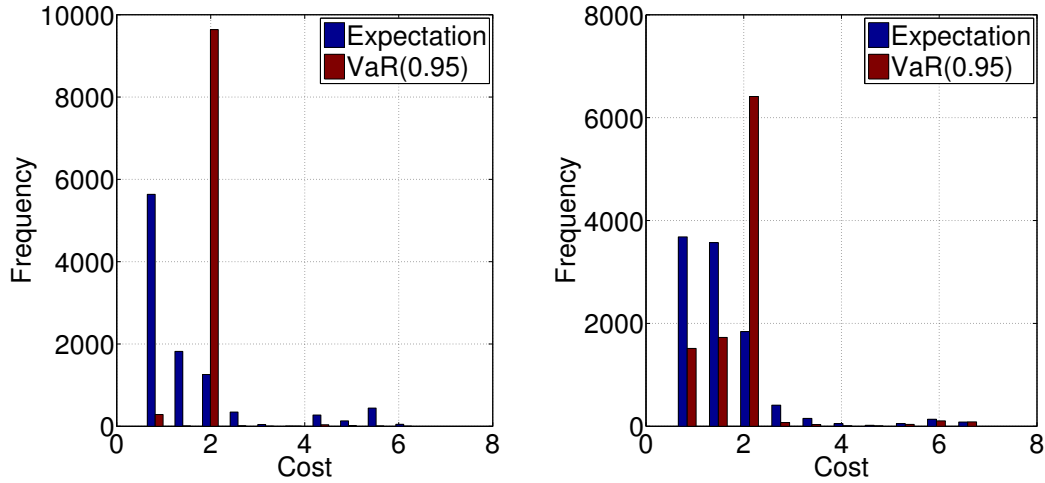


Figure 3.2: Cost distributions for the policies learned by the chance-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

### 3.6.2 A Personalized Ad-Recommendation System

Many companies such as banks and retailers use user-specific targeting of advertisements to attract more customers and increase their revenue. When a user requests a webpage that contains a box for an advertisement, the system should decide which advertisement (among those in the current campaign) to show to this particular user based on a vector containing all her features, often collected by a cookie. Our goal here is to generate

a strategy that for each user of the website selects an ad that when it is presented to her has the highest probability to be clicked on. These days, almost all the industrial personalized ad recommendation systems use supervised learning or contextual bandits algorithms. These methods are based on the i.i.d. assumption of the visits (to the website) and do not discriminate between a visit and a visitor, i.e., each visit is considered as a new visitor that has been sampled i.i.d. from the population of the visitors. As a result, these algorithms are myopic and do not try to optimize for the long-term performance. Despite their success, these methods seem to be insufficient as users establish longer-term relationship with the websites they visit, i.e., the ad recommendation systems should deal with more and more returning visitors. The increase in returning visitors violates (more) the main assumption underlying the supervised learning and bandit algorithms, i.e., there is no difference between a visit and a visitor, and thus, shows the need for a new class of solutions.

The reinforcement learning (RL) algorithms that have been designed to optimize the long-term performance of the system (expected sum of rewards/costs) seem to be suitable candidates for ad recommendation systems [128]. The nature of these algorithms allows them to take into account all the available knowledge about the user at the current visit, and then selects an offer to maximize the total number of times she will click over multiple visits, also known as the user's life-time value (LTV). Unlike myopic approaches, RL algorithms differentiate between a visit and a visitor, and consider all the visits of a user (in chronological order) as a system trajectory generated by her. In this approach, while the visitors are i.i.d. samples from the population of the users, their visits are not. This long-term approach to the ad recommendation problem allows us to make decisions that are not usually possible with myopic techniques, such as to propose an offer to a user that might be a loss to the company in the short term, but has the effect that makes the user engaged with the website/company and brings her back to spend more money in the future.

For our second case study, we use an Adobe personalized ad-recommendation [148] simulator that has been trained based on real data captured with permission from the website of a Fortune 50 company that receives hundreds of visitors per day. The simulator produces a vector of 31 real-valued features that provide a compressed representation of all of the available information about a user. The advertisements are clustered into four high-level classes that the agent must select between. After the agent selects an advertisement, the user either clicks (reward of +1) or does not click (reward of 0) and the feature vector describing the user is updated. In this case, we test our algorithm by maximizing the customers' life-time value in 15 time steps subject to a bounded tail risk.

Instead of using the cost-minimization framework from the main paper, by defining the return random variable (under a fixed policy  $\theta$ )  $\mathcal{R}^\theta(x^0)$  as the (discounted) total number of clicks along a user's trajectory, here we formulate the personalized ad-recommendation problem as a return maximization problem where the tail risk corresponds to the worst case return distribution:

$$\max_{\theta} \mathbb{E} [\mathcal{R}^\theta(x^0)] \quad \text{subject to} \quad \text{CVaR}_{1-\alpha}(-\mathcal{R}^\theta(x^0)) \leq \beta. \quad (3.38)$$

We set the parameters of the MDP as  $T = 15$  and  $\gamma = 0.98$ , the confidence interval and constraint threshold as  $\alpha = 0.05$  and  $\beta = 0.12$ , the number of sample trajectories  $N$  to 1,000,000, and the parameter bounds

as  $\lambda_{\max} = 5,000$  and  $\Theta = [-60, 60]^{\kappa_1}$ , where the dimension of the basis functions is  $\kappa_1 = 4096$ . Similar to the optimal stopping problem, we implement both the trajectory based algorithm (PG, PG-CVaR) and the actor-critic algorithms (AC, AC-CVaR) for risk-neutral and risk sensitive optimal control. Here we used the 3<sup>rd</sup> order Fourier basis with cross-products in [70] as features and search over the family of Boltzmann policies. We compared the performance of PG-CVaR and AC-CVaR, our risk-constrained policy gradient (Algorithm 2) and actor-critic (Algorithms 3) algorithms, with their risk-neutral counterparts (PG and AC). Figure 3.3 shows the distribution of the discounted cumulative return  $\mathcal{R}^\theta(x^0)$  for the policy  $\theta$  learned by each of these algorithms. The results indicate that the risk-constrained algorithms yield a lower expected reward, but have higher left tail (worst-case) reward distributions. Table 3.2 summarizes the findings of this experiment.

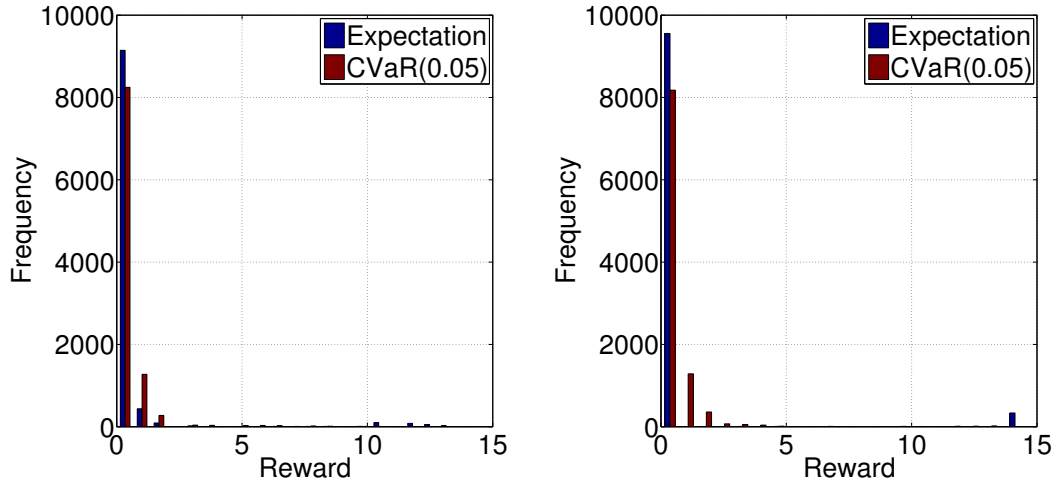


Figure 3.3: Reward distributions for the policies learned by the CVaR-constrained and risk-neutral policy gradient and actor-critic algorithms. The left figure corresponds to the PG methods and the right figure corresponds to the AC algorithms.

	$\mathbb{E}(\mathcal{R}^\theta(x^0))$	$\sigma(\mathcal{R}^\theta(x^0))$	$\text{CVaR}(\mathcal{R}^\theta(x^0))$	$\text{VaR}(\mathcal{R}^\theta(x^0))$
PG	0.396	1.898	0.037	1.000
PG-CVaR	0.287	0.914	0.126	1.795
AC	0.581	2.778	0	0
AC-CVaR	0.253	0.634	0.137	1.890

Table 3.2: Performance comparison of the policies learned by the CVaR-constrained and risk-neutral algorithms. In this table  $\sigma(\mathcal{R}^\theta(x^0))$  stands for the standard deviation of the total reward.

### 3.7 Conclusion

In this chapter we proposed several policy gradient and actor-critic algorithms for CVaR-constrained and chance-constrained optimization in MDPs, and proved their convergence. Using an optimal stopping problem



and a personalized ad-recommendation problem, we showed that our algorithms resulted in policies whose cost distributions have lower right-tail compared to their risk-neutral counterparts. This is important for a risk-averse decision-maker, especially if the right-tail contains catastrophic costs.

In the next chapter we will study the model predictive control approach for risk-sensitive decision-making, where the objective function is characterized by the a general class of time-consistent coherent risk measures. As discussed in Chapter 1, the main advantage of adopting these objective functions in planning is that the resultant policies are always guaranteed to be rational and time-consistent.

## Chapter 4

# Risk Sensitive Model Predictive Control

### 4.1 Introduction

#### 4.1.1 Model Predictive Control

Model Predictive Control (MPC) is one of the most popular methods to address optimal control problems in an online setting [108, 156]. The key idea behind MPC is to obtain the control action by repeatedly solving, at each sampling instant, a finite horizon open-loop optimal control problem, using the current state of the plant as the initial state; the result of the optimization is an (open-loop) control sequence, whose first element is applied to control the system [85].

The classic MPC framework does not provide a systematic way to address model uncertainties and disturbances [15]. Accordingly, one of the main research thrusts for MPC is to find techniques to guarantee persistent feasibility and stability in the presence of disturbances. Essentially, current techniques fall into two categories: (1) min-max (or worst-case) formulations, where the performance indices to be minimized are computed with respect to the worst possible disturbance realization [72, 50, 98], and (2) stochastic formulations, where *risk-neutral expected* values of performance indices (and possibly constraints) are considered [15, 107].

The main drawback of the worst-case approach is that the control law may be too conservative, since the MPC law is required to guarantee stability and constraint fulfillment under the worst-case scenario. On the other hand, stochastic formulations, whereby the assessment of future random outcomes is accomplished through a risk-neutral expectation, may be unsuitable in scenarios where one takes risk-aversion into account and desires to protect the system from large deviations.

#### 4.1.2 MPC with Time Consistent Risk Measures

In this chapter, as a radical departure from traditional approaches, we leverage recent strides in the theory of *dynamic* risk metrics developed by the operations research community [122, 119] to include risk-aversion in

MPC. The key property of *dynamic* risk metrics is that, by assessing risk at multiple points in time, one can guarantee *time-consistency* of risk preferences over time [122, 119]. In particular, the essential requirement for time consistency is that if a certain outcome is considered less risky in all states of the world at stage  $k + 1$ , then it should also be considered less risky at stage  $k$ . Remarkably, in [119], it is proven that any risk measure that is time consistent can be represented as a *composition* of one-step risk metrics. In other words, in multi-period settings, risk (as expected) should be compounded over time.

### 4.1.3 Chapter Contribution

The contribution of this chapter is threefold. First, we introduce a notion of dynamic risk metric, referred to as Markov dynamic polytopic risk metric, that captures a full range of risk assessments and enjoys a geometrical structure that is particularly favorable from a computational standpoint. Second, we present and analyze a *risk-averse* MPC algorithm that minimizes in a receding-horizon fashion a Markov dynamic polytopic risk metric, under the assumption that the system's model is linear and is affected by multiplicative uncertainty. Finally, by exploring the “geometrical” structure of Markov dynamic polytopic risk metrics, we present a convex programming formulation for risk-averse MPC that is amenable to a real-time implementation (for moderate horizon lengths). Our framework has three main advantages: (1) it is axiomatically justified, in the sense that risk, by construction, is assessed in a time-consistent fashion; (2) it is amenable to dynamic and convex optimization, primarily due to the compositional form of Markov dynamic polytopic risk metrics and their geometry; and (3) it is general, in that it captures a full range of risk assessments from risk-neutral to worst-case. In this respect, our formulation represents a *unifying* approach for risk-averse MPC.

### 4.1.4 Chapter Organization

The rest of the chapter is organized as follows. In Section 4.2 we discuss the stochastic model we address in this chapter. In Section 4.3 we introduce and discuss the notion of Markov dynamic polytopic risk metrics. In Section 4.4 we state the infinite horizon optimal control problem we wish to address and in Section 4.5 we derive conditions for risk-averse closed-loop stability. From Section 4.6 to 4.8 we present a risk-averse model predictive control law, its performance analysis and various approaches for its computation, respectively. Numerical experiments are presented and discussed in Section 4.9. Finally, complete proofs of the technical results and further extensions can be found in Section 7.4.

## 4.2 Model Description

Consider the discrete time system:

$$x_{k+1} = A(w_k)x_k + B(w_k)a_k, \quad (4.1)$$

where  $k \in \mathbb{N}$  is the time index,  $x_k \in \mathbb{R}^{N_x}$  is the state,  $a_k \in \mathbb{R}^{N_a}$  is the (unconstrained) control input, and  $w_k \in \mathcal{W}$  is the process disturbance. We assume that the initial condition  $x_0$  is deterministic. We assume that  $\mathcal{W}$  is a finite set of cardinality  $L$ , i.e.,  $\mathcal{W} = \{w^{[1]}, \dots, w^{[L]}\}$ . For each stage  $k$  and state-control pair  $(x_k, a_k)$ , the process disturbance  $w_k$  is drawn from set  $\mathcal{W}$  according to the probability mass function  $p = [p(1), p(2), \dots, p(L)]^\top$ , where  $p(j) = \mathbb{P}(w_k = w^{[j]})$ ,  $j \in \{1, \dots, L\}$ . Without loss of generality, we assume that  $p(j) > 0$  for all  $j$ . Note that the probability mass function for the process disturbance is time-invariant, and that the process disturbance is *independent* of the process history and of the state-control pair  $(x_k, a_k)$ . Under these assumptions, the stochastic process  $\{x_k\}$  is clearly a Markov process.<sup>1</sup>

By enumerating all  $L$  realizations of the process disturbance  $w_k$ , system (4.1) can be rewritten as:

$$x_{k+1} = \begin{cases} A_1 x_k + B_1 a_k & \text{if } w_k = w^{[1]}, \\ \vdots & \vdots \\ A_L x_k + B_L a_k & \text{if } w_k = w^{[L]}, \end{cases}$$

where  $A_j := A(w^{[j]})$  and  $B_j := B(w^{[j]})$ ,  $j \in \{1, \dots, L\}$ .

The results presented in this chapter can be immediately extended to the time-varying case (i.e., where the probability mass function for the process disturbance is time-varying). To simplify notation, however, we prefer to focus this chapter on the time-invariant case.

### 4.3 Markov Polytopic Risk Measures

In this section we *refine* the notion of Markov (dynamic and time-consistent) risk metrics (as defined in Theorem 1.3.8) by adding a polytopic structure to the dual representation of coherent risk metrics. This will lead to the definition of Markov dynamic polytopic risk metrics, which enjoy favorable computational properties and, at the same time, maintain most of the generality of dynamic time-consistent risk metrics.

#### 4.3.1 Polytopic Risk Measures

According to the discussion in Section 4.2, the probability space for the process disturbance has a finite number of elements. Accordingly, consider Theorem 1.3.3; by definition of expectation, one has  $\mathbb{E}_\zeta[Z] = \sum_{j=1}^L Z(j)p(j)\zeta(j)$ . In our framework (inspired by [53]), we consider coherent risk measures where the risk envelope  $\mathcal{U}$  is a *polytope*, i.e., there exist matrices  $S^I, S^E$  and vectors  $T^I, T^E$  of appropriate dimensions such that

$$\mathcal{U}^{\text{poly}} = \{\zeta \in \mathcal{B} \mid S^I \zeta \leq T^I, S^E \zeta = T^E\}.$$

<sup>1</sup>In the context of MDPs (Section 1.2), in this problem the state space  $\mathcal{X}$  is  $\mathbb{R}^{N_x}$ , the action space  $\mathcal{A}$  is  $\mathbb{R}^{N_a}$ , and the state evolution follows from (4.1). According to the problem formulation in Section 4.4, the discounting factor  $\gamma$  equals to 1, and the immediate cost  $c(x, a)$  is given by  $x^\top Qx + u^\top Ru$ , where  $(Q, R)$  is a set of state and control weighting matrices.

We will refer to coherent risk measures representable with a polytopic risk envelope as *polytopic risk measures*. Consider the bijective map  $q(j) := p(j)\zeta(j)$  (recall that, in our model,  $p(j) > 0$ ). Then, by applying such map, one can easily rewrite a polytopic risk measure as

$$\rho(Z) = \max_{q \in \mathcal{U}^{\text{poly}}} \mathbb{E}_q[Z],$$

where  $q$  is a *probability mass function* belonging to a polytopic subset of the standard simplex, i.e.:

$$\mathcal{U}^{\text{poly}} = \left\{ q \in \Delta^L \mid S^I q \leq T^I, S^E q = T^E \right\}, \quad (4.2)$$

where  $\Delta^L := \{q \in \mathbb{R}^L : \sum_{j=1}^L q(j) = 1, q \geq 0\}$ . Accordingly, one has  $E_q[Z] = \sum_{j=1}^L Z(j)q(j)$  (note that, with a slight abuse of notation, we are using the same symbols as before for  $\mathcal{U}^{\text{poly}}$ ,  $S^I$ , and  $S^E$ ).

The class of polytopic risk measures is large: we give below some examples (also note that any comonotonic risk measure is a polytopic risk measure [62]).

**Example 4.3.1** (Examples of Polytopic Risk Measures). *As a first example, the expected value of a random variable  $Z$  can be represented according to equation (4.2) with polytopic risk envelope*

$$\mathcal{U}^{\text{poly}} = \left\{ q \in \Delta^L \mid q(j) = p(j) \text{ for all } j \in \{1, \dots, L\} \right\}.$$

*A second example is represented by the average upper semi-deviation risk metric, defined as*

$$\rho_{\text{AUS}}(Z) := \mathbb{E}[Z] + c \mathbb{E}[(Z - \mathbb{E}[Z])^+],$$

where  $0 \leq c \leq 1$  and  $(x)^+ := \max(0, x)$ . This metric can be represented according to equation (4.2) with polytopic risk envelope ([93, 132]):

$$\mathcal{U}^{\text{poly}} = \left\{ q \in \Delta^L \mid q(j) = p(j) \left( 1 + h(j) - \sum_{j=1}^L h(j)p(j) \right), 0 \leq h(j) \leq c, j \in \{1, \dots, L\} \right\}.$$

*A related risk metric is the mean absolute semi-deviation risk metric, defined as*

$$\rho_{\text{AUS}}(Z) = \mathbb{E}[Z] + c \mathbb{E} \left[ \left| Z - \mathbb{E}[Z] \right| \right],$$

where  $0 \leq c \leq 1$ . This metric can be represented according to equation (4.2) with polytopic risk envelope ([93]):

$$\mathcal{U}^{\text{poly}} = \left\{ q \in \Delta^L \mid q_l = p_l \left( 1 + h_l - \sum_{l=1}^L h_l p_l \right), -c \leq h_l \leq c, l \in \{1, \dots, L\} \right\}.$$

A risk metric that is very popular in the finance industry is the Conditional Value-at-Risk (CVaR), defined as ([112])

$$\text{CVaR}_\alpha(Z) := \inf_{y \in \mathbb{R}} \left[ y + \frac{1}{\alpha} \mathbb{E}[(Z - y)^+] \right], \quad (4.3)$$

where  $\alpha \in (0, 1]$ .  $\text{CVaR}_\alpha$  can be represented according to equation (4.2) with the polytopic risk envelope (see [132]):

$$\mathcal{U}^{\text{poly}} = \left\{ q \in \Delta^L \mid 0 \leq q(j) \leq \frac{p(j)}{\alpha} \text{ for all } j \in \{1, \dots, L\} \right\}.$$

As a special case of the Conditional Value-at-Risk, by setting  $\alpha = 0$ , the worst case risk defined as

$$\text{WCR}(Z) = \max\{Z(j) : j \in \{1, \dots, L\}\}, \quad (4.4)$$

can be trivially represented according to (4.2) with polytopic risk envelope  $\mathcal{U}^{\text{poly}} = \Delta^L$ .

Other important examples include the spectral risk measures [21], the optimized certainty equivalent and expected utility [13, 132], and the distributionally-robust risk [15]. The key point is that the notion of polytopic risk metric covers a full gamut of risk assessments, ranging from risk-neutral to worst case.

### 4.3.2 Markov Dynamic Polytopic Risk Metrics

Note that in the definition of dynamic, time-consistent risk measures, since at stage  $k$  the value of  $\rho_k$  is  $\mathcal{F}_k$ -measurable, the evaluation of risk can depend on the *whole* past, see [119, Section IV]. For example, the weight  $c$  in the definition of the average upper mean semi-deviation risk metric can be an  $\mathcal{F}_k$ -measurable random variable (see [119, Example 2]). This generality, which appears of little practical value in many cases, leads to optimization problems that are intractable. This motivates us to add a *Markovian structure* to dynamic, time-consistent risk measures (similarly as in [119]). We start by introducing the notion of Markov polytopic risk measure (similar to [119, Definition 6]).

**Definition 4.3.2** (Markov Polytopic Risk Measures). *Consider the Markov process  $\{x_k\}$  that evolves according to equation (4.1). A coherent one-step conditional risk measure  $\rho_k(\cdot)$  is a Markov polytopic risk measure with respect to  $\{x_k\}$  if it can be written as*

$$\rho_k(Z(x_{k+1})) = \max_{q \in \mathcal{U}_k^{\text{poly}}(x_k, p)} \mathbb{E}_q[Z(x_{k+1})]$$

where  $\mathcal{U}_k^{\text{poly}}(x_k, p) = \{q \in \Delta^L \mid S_k^I(x_k, p)q \leq T_k^I(x_k, p), S_k^E(x_k, p)q = T_k^E(x_k, p)\}$  is the polytopic risk envelope.

In other words, a Markov polytopic risk measure is a coherent one-step conditional risk measure where the evaluation of risk is not allowed to depend on the whole past (for example, the weight  $c$  in the definition of the average upper mean semi-deviation risk metric can depend on the past only through  $x_k$ ), and the risk envelope is a polytope. Correspondingly, we define a Markov dynamic polytopic risk metric as follows.

**Definition 4.3.3** (Markov Dynamic Polytopic Risk Measures). *Consider the Markov process  $\{x_k\}$  that evolves according to equation (4.1). A Markov dynamic polytopic risk measure is a set of mappings  $\rho_{k,N} : \mathcal{Z}_{k,N} \rightarrow \mathcal{Z}_k$  defined as*

$$\rho_{k,N} = Z(x_k) + \rho_k(Z(x_{k+1}) + \dots + \rho_{N-2}(Z(x_{N-1}) + \rho_{N-1}(Z(x_N))) \dots),$$

for  $k \in \{0, \dots, N\}$ , where  $\rho_k$  are single-period Markov polytopic risk measures.

Clearly, a Markov dynamic polytopic risk metric is time consistent. Definition 4.3.3 can be extended to the case where the probability distribution for the disturbance depends on the current state and control action. We avoid this generalization to keep the exposition simple and consistent with model (4.1).

### 4.3.3 Computational Aspects of Markov Dynamic Polytopic Risk Metrics

According to Definition 4.3.3, Markov dynamic polytopic risk measures are obtained by compounding coherent one-step conditional risk measures, whose risk envelope is a polytope. Some of the algorithms presented in Section 4.8 require a vertex representation of such polytopes (rather than the hyperplane representation in Definition 4.3.2). Several methods are available to enumerate the vertices of a polytope, such as the Fourier-Motzkin elimination method, the simplex method, and the iterative linear programming method, see [89, Section 5] and references therein. In our implementation, we use the vertex enumeration function included in the MPT toolbox [75], which relies on the simplex method.

## 4.4 Problem Formulation

In light of Sections 4.2 and 4.3, we are now in a position to state the risk-averse optimization problem we wish to solve in this chapter. Our problem formulation relies on Markov dynamic polytopic risk metrics that satisfy the following stationarity assumption.

**Assumption 4.4.1** (Time-invariance of Risk Assessments). *The polytopic risk envelopes  $\mathcal{U}_k^{\text{poly}}$  are independent of time  $k$  and state  $x_k$ , i.e.  $\mathcal{U}_k^{\text{poly}}(x_k, p) = \mathcal{U}^{\text{poly}}(p)$  for all  $k$ .*

This assumption is crucial for the well-posedness of our formulation and to devise a tractable MPC algorithm that relies on linear matrix inequalities. We next introduce a notion of stability tailored to our risk-averse context.

**Definition 4.4.2** (Uniform Global Risk-Sensitive Exponential Stability). *System (4.1) is said to be Uniformly Globally Risk-Sensitive Exponentially Stable (UGRSES) if there exist constants  $c \geq 0$  and  $\lambda \in [0, 1)$  such that for all initial conditions  $x_0 \in \mathbb{R}^{N_x}$ ,*

$$\rho_{0,k}(0, \dots, 0, x_k^\top x_k) \leq c \lambda^k x_0^\top x_0, \quad \text{for all } k \in \mathbb{N}, \quad (4.5)$$

where  $\{\rho_{0,k}\}$  is a Markov dynamic polytopic risk measure satisfying Assumption 4.4.1. If condition (4.5) only holds for initial conditions within some bounded neighborhood  $\Omega$  of the origin, the system is said to be *Uniformly Locally Risk-Sensitive Exponentially Stable (ULRSES)* with domain  $\Omega$ .

Note that, in general, UGRSES is a *more restrictive* stability condition than mean-square stability, as clarified by the following example.

**Example 4.4.3** (Mean-Square Stability versus Risk-Sensitive Stability). *System (4.1) is said to be Uniformly Globally Mean-Square Exponentially Stable (UGMSES) if there exist constants  $c \geq 0$  and  $\lambda \in [0, 1)$  such that for all initial conditions  $x_0 \in \mathbb{R}^{N_x}$ ,*

$$\mathbb{E}[x_k^\top x_k] \leq c \lambda^k x_0^\top x_0, \quad \text{for all } k \in \mathbb{N},$$

see [123, Definition 1] and [15, Definition 1]. Consider the discrete time system

$$x_{k+1} = \begin{cases} \sqrt{0.5} x_k & \text{with probability } 0.2, \\ \sqrt{1.1} x_k & \text{with probability } 0.8. \end{cases} \quad (4.6)$$

A sufficient condition for system (4.6) to be UGMSES is that there exist positive definite matrices  $P = P^\top \succ 0$  and  $L = L^\top \succ 0$  such that

$$\mathbb{E}[x_{k+1}^\top P x_{k+1}] - x_k^\top P x_k \leq -x_k^\top L x_k,$$

for all  $k \in \mathbb{N}$ , see [15, Lemma 1]. One can easily check that with  $P = 100$  and  $L = 1$  the above inequality is satisfied, and, hence system (4.6) is UGMSES.

Assuming risk is assessed according to the Markov dynamic polytopic risk metric  $\rho_{0,k} = \text{CVaR}_{0.5} \circ \dots \circ \text{CVaR}_{0.5}$ , we next show that system (4.6) is not UGRSES. In fact, using the dual representation given in Example 4.3.1, one can write

$$\text{CVaR}_{0.5}(Z(x_{k+1})) = \max_{q \in \mathcal{U}^{\text{poly}}} \mathbb{E}_q[Z(x_{k+1})], \text{ where } \mathcal{U}^{\text{poly}} = \{q \in \Delta^2 \mid 0 \leq q_1 \leq 0.4, 0 \leq q_2 \leq 1.6\}.$$

Consider the probability mass function  $\bar{q} = [0.1/1.1, 1/1.1]^\top$ . Since  $\bar{q} \in \mathcal{U}^{\text{poly}}$ , one has

$$\text{CVaR}_{0.5}(x_{k+1}^2) \geq 0.5 x_k^2 \frac{0.1}{1.1} + 1.1 x_k^2 \frac{1}{1.1} = 1.0455 x_k^2.$$

By repeating this argument, one can then show that

$$\rho_{0,k}(x_{k+1}^2) = \text{CVaR}_{0.5} \circ \dots \circ \text{CVaR}_{0.5}(x_{k+1}^2) \geq a^{k+1} x_0^\top x_0,$$

where  $a = 1.0455$ . Hence, one cannot find constants  $c$  and  $\lambda$  that satisfy equation (4.5). Consequently, system (4.6) is UGMSES but not UGRSES.

Consider the MDP described in Section 4.2 and let  $\Pi$  be the set of all stationary feedback control policies,



i.e.,  $\Pi := \{\pi : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_a}\}$ . Consider the quadratic cost function  $C : \mathbb{R}^{N_x} \times \mathbb{R}^{N_a} \rightarrow \mathbb{R}_{\geq 0}$  defined as  $C(x, a) := x^\top Q x + u^\top R u$ , where  $Q = Q^\top \succ 0$  and  $R = R^\top \succ 0$  are given state and control penalties. Define the multi-stage cost function:

$$J_{0,k}(x_0, \pi) := \rho_{0,k}\left(C(x_0, \pi(x_0)), \dots, C(x_k, \pi(x_k))\right),$$

where  $\{\rho_{0,k}\}$  is a Markov dynamic polytopic risk measure satisfying Assumption 4.4.1. The problem we wish to address is as follows.

**Optimization problem  $\mathcal{OPT}_{\text{RS}}$**  — Given an initial state  $x_0 \in \mathbb{R}^{N_x}$ , solve

$$\begin{aligned} & \inf_{\pi \in \Pi} \quad \limsup_{k \rightarrow \infty} J_{0,k}(x_0, \pi) \\ \text{such that} \quad & x_{k+1} = A(w_k)x_k + B(w_k)\pi(x_k) \\ & \|T_a\pi(x_k)\|_2 \leq a_{\max}, \quad \|T_x x_k\|_2 \leq x_{\max} \\ & \text{System is UGRSES} \end{aligned}$$

where  $(T_a, a_{\max})$  and  $(T_x, x_{\max})$  describe the second order cone constraints for the control and state, respectively.

We denote the optimal cost function as  $J_{0,\infty}^*(x_0)$ . Note that the risk measure in the definition of UGRSES is assumed to be identical to the risk measure used to evaluate the cost of a policy. Also, by Assumption 4.4.1, the single-period risk metrics are time-invariant, hence one can write

$$\rho_{0,k}\left(C(x_0, \pi(x_0)), \dots, C(x_k, \pi(x_k))\right) = C(x_0, \pi(x_0)) + \rho(C(x_1, \pi(x_1)) + \dots + \rho(C(x_k, \pi(x_k))) \dots), \quad (4.7)$$

where  $\rho$  is a given Markov polytopic risk metric that models the “amount” of risk aversion. This chapter addresses problem  $\mathcal{OPT}_{\text{RS}}$  along three main dimensions:

1. Find lower bounds for the optimal cost of problem  $\mathcal{OPT}_{\text{RS}}$ .
2. Find sufficient conditions for *risk-sensitive* stability (i.e., for UGRSES).
3. Design a model predictive control algorithm to efficiently compute a suboptimal state-feedback control policy.

In the next section, we provide sufficient conditions for (4.1) to be UGRSES, thereby leading to the discussion on the MPC adaption of problem  $\mathcal{OPT}_{\text{RS}}$ .

## 4.5 Risk-Sensitive Stability

In this section we provide a sufficient condition for system (4.1) to be UGRSES, under the assumptions of Section 4.4. This condition relies on Lyapunov techniques and is inspired by [15] (Lemma 4.5.1 indeed reduces to Lemma 1 in [15] when the risk measure is simply an expectation).

**Lemma 4.5.1** (Sufficient Conditions for UGRSES). *Consider a policy  $\pi \in \Pi$  and the corresponding closed-loop dynamics for system (4.1), denoted by  $x_{k+1} = f(x_k, w_k)$ . The closed-loop system is UGRSES if there exists a function  $V(x) : \mathbb{R}^{N_x} \rightarrow \mathbb{R}$  and scalars  $b_1, b_2, b_3 > 0$ , such that for all  $x \in \mathbb{R}^{N_x}$ ,*

$$\begin{aligned} b_1 \|x\|^2 &\leq V(x) \leq b_2 \|x\|^2, \text{ and} \\ \rho(V(f(x, w))) - V(x) &\leq -b_3 \|x\|^2. \end{aligned} \tag{4.8}$$

**Remark 4.5.2** (Sufficient Conditions for ULRSES). *The closed-loop system is ULRSES with domain  $\Omega$  if the conditions in (4.8) only hold within the bounded set  $\Omega$ .*

## 4.6 Model Predictive Control Problem

### 4.6.1 The Unconstrained Case

In this section we set up the receding horizon version of problem  $\mathcal{OPT}_{RS}$ , under the assumption that there are no constraints. This will lead to a model predictive control algorithm for the (suboptimal) solution of problem  $\mathcal{OPT}_{RS}$ . Consider the following receding-horizon cost function for  $N \geq 1$ :

$$\begin{aligned} &J(x_{k|k}, \pi_{k|k}, \dots, \pi_{k+N-1|k}, P) \\ &:= \rho_{k,k+N} \left( C(x_{k|k}, \pi_{k|k}(x_{k|k})), \dots, C(x_{k+N-1|k}, \pi_{k+N-1|k}(x_{k+N-1|k})), x_{k+N}^\top P x_{k+N} \right), \end{aligned} \tag{4.9}$$

where  $x_{h|k}$  is the state at time  $h$  predicted at stage  $k$  (a *discrete* random variable),  $\pi_{h|k}$  is the control policy to be applied at time  $h$  as determined at stage  $k$  (i.e.,  $\pi_{h|k} : \mathbb{R}^{N_x} \rightarrow \mathbb{R}^{N_a}$ ), and  $P = P^\top \succ 0$  is a terminal weight matrix. We are now in a position to state the model predictive control problem.

**Optimization problem  $\mathcal{MPC}$**  — Given an initial state  $x_{k|k} \in \mathbb{R}^{N_x}$  and a prediction horizon  $N \geq 1$ , solve

$$\begin{aligned} &\min_{\pi_{k|k}, \dots, \pi_{k+N-1|k}} J(x_{k|k}, \pi_{k|k}, \dots, \pi_{k+N-1|k}, P) \\ &\text{such that } x_{k+h+1|k} = A(w_{k+h})x_{k+h|k} + B(w_{k+h})\pi_{k+h|k}(x_{k+h|k}) \end{aligned}$$

for  $h \in \{0, \dots, N-1\}$ .

Note that a Markov policy is guaranteed to be optimal for problem  $\mathcal{MPC}$  (see [119, Theorem 2]). The optimal cost function for problem  $\mathcal{MPC}$  is denoted by  $J_k^*(x_{k|k})$ , and a minimizing policy is denoted by

$\{\pi_{k|k}^*, \dots, \pi_{k+N-1|k}^*\}$  (if multiple minimizing policies exist, then one of the minimizing policies is selected arbitrarily). For each state  $x_k$ , we set  $x_{k|k} = x_k$  and the (time-invariant) model predictive control law is then defined as

$$\pi^{MPC}(x_k) = \pi_{k|k}^*(x_{k|k}). \quad (4.10)$$

Note that the model predictive control problem  $\mathcal{MPC}$  involves an optimization over *time-varying closed-loop policies*, as opposed to the classical deterministic case where the optimization is over open-loop sequences. A similar approach is taken in [107, 15]. We will show in Section 4.8 how to solve problem  $\mathcal{MPC}$  efficiently.

The following theorem shows that the model predictive control law (4.18), with a proper choice of the terminal weight  $P$ , is risk-sensitive stabilizing, i.e., the closed-loop system (4.1) is UGRSES.

**Theorem 4.6.1** (Stochastic Stability for Model Predictive Control Law, Unconstrained Case). *Consider the model predictive control law in equation (4.18) and the corresponding closed-loop dynamics for system (4.1) with initial condition  $x_0 \in \mathbb{R}^{N_x}$ . Suppose that  $P = P^\top \succ 0$ , and there exists a matrix  $F$  such that:*

$$\sum_{j=1}^L q_l(j) (A_j + B_j F)^\top P (A_j + B_j F) - P + Q + F^\top R F \prec 0, \quad (4.11)$$

for all  $l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}$ , where  $\mathcal{U}^{\text{poly}, V}(p)$  is the set of vertices of polytope  $\mathcal{U}^{\text{poly}}(p)$ ,  $q_l$  is the  $l$ th element in set  $\mathcal{U}^{\text{poly}, V}(p)$ , and  $q_l(j)$  denotes the  $j$ th component of vector  $q_l$ ,  $j \in \{1, \dots, L\}$ . Then, the closed loop system (4.1) is UGRSES.

## 4.6.2 The Constrained Case

We now enforce the state and control constraints introduced in problem  $\mathcal{OPT}_{\text{RS}}$  within the receding horizon framework. Consider the time-invariant ellipsoids:

$$\mathbb{A} := \{a \in \mathbb{R}^{N_a} \mid \|T_a a\|_2 \leq a_{\max}\}, \quad \mathbb{X} := \{x \in \mathbb{R}^{N_x} \mid \|T_x x\|_2 \leq x_{\max}\}.$$

While we focus on ellipsoidal state and control constraints in this chapter, our methodology can readily accommodate component-wise and polytopic constraints via suitable LMI representations, for example, see [125, 44, 6] for detailed derivations.

Our receding horizon framework may be decomposed into two steps. First, offline, we search for an ellipsoidal set  $\mathcal{E}_{\max}$  and a *local* feedback control law  $a(x) = Fx$  that renders  $\mathcal{E}_{\max}$  control invariant and ensures satisfaction of state and control constraints. Additionally, within the offline step, we search for a terminal cost matrix  $P$  (for the online MPC problem) to ensure that the closed-loop dynamics under the model predictive controller are risk-sensitive exponentially stable. The online MPC optimization then constitutes

the second step of our framework. Consider first, the offline step. We parameterize  $\mathcal{E}_{\max}$  as follows:

$$\mathcal{E}_{\max}(W) := \{x \in \mathbb{R}^{N_x} \mid x^\top W^{-1} x \leq 1\}, \quad (4.12)$$

where  $W$  (and hence  $W^{-1}$ ) is a positive definite matrix. The (offline) optimization problem to solve for  $W$ ,  $F$ , and  $P$  is presented below.

**Optimization problem  $\mathcal{PE}$  — Solve**

$$\begin{aligned} \max_{W=W^\top \succ 0, F, P=P^\top \succ 0} \quad & \log \det(W) \\ \text{such that} \quad & F^\top \frac{T_a^\top T_a}{a_{\max}^2} F - W^{-1} \preceq 0, \end{aligned} \quad (4.13)$$

$$(A_j + B_j F)^\top \frac{T_x^\top T_x}{x_{\max}^2} (A_j + B_j F) - W^{-1} \preceq 0, \forall j \in \{1, \dots, L\}, \quad (4.14)$$

$$(A_j + B_j F)^\top W^{-1} (A_j + B_j F) - W^{-1} \preceq 0, \forall j \in \{1, \dots, L\}, \quad (4.15)$$

$$\begin{aligned} \sum_{j=1}^L q_l(j) (A_j + B_j F)^\top P (A_j + B_j F) - P + (F^\top R F + Q) &\prec 0 \\ \forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}. \end{aligned} \quad (4.16)$$

The objective in problem  $\mathcal{PE}$  is to maximize the volume of the control invariant ellipsoid  $\mathcal{E}_{\max}(W)$ . Note that  $\mathcal{E}_{\max}(W)$  may contain states outside of  $\mathbb{X}$ , however, we restrict our domain of interest to the intersection  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$ . The bi-linear semi-definite inequality in (7.73) defines the terminal cost matrix  $P$ , and will be instrumental in proving risk-sensitive stability for system (4.1) under the model predictive control law. We first analyze the properties of the state feedback control law  $a(x) = Fx$  within the set  $\mathcal{E}_{\max}(W)$ .

**Lemma 4.6.2** (Properties of  $\mathcal{E}_{\max}$ ). *Suppose problem  $\mathcal{PE}$  is feasible and  $x \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$ . Let  $a(x) = Fx$ . Then, the following statements are true:*

1.  $\|T_a a\|_2 \leq a_{\max}$ , i.e., the control constraint is satisfied.
2.  $\|T_x (A(w)x + B(w)a)\|_2 \leq x_{\max}$  surely, i.e., the state constraint is satisfied at the next step surely.
3.  $A(w)x + B(w)a \in \mathcal{E}_{\max}(W)$  surely, i.e., the set  $\mathcal{E}_{\max}(W)$  is robust control invariant under the control law  $a(x) = Fx$ .

Thus,  $a(x) \in \mathbb{A}$  and  $A(w)x + B(w)a \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$  surely.

Lemma 4.6.2 establishes  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$  as a robust control invariant set under the feasible local feedback control law  $a(x) = Fx$ . This result will be crucial to ascertain the persistent feasibility properties of the online optimization algorithm and the resulting closed-loop stability.

We are now ready to formalize the MPC problem. Suppose the feasible set of solutions in problem  $\mathcal{PE}$  is non-empty and define  $W = W^*$  and  $P = P^*$ , where  $W^*, P^*$  are the maximizers for problem  $\mathcal{PE}$ . Consider the following online optimization problem:

**Optimization problem  $\mathcal{MPC}$**  — Given an initial state  $x_{k|k} \in \mathbb{X}$  and a prediction horizon  $N \geq 1$ , solve

$$\begin{aligned} \min_{\pi_{k|k}, \dots, \pi_{k+N-1|k}} & J(x_{k|k}, \pi_{k|k}, \dots, \pi_{k+N-1|k}, P) \\ \text{such that} & x_{k+h+1|k} = A(w_{k+h})x_{k+h|k} + B(w_{k+h})\pi_{k+h|k}(x_{k+h|k}), \\ & \|T_a \pi_{k+h|k}(x_{k+h|k})\|_2 \leq a_{\max}, \|T_x x_{k+h+1|k}\|_2 \leq x_{\max}, h \in \{0, \dots, N-1\}, \\ & x_{k+N|k} \in \mathcal{E}_{\max}(W) \text{ surely.} \end{aligned} \quad (4.17)$$

Note that a Markov policy is guaranteed to be optimal for problem  $\mathcal{MPC}$  (see [119, Theorem 2]). The optimal cost function for problem  $\mathcal{MPC}$  is denoted by  $J_k^*(x_{k|k})$ , and a minimizing policy is denoted by  $\{\pi_{k|k}^*, \dots, \pi_{k+N-1|k}^*\}$ . For each state  $x_k$ , we set  $x_{k|k} = x_k$  and the (time-invariant) model predictive control law is then defined as

$$\pi^{\mathcal{MPC}}(x_k) = \pi_{k|k}^*(x_{k|k}). \quad (4.18)$$

**Remark 4.6.3.** While the problem formulation in this chapter considers hard state and control constraints, the approach may be readily adapted to accommodate probabilistic constraints using the method described in [38, 39]. The key idea is to consider control laws of the form  $a(x) = Fx + c$ , where  $F$  is fixed and solved offline and  $c$  is computed online within the MPC algorithm. Additionally, the offline step solves for a set in which state and control constraints are satisfied surely, and the set is probabilistically control invariant with some desired confidence level. The online MPC problem then tries to either retain the state within this set, or return the state back to this set. State and control constraint satisfaction can then be assured with the desired probabilistic confidence.

Note that problem  $\mathcal{MPC}$  involves an optimization over *time-varying closed-loop policies*, as opposed to the classical deterministic case where the optimization is over open-loop control inputs. A similar approach is taken in [107, 15]. We will show in Section 4.8 how to solve problem  $\mathcal{MPC}$  efficiently. We now address the persistent feasibility and stability properties for problem  $\mathcal{MPC}$ .

#### 4.6.2.1 Persistent Feasibility for problem $\mathcal{MPC}$

The following theorem proves that problem  $\mathcal{MPC}$  is persistently feasible:

**Theorem 4.6.4** (Persistent Feasibility). *Define  $\mathcal{X}_N$  to be the set of initial states for which problem  $\mathcal{MPC}$  is feasible. Assume  $x_{k|k} \in \mathcal{X}_N$  and the control law is given by (4.18). Then, it follows that  $x_{k+1|k+1} \in \mathcal{X}_N$  surely.*

**Remark 4.6.5** (Compactness of  $\mathcal{X}_N$ ). *By leveraging the finite cardinality of the disturbance set  $\mathcal{W}$  and the set closure preservation property attributed to the inverse of continuous functions, it is possible to show that  $\mathcal{X}_N$  is closed. Then, since  $\mathcal{X}_N$  is necessarily a subset of the bounded set  $\mathbb{X}$ , it follows that  $\mathcal{X}_N$  is compact.*

#### 4.6.2.2 ULRSES Stability for Problem MPC

The following theorem demonstrates that the closed-loop system under the MPC control law is ULRSES:

**Theorem 4.6.6** (Stochastic Stability with MPC). *Suppose the initial state  $x_0$  lies within  $\mathcal{X}_N$ . Then, under the model predictive control law given in (4.18), the closed-loop system is ULRSES with domain  $\mathcal{X}_N$ .*

**Remark 4.6.7** (Performance Comparisons). *The two-step optimization methodology proposed via Problems  $\mathcal{PE}$  and MPC is similar to the approach described in [15] in that both the control invariant ellipsoid ( $\mathcal{E}_{\max}$ ) and the conditions to ensure stability are computed offline, while problem MPC is solved online. This hybrid procedure is more computationally efficient than the online algorithm given in [98], and boasts better performance as compared with the offline algorithm in [72]. On the other hand, the stability analysis here differs from [15] since we use  $J_k^*$  as the Lyapunov function instead of the fixed quadratic form described in [15]. This allows us to explicitly characterize the cost function performance of the closed-loop dynamics under the model predictive control law with respect to  $J_{0,\infty}^*$  and  $J_k^*$ . To gain additional insight into this comparison, we present an alternative formulation of problems  $\mathcal{PE}$  and MPC in Section 7.4.12, analogous to the approach in [15].*

Having proven persistent feasibility for the online MPC algorithm and ULRSES stability for the resulting closed-loop dynamics, we now present both lower and upper bounds for the infinite horizon cost function associated with the MPC algorithm. This in-turn allows us to quantify the sub-optimality of the receding horizon adaptation of problem  $\mathcal{OPT}_{\text{RS}}$ .

## 4.7 Bounds on Optimal Cost

In this section, by leveraging semi-definite programming, we provide a lower bound for the optimal cost of problem  $\mathcal{OPT}_{\text{RS}}$  and an upper bound for the optimal cost using the MPC algorithm. These bounds will be used in Section 4.9 to bound the factor of sub-optimality for our MPC control algorithm. On top of these results, a complete theoretical analysis of MPC sub-optimality performance can be found in Section 7.4.13.

### 4.7.1 Lower Bound

In the following, let

$$\overline{A} := \begin{bmatrix} A_1^\top & \dots & A_L^\top \end{bmatrix}^\top, \quad \text{and} \quad \overline{B} := \begin{bmatrix} B_1^\top & \dots & B_L^\top \end{bmatrix}^\top.$$

Furthermore, let

$$\Sigma_l := \text{diag}(q_l(1), \dots, q_l(L)) \succ 0,$$

for all  $l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}$ , where  $\mathcal{U}^{\text{poly}, V}(p)$  is the set of vertices of polytope  $\mathcal{U}^{\text{poly}}(p)$ ,  $q_l$  is the  $l$ th element in set  $\mathcal{U}^{\text{poly}, V}(p)$ , and  $q_l(j)$  denotes the  $j$ th component of vector  $q_l$ ,  $j \in \{1, \dots, L\}$ .

**Theorem 4.7.1** (Lower Bound for Problem  $\mathcal{OPT}_{RS}$ ). *Suppose there exists a symmetric matrix  $X \succ 0$  such that the following Linear Matrix Inequality holds:*

$$\begin{bmatrix} R + \bar{B}^\top (\Sigma_l \otimes X) \bar{B} & \bar{B}^\top (\Sigma_l \otimes X) \bar{A} \\ * & \bar{A}^\top (\Sigma_l \otimes X) \bar{A} - (X - Q) \end{bmatrix} \succeq 0, \quad \forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}. \quad (4.19)$$

Then, the optimal cost of problem  $\mathcal{OPT}_{RS}$  can be lower bounded as

$$J_{0,\infty}^*(x_0) \geq \max\{x_0^\top X x_0 : X \text{ satisfies LMI in equation (4.19)}\}.$$

## 4.7.2 Upper Bound

The following theorem presents an upper bound for the infinite horizon cost incurred when executing the MPC policy.

**Theorem 4.7.2** (Upper Bound on MPC Performance). *Suppose problem  $\mathcal{PE}$  is feasible. Recall our definition for  $\mathcal{X}_N$  as the set of initial states for which problem  $\mathcal{MPC}$  is feasible. Then for all  $x_0 \in \mathcal{X}_N$ , the value  $J_0^*(x_0)$  provides an upper bound for the infinite horizon cost under the MPC policy, that is*

$$J_0^*(x_0) \geq \limsup_{k \rightarrow \infty} \rho_{0,k} (C(x_0, \pi^{\text{MPC}}(x_0)), \dots, C(x_k, \pi^{\text{MPC}}(x_k))) = \limsup_{k \rightarrow \infty} J_{0,k}(x_0, \pi^{\text{MPC}}).$$

## 4.8 Solution Algorithms

In this section we discuss two solution approaches, the first via dynamic programming, the second via convex programming.

### 4.8.1 Dynamic Programming Approach

While problem  $\mathcal{MPC}$  can be solved via dynamic programming (see [119, Theorem 2]), one would first need to find a matrix  $P$  that satisfies (7.73). Expression (7.73) is a bilinear semi-definite inequality in  $(P, F)$ . It is well known that feasibility checks in bilinear semi-definite inequality constraints is an NP-hard problem [150]. Nonetheless, one can transform this bilinear semi-definite inequality constraint into a linear matrix inequality by applying the Projection Lemma [135]. The next theorem presents a linear matrix inequality characterization of condition (7.73). Due to space limits, the proof of this theorem is included in Section 7.1 7.4.9.

**Theorem 4.8.1** (LMI Characterization of Stability Constraint). *Define  $\bar{A} = \begin{bmatrix} A_1^\top & \dots & A_L^\top \end{bmatrix}^\top$ ,  $\bar{B} = \begin{bmatrix} B_1^\top & \dots & B_L^\top \end{bmatrix}^\top$  and for each  $q_l \in \mathcal{U}^{\text{poly},V}(p)$  define  $\Sigma_l = \text{diag}(q_l(1)I, \dots, q_l(L)I) \succeq 0$ . Consider the following set of LMIs with decision variables  $Y, G, \bar{Q} = \bar{Q}^\top \succ 0$ :*

$$\begin{bmatrix} I_{L \times L} \otimes \bar{Q} & 0 & 0 & -\Sigma_l^{\frac{1}{2}}(\bar{A}G + \bar{B}Y) \\ * & R^{-1} & 0 & -Y \\ * & * & I & -Q^{\frac{1}{2}}G \\ * & * & * & -\bar{Q} + G + G^\top \end{bmatrix} \succ 0, \forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly},V}(p))\}. \quad (4.20)$$

The expression in (7.73) is equivalent to the set of LMIs in (4.20) by setting  $F = YG^{-1}$  and  $P = \bar{Q}^{-1}$ .

Furthermore, by the application of the Projection Lemma to the expressions in (7.70), (7.71) and (7.72), we obtain the following corollary:

**Corollary 4.8.2.** *Let  $Y$  and  $G$  be the decision variables in the set of LMIs in (4.20). Suppose the following set of LMIs with decision variables  $Y, G$ , and  $W = W^\top \succ 0$  are satisfied:*

$$\begin{aligned} & \begin{bmatrix} x_{\max}^2 I & -T_x(A_j G + B_j Y) \\ * & -W + G + G^\top \end{bmatrix} \succ 0, \\ & \begin{bmatrix} a_{\max}^2 I & -T_a Y \\ * & -W + G + G^\top \end{bmatrix} \succ 0, \\ & \begin{bmatrix} W & -(A_j G + B_j Y) \\ * & -W + G + G^\top \end{bmatrix} \succ 0. \end{aligned} \quad (4.21)$$

Then, the LMIs in (7.70), (7.71) and (7.72) with  $F = YG^{-1}$  are also satisfied. That is, by setting  $F = YG^{-1}$ , the inequalities above represent sufficient conditions for the LMIs in (7.70), (7.71) and (7.72).

**Remark 4.8.3** (Projection Lemma with Non-strict Inequalities). *The LMIs in Corollary 4.8.2 represent sufficient conditions for the invariance of the set  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$  under the feasible local control law  $a(x) = Fx$ . In these LMIs, strict inequalities are imposed only for the sake of analytical simplicity when applying the Projection Lemma (Lemma 7.4.1). Using similar arguments as in [72], non-strict versions of the above LMIs may also be derived, for example, leveraging some additional technicalities, [124] presents conditions that extend the Projection Lemma to encompass non-strict inequalities.*

A solution approach for the receding horizon adaptation of problem  $\mathcal{OPT}_{\text{RS}}$  is to first solve the LMIs in Theorem 4.8.1 and Corollary 4.8.2. If a solution for  $(P, Y, G, W)$  is found, apply dynamic programming (after state and action *discretization*, see, e.g., [47, 45]). Note that the discretization process might yield a large-scale dynamic programming problem for which the computational complexity scales exponentially with the resolution of discretization. This motivates the convex programing approach presented next.



### 4.8.2 Convex Programming Approach

Consider the following parameterization of *history-dependent* control policies. Let  $j_0, \dots, j_h \in \{1, \dots, L\}$  be the realized indices for the disturbances in the first  $h + 1$  steps of the MPC problem, where  $h \in \{1, \dots, N - 1\}$ . The control to be exerted at stage  $h$  is denoted by  $\bar{a}_h(j_0, \dots, j_{h-1})$ . Similarly, we refer to the state at stage  $h$  as  $\bar{x}_h(j_0, \dots, j_{h-1})$ . The quantities  $\bar{x}_h(j_0, \dots, j_{h-1})$  and  $\bar{a}_h(j_0, \dots, j_{h-1})$  enable us to keep track of the growth of the scenario tree. In terms of this new notation, the system dynamics (4.1) can be rewritten as:

$$\begin{aligned} \bar{x}_0 &:= x_{k|k}, \bar{a}_0 \in \mathbb{A}, \bar{x}_1(j_0) = A_{j_0}\bar{x}_0 + B_{j_0}\bar{a}_0, \text{ for } h = 1, \\ \bar{x}_h(j_0, \dots, j_{h-1}) &= A_{j_{h-1}}\bar{x}_{h-1}(j_0, \dots, j_{h-2}) + B_{j_{h-1}}\bar{a}_{h-1}(j_0, \dots, j_{h-2}), \text{ for } h \geq 2. \end{aligned} \quad (4.22)$$

While problem MPC is defined as an optimization over *Markov* control policies, in the convex programming approach, we re-define the problem as an optimization over *history-dependent* policies. One can show (with a virtually identical proof, see Section 7.4.1.1 for more details) that the stability Theorem 4.6.6 still holds when history-dependent policies are considered. Furthermore, since Markov policies are optimal in our setup (see [119, Theorem 2]), the value of the optimal cost stays the same. The key advantage of history-dependent policies is that their additional flexibility leads to a convex optimization problem for the determination of the model predictive control law. This is illustrated by the following solution algorithm:

**Algorithm MPC** — Given an initial state  $x_0 \in \mathbb{X}$  and a prediction horizon  $N \geq 1$ , solve

- **Offline step:** Solve

$$\max_{W=W^\top \succ 0, G, Y, \bar{Q}=\bar{Q}^\top \succ 0} \log \det(W)$$

subjected to the LMIs in expressions (4.20) and (4.21).

- **Online Step:** Suppose the feasible set of solutions in the offline step is non-empty. Define:  $W = W^*$  and  $P = (\bar{Q}^*)^{-1}$  where  $W^*$  and  $\bar{Q}^*$  are the maximizers for the offline step. Now at each step  $k \in \{0, 1, \dots\}$ , solve:

$$\begin{aligned} \min_{\gamma_2(j_0, \dots, j_{N-1}), \bar{x}_h(j_0, \dots, j_{h-1}), \bar{a}_0, \bar{a}_h(j_0, \dots, j_{h-1}),} & \rho_{k, k+N}(C(x_{k|k}, \bar{a}_0), \dots, C(\bar{x}_{N-1}, \bar{a}_{N-1}), \gamma_2) \\ & h \in \{1, \dots, N\}, j_0, \dots, j_{N-1} \in \{1, \dots, L\} \end{aligned} \quad (4.23)$$

subject to

- the LMIs

$$\begin{bmatrix} 1 & \bar{x}_N(j_0, \dots, j_{N-1})^\top \\ * & W \end{bmatrix} \succeq 0; \quad (4.24)$$

$$\begin{bmatrix} \gamma_2(j_0, \dots, j_{N-1})I & \bar{x}_N(j_0, \dots, j_{N-1})^\top \\ * & P^{-1} \end{bmatrix} \succeq 0; \quad (4.25)$$

- the system dynamics in equation (4.22);
- the control constraints for  $h \in \{1, \dots, N\}$ :

$$\|T_a \bar{a}_0\|_2 \leq a_{\max}, \|T_a \bar{a}_h(j_0, \dots, j_{h-1})\|_2 \leq a_{\max}; \quad (4.26)$$

- the state constraints for  $h \in \{1, \dots, N\}$ :

$$\|T_x \bar{x}_h(j_0, \dots, j_{h-1})\|_2 \leq x_{\max}; \quad (4.27)$$

Then, set  $\pi^{MPC}(x_{k|k}) = \bar{a}_0$ .

Note that the terminal cost has been *equivalently* reformulated via the epigraph constraint in (4.25) using the variable  $\gamma_2$ . Details of this analysis can be found in Section 7.4.10. This algorithm is clearly suitable only for “moderate” values of  $N$ , given the combinatorial explosion of the scenario tree. As a degenerate case, when we exclude all lookahead steps, problem  $\mathcal{MPC}$  is reduced to an *offline* optimization. By trading off performance, one can compute the control policy offline and implement it directly online without further optimization:

**Algorithm  $\mathcal{MPC}^0$**  — Given an initial state  $x_0 \in \mathbb{X}$ , solve:

$$\min_{\gamma_2, W = W^\top \succ 0, G, Y, \bar{Q} = \bar{Q}^\top \succ 0} \gamma_2$$

subjected to the LMIs in expressions (4.20), (4.21) and

$$\begin{bmatrix} 1 & x_0^\top \\ * & W \end{bmatrix} \succeq 0, \quad \begin{bmatrix} \gamma_2 I & x_0^\top \\ * & \bar{Q} \end{bmatrix} \succeq 0.$$

Then, set  $\pi^{MPC}(x_k) = YG^{-1}x_k$ .

Note that the domain of feasibility for  $\mathcal{MPC}^0$  is the original control invariant set  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$ . Showing ULRSSES for algorithm  $\mathcal{MPC}^0$  is more straightforward than the corresponding analysis for problem  $\mathcal{MPC}$  and is summarized within the following corollary.

**Corollary 4.8.4** (Quadratic Lyapunov Function). *Suppose problem  $\mathcal{MPC}^0$  is feasible. Then, system (4.1) under the offline MPC policy:  $\pi^{MPC}(x_k) = YG^{-1}x_k$  is ULRSSES with domain  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$ .*

## 4.9 Numerical Experiments

In this section we present several numerical experiments that were run on a 2.3 GHz Intel Core i5, MacBook Pro laptop, using the MATLAB YALMIP Toolbox (version 2.6.3 [77]) with the SDPT3 solver. All measurements of computation time are given in seconds.

### 4.9.1 Effects due to Risk Aversion

Consider the stochastic system:  $x_{k+1} = A(w_k)x_k + B(w_k)u_k$ , where  $w_k \in \{1, 2, 3\}$  and

$$A_1 = \begin{bmatrix} 2 & 0.5 \\ -0.5 & 2 \end{bmatrix}, A_2 = \begin{bmatrix} 0.01 & 0.1 \\ 0.05 & 0.01 \end{bmatrix}, A_3 = \begin{bmatrix} 1.5 & -0.3 \\ 0.2 & 1.5 \end{bmatrix},$$

$$B_1 = \begin{bmatrix} 3 & 0.1 \\ 0.1 & 3 \end{bmatrix}, B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, B_3 = \begin{bmatrix} 2 & 0.3 \\ 0.3 & 2 \end{bmatrix}.$$

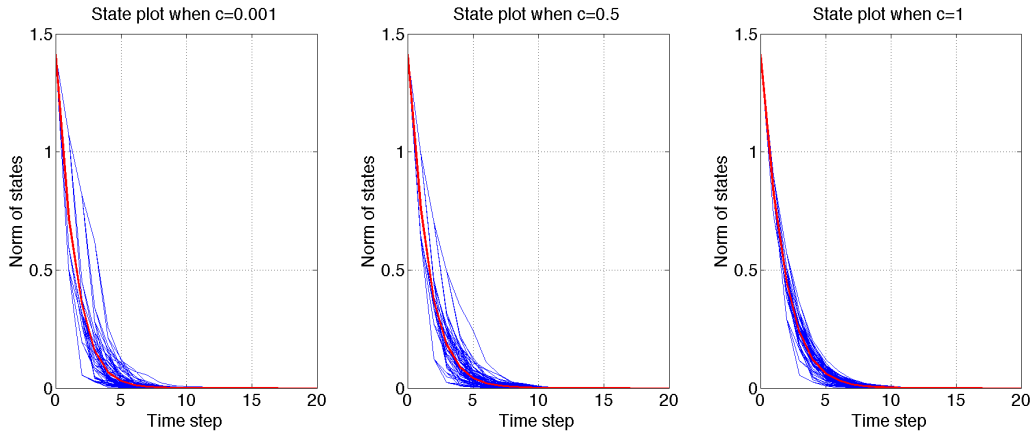


Figure 4.1: Effect of semi-deviation parameter  $c$

The probability mass function for the process disturbance is uniformly distributed, i.e.,  $\mathbb{P}(w_k = i) = 1/3$ , for  $i \in \{1, 2, 3\}$ . In this example, the goal is to explore the risk aversion capability of the risk-averse MPC algorithm presented in Section 4.6 (the solution relies on the convex programming approach). We consider as risk-aversion metric, the mean upper semi-deviation metric, where  $c$  ranges in the set  $\{0, 0.25, 0.5, 0.75, 1\}$ . The initial condition is  $x_0(1) = x_0(2) = 1$  and the number of lookahead steps is 3. We do not impose additional state and control constraints and set  $Q = I_{2 \times 2}$ ,  $R = 10^{-4}I_{2 \times 2}$ .

We performed 100 Monte Carlo simulations for each value of  $c$ . When  $c = 0$ , the problem reduces to risk-neutral minimization. On the other hand, one enforces maximum emphasis on regulating semi-deviation (dispersion) by setting  $c = 1$ . Table 4.1 and Figure 4.1 summarize our results. When  $c \approx 0$  (risk neutral formulation), the average cost is the lowest (with respect to the different choices for  $c$ ), but the dispersion is the largest. Conversely, when  $c = 1$ , the dispersion is the lowest, but the average cost is the largest. In the figure, it can be noted that the dispersion above the mean (given by the red curve) decreases as the value of  $c$  increases, as expected.

Table 4.1: Statistics for Risk-Averse MPC.

Level of Risk	Empirical Risk Cost	Dispersion	Standard Deviation	Mean (Variance) of Time per Itr.
c=0.001	2.5908	0.2813	0.6758	Offline: 0.3291 (0.0038) Online: 0.0699 (0.0033)
c=0.25	2.6910	0.2667	0.5210	Offline: 0.3761 (0.0073) Online: 0.0801 (0.0031)
c=0.5	2.8911	0.2271	0.4579	Offline: 0.4199 (0.0045) Online: 0.0865 (0.0029)
c=0.75	2.9310	0.1683	0.3877	Offline: 0.4249 (0.0071) Online: 0.0891 (0.0030)
c=1	3.0317	0.1145	0.2305	Offline: 0.4003 (0.0082) Online: 0.0903 (0.0034)

### 4.9.2 A 2-state, 2-input Stochastic System

Consider the following stochastic system with 6 scenarios:  $x_{k+1} = A(w)x_k + B(w)u_k$  where  $w \in \{1, 2, 3, 4, 5, 6\}$  and

$$\begin{aligned}
A_1 &= \begin{bmatrix} 2.0000 & 0.5000 \\ -0.5000 & 2.0000 \end{bmatrix}, A_2 = \begin{bmatrix} -0.1564 & -0.0504 \\ -0.0504 & -0.1904 \end{bmatrix}, A_3 = \begin{bmatrix} 1.5000 & -0.3000 \\ 0.2000 & 1.5000 \end{bmatrix}, \\
A_4 &= \begin{bmatrix} 0.5768 & 0.2859 \\ 0.2859 & 0.7499 \end{bmatrix}, A_5 = \begin{bmatrix} 1.8000 & 0.3000 \\ -0.2000 & 1.8000 \end{bmatrix}, A_6 = \begin{bmatrix} 0.2434 & 0.3611 \\ 0.3611 & 0.3630 \end{bmatrix}, \\
B_1 &= \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, B_2 = \begin{bmatrix} -0.9540 & 0 \\ -0.7773 & 0.1852 \end{bmatrix}, B_3 = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}, \\
B_4 &= \begin{bmatrix} -0.2587 & -0.9364 \\ 0.4721 & 0 \end{bmatrix}, B_5 = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}, B_6 = \begin{bmatrix} -1.6915 & 0 \\ 1.0249 & -0.3834 \end{bmatrix}.
\end{aligned}$$

The transition probabilities between the scenarios are uniformly distributed, i.e.,  $\mathbb{P}(w = i) = 1/6$ ,  $i \in \{1, 2, 3, 4, 5, 6\}$ . Clearly there exists a switching sequence such that this open loop stochastic system is unstable. The objectives of the model predictive controller are to 1) guarantee closed-loop ULRSES, 2) satisfy the control input constraints, with  $T_u = I_{2 \times 2}$ ,  $u_{\max} = 2.5$ , and 3) satisfy the state constraints, with  $T_x = I_{2 \times 2}$ ,  $x_{\max} = 5$ . The initial state is  $x_0(1) = x_0(2) = 2.5$ . The objective cost function follows expression (4.9), with  $Q = R = 0.01 \times I_{2 \times 2}$  and the one-step Markov polytopic risk metric is  $\text{CVaR}_{0.75}$ .

We simulated the state trajectories with 100 Monte Carlo samples, varying the number of lookahead steps  $N$  from 1 to 6, and compared the closed-loop performance from algorithms  $\mathcal{MPC}$  and  $\mathcal{MPC}^0$ . Since at every time step we can only access the realizations of the stochastic system in the current simulation, we cannot compare the performance of the model predictive controller with Problem  $\mathcal{OPT}$  exactly. Instead, for each simulation, the MPC algorithm was run until a stage  $k'$  such that  $\|x_{k'}\|_2 \leq x_{\max}$ . We then computed the *empirical* risk from all Monte Carlo simulations for a given horizon length using the cost function  $J_{0,k'}$ . Additionally, Theorem 4.7.2 shows that the MPC cost function evaluated at the first time step (i.e.,  $J_0^*(x_0)$  for Algorithm  $\mathcal{MPC}$ , and  $x_0^\top P^* x_0$  for Algorithm  $\mathcal{MPC}^0$ ) is an upper bound for the infinite horizon cost for

Table 4.2: Performance of Different Algorithms.

Algorithms	Empirical Risk Cost (Cost Upper Bound)	Mean (Variance) of Time per Itr.
$C - \mathcal{MPC}^0$	4.016 (4.983)	Offline: 0.3574 (0.0091) Online: 0 (0)
$C - \mathcal{MPC}, N = 1$	2.882 (3.481)	Offline: 0.3252 (0.0023) Online: 0.1312 (0.0032)
$C - \mathcal{MPC}, N = 2$	1.686 (2.288)	Offline: 0.3241 (0.0042) Online: 0.8214 (0.0256)
$C - \mathcal{MPC}, N = 3$	1.105 (1.525)	Offline: 0.3380 (0.0133) Online: 2.9984 (0.3410)
$C - \mathcal{MPC}, N = 4$	0.898 (1.063)	Offline: 0.3421 (0.0117) Online: 40.9214 (2.4053)
$C - \mathcal{MPC}, N = 5$	0.676 (0.794)	Offline: 0.3989 (0.0091) Online: 498.9214 (15.4921)
$C - \mathcal{MPC}, N = 6$	0.440 (0.487)	Offline: 0.4011 (0.0154) Online: 7502.90075 (98.4104)

Problem  $\mathcal{OPT}$  under the MPC policy. Thus, the performances of the MPC algorithms are evaluated based on the empirical risk and the upper bounds, summarized in Table 4.2.

Solving the MPC problem with more lookahead steps decreases the performance index ( $J_0^*(x_0)$ ), i.e., the sub-optimality gap of the MPC controller decreases. However, since the size of the online MPC problem scales exponentially with the number of lookahead steps, we can see that the online computation time scales exponentially from about 4 seconds at  $N = 3$  to over 7300 seconds at  $N = 6$ . Due to this drastic increase in computation complexity, we are only able to run 5 Monte Carlo trials for each case at  $N \in \{4, 5, 6\}$  for illustration. Note that the offline computation time is almost constant in all cases as the complexity of the offline problem is independent of the number of lookahead steps. Finally, using Lemma 4.7.1 we obtain a lower bound value of 0.1276 for the optimal solution of problem  $\mathcal{OPT}$ . The looseness in the sub-optimality gap may be attributed to neglecting stability and state/control constraint guarantees in the lower bound derivation.

### 4.9.3 Comparison with Bernadini and Bemporad's Algorithm [15]

In this experiment we compare the performance of algorithm  $\mathcal{MPC}$  with the risk-sensitive MPC algorithm in [15] (the problem formulation is given in Appendix 7.4.12). Define the following stochastic system

$$A_1 = \begin{bmatrix} -0.8 & 1 \\ 0 & 0.8 \end{bmatrix}, A_2 = \begin{bmatrix} -0.8 & 1 \\ 0 & 1.2 \end{bmatrix}, A_3 = \begin{bmatrix} -0.8 & 1 \\ 0 & -0.4 \end{bmatrix}, B_1 = B_2 = B_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

Table 4.3: Statistics for Risk-Averse MPC.

Method	Empirical Risk Cost	Standard Deviation	Mean (Variance) of Time per Itr.
Algorithm $\mathcal{MPC}$	82.9913	78.0537	Offline: 0.1448 (0.0299) Online: 10.8312 (2.3914)
MPC algorithm in [15]	90.3348	98.0748	Offline: 0.1309 (0.0099) Online: 13.5374 (4.0046)

where the initial state is  $x_0 = [5, 5]^\top$ , and uncertainty  $w_k$  is governed by an unknown probability mass function (different at each time step  $k$ ), which belongs to the set of distributions

$$\mathcal{M} = \{m = \delta_1[0.5, 0.3, 0.2] + \delta_2[0.1, 0.6, 0.3] + \delta_3[0.2, 0.1, 0.7] : [\delta_1, \delta_2, \delta_3] \in \mathcal{B}\}.$$

The cost matrices used in this test are  $Q = \begin{bmatrix} 1 & 0 \\ 0 & 5 \end{bmatrix}$  and  $R = 1$ . The state constraint matrix and threshold are given by  $T_x = I_{2 \times 2}$ ,  $x_{\max} = 12$ , and the control constraint matrix and threshold are given by  $T_u = 1$ ,  $u_{\max} = 2$ . While the MPC algorithm in [15] implemented scenario tree optimization techniques to reduce numerical complexity (to less than 20 nodes in their example), it is beyond the scope of this paper. For this reason, we choose  $N = 3$  (giving 27 leaves in the scenario tree) to ensure that the above problems have similar online complexity. Table 4.3 shows the results from 100 Monte Carlo trials. Due to the additional complexity of the LMI conditions in algorithm  $\mathcal{MPC}$ , the offline computation time for our algorithm is slightly longer. Nevertheless, the resulting policy yields a lower empirical risk (the one-step dynamic coherent risk in this example is defined as a distributionally robust expectation operator over the set distributions  $\mathcal{M}$ , i.e.  $\rho(Z) = \max_{m \in \mathcal{M}} \mathbb{E}_m[Z]$ ), lower standard deviation, and has a shorter online computation time as compared with its counterpart in [15]. This clearly demonstrates the advantages of using our risk averse approach to MPC.

#### 4.9.4 Safety Brake in Adaptive Cruise Control

Adaptive cruise control (ACC) [76, 28] extends the functionalities of conventional cruise control. In addition to tracking the reference velocity of the driver, ACC also enforces a separation distance between the leading vehicle (the host) and the follower (the vehicle that is equipped with the ACC system) to improve passenger comfort and safety. This crucial safety feature prevents a car crash when the host stops abruptly due to unforeseeable hazards.

In this experiment, we design a risk-sensitive controller for the ACC system that guarantees a safe separation distance between vehicles even when the host stops abruptly. As a prediction model for the MPC control problem, we define  $v_k$  and  $a_k$  to be the speed and the acceleration of the follower respectively, and  $v_{l,k}$ ,  $a_{l,k}$  as the velocity and acceleration of the leader. The acceleration  $a_k$  is modeled as the integrator

$$a_{k+1} = a_k + T_s u_k,$$

where  $T_s$  is the sampling period, and the control input  $u_k$  is the rate of change of acceleration (jerk) which is assumed to be constant over the sampling interval. The leader and follower velocities are given by

$$v_{k+1} = v_k + T_s a_k, \quad v_{l,k+1} = w_k v_{l,k},$$

where  $w_k$  is the leader's geometric deceleration rate. Since this rate captures the degree of abrupt stopping, its evolution has a stochastic nature. Here we assume  $w_k$  belongs to the sample space  $\mathcal{W} = \{0.5, 0.7, 0.9\}$  whose transition follows a uniform distribution. Furthermore, the distance  $d_k$  between the leader and the follower evolves as

$$d_{k+1} = d_k + T_s(v_{l,k} - v_k).$$

In order to ensure safety, we also set the reference distance to be velocity dependent, which can be modeled as  $d_{\text{ref},k} = \delta_{\text{ref}} + \gamma_{\text{ref}} v_k$  with  $\delta_{\text{ref}} = 4m$  and  $\gamma_{\text{ref}} = 3s$ . Together, the system dynamics are modeled by  $x_{k+1} = A(w_k)x_k + B(w_k)u_k$ , where  $w_k \in \{1, 2, 3\}$ ,  $x_k = [d_k - d_{\text{ref},k}, v_k, a_k, v_{l,k}]$ , and

$$A(w_k) = \begin{bmatrix} 1 & -T_s & -\gamma_{\text{ref}}T_s & T_s \\ 0 & 1 & T_s & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & w_k \end{bmatrix}, \quad B(w_k) = \begin{bmatrix} 0 \\ 0 \\ T_s \\ 0 \end{bmatrix}.$$

In order to guarantee comfort and safety, we assume the constraints  $|u_k| \leq 3ms^{-3}$  (bounded jerk), and  $|v_k| \leq 12ms^{-1}$  (bounded speed), and the state and control weighting matrices within the quadratic cost are  $Q = \text{diag}(Q_d, Q_v, 0, 0)$ ,  $R = Q_u$ , where  $Q_d$ ,  $Q_v$ ,  $Q_u$  are the weights on the separation distance tracking error, velocity, and jerk, respectively. To study the risk-averse behavior of the safety brake mechanism, we design a risk-sensitive MPC controller based on the dynamic risk compounded by the mean absolute semi-deviation with  $c = 1$ . For demonstrative purposes, the MPC lookahead step is simply set to one ( $N = 1$ ). The performance of the risk sensitive ACC system is illustrated by the state trajectories in Figure ?? . It can be seen that the controller is able to stabilize the stochastic error  $d_k - d_{\text{ref},k}$  (in the risk-sensitive sense) such that the speed of the follower vehicle gradually vanishes, and the separation distance  $d_k$  between the two cars converges to the constant  $\delta_{\text{ref}}$ . Notice that besides error tracking, the dynamic mean semi-deviation risk sensitive objective function also regulates the variability of distance separation of the follower vehicle, as shown in Table 4.4. Compared with the risk-neutral MPC approach, this risk-sensitive ACC system results in a lower variance in jerk and separation distance, suggesting a more comfortable passenger experience.

## 4.10 Conclusion

In this chapter we presented a framework for risk-averse MPC. Advantages of this framework include: (1) it is axiomatically justified; (2) it is amenable to dynamic and convex optimization; and (3) it is general, in that it captures a full range of risk assessments from risk-neutral to worst case (given the generality of Markov

Table 4.4: Statistics for Risk-Sensitive ACC System (with Mean Absolute Semi-deviation Risk).

Method	Mean Cost	Standard Deviation	Mean (Variance) of Time per Itr.
$c = 1$	451.3442	9.5854	Offline: 0.3944 (0.0036) Online: 0.0727 (0.0054)
$c = 0$ (Risk Neutral)	423.8701	12.7447	Offline: 0.4011 (0.0024) Online: 0.0450 (0.0067)

polytopic risk metrics).

In the next chapter we will further extend the risk-constrained optimal control framework to include multi-stage constraints that are induced by time-consistent, Markov coherent risk measures.



## Chapter 5

# Stochastic Optimal Control with Dynamic Risk Constraints

### 5.1 Introduction

#### 5.1.1 An Overview on Constrained Stochastic Optimal Control

Constrained stochastic optimal control problems naturally arise in several domains, including engineering, finance, and logistics. For example, in a telecommunication setting, one is often interested in the maximization of the throughput of some traffic subject to constraints on delays [4, 71], or seeks to minimize the average delays of some traffic types, while keeping the delays of other traffic types within a given bound [97]. Arguably, the most common setup is the optimization of a *risk-neutral expectation* criterion subject to a *risk-neutral* constraint [40, 104, 41]. This model, however, is not suitable in scenarios where risk-aversion is a key feature of the problem setup. For example, financial institutions are interested in trading assets while keeping the *riskiness* of their portfolios below a threshold; or, in the optimization of rover planetary missions, one seeks to find a sequence of divert and driving maneuvers so that the rover drive is minimized and the *risk* of a mission failure (e.g., due to a failed landing) is below a user-specified bound [74].

A common strategy to include risk-aversion in constrained problems is to have constraints where a static, single-period risk metric is applied to the future stream of costs; typical examples include variance-constrained stochastic optimal control problems (see, e.g., [104, 136, 82]), or problems with probability constraints [40, 104]. However, using static, single-period risk metrics in multi-period decision processes can lead to an over or under-estimation of the true dynamic risk, as well as to a potentially “inconsistent” behavior (whereby risk preferences change in a seemingly irrational fashion between consecutive assessment periods), see Section 1.3 and references therein. In [118], the authors provide an example of a portfolio selection problem where the application of a static risk metric in a multi-period context leads a risk-averse

decision maker to (erroneously) show risk neutral preferences at intermediate stages.

Indeed, in the recent past, the topic of *time-consistent* risk assessment in multi-period decision processes has been heavily investigated [120, 122, 121, 119, 129, 131, 130, 42, 1]. The key idea behind time consistency is that if a certain outcome is considered less risky in all states of the world at stage  $k$ , then it should also be considered less risky at stage  $k$  [62]. Remarkably, in [119], it is proven that any risk measure that is time consistent can be represented as a composition of one-step conditional risk mappings, in other words, in multi-period settings, risk (as expected) should be compounded over time.

### 5.1.2 Chapter Contribution

Despite the widespread usage of constrained stochastic optimal control and the significant strides in the theory of dynamic, time-consistent risk metrics, their integration within constrained stochastic optimal control problems has received little attention. The purpose of this chapter is to bridge this gap. Specifically, the contribution is threefold. First, equipped with the notion of dynamic, time-consistent risk metrics in Section 1.3, we formulate a risk constrained MDP problem whose constraint is modeled by such risk metric. Second, we develop a dynamic programming approach for the solution, which allows efficient computation of the optimal costs by value iteration. There are two main reasons behind our choice of a dynamic programming approach: (a) the dynamic programming approach can be used as an analytical tool in special cases and as the basis for the development of either exact or approximate solution algorithms; and (b) in the risk-neutral setting (i.e., both objective and constraints given as expectations of the sum of stage-wise costs) the dynamic programming approach appears numerical convenient with respect to other approaches (e.g., with respect to the convex analytic approach [4]) and allows to build all (Markov) optimal control strategies [104]. While the dynamic programming algorithm provides a theoretically sound methodology to tackle the risk constrained MDP problem, its implementation presents several computation challenges. Thus third, we present two approximate dynamic programming solution approaches to (approximately) solve the risk constrained MDP problems for medium to large scale systems.

### 5.1.3 Chapter Organization

The rest of the chapter is structured as follows. In Section 5.2 we formulate the problem we wish to solve as a risk constrained MDP problem, while in Section 5.3 we propose a dynamic programming approach for computing the solution (the value function of the risk constrained MDP problem) and a procedure to construct optimal policies. In Section 5.4, we develop two novel approximate dynamic programming approaches to solve for the value function and provide the corresponding error bound guarantees. All technical results of this chapter will be given in Section 7.5.

## 5.2 Problem Formulation

In this section we formally state the problem we wish to solve. According to Section 1.2, consider the CMDP framework (a CMDP with finite state and action spaces)  $(\mathcal{X}, \mathcal{A}, C, D, P, \gamma, x_0, d_0)$ . Given a policy  $\pi \in \Pi_H$ , an initial state  $x_0 \in \mathcal{X}$ , the cost function is defined as

$$\mathcal{C}^\pi(x_0) := \lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=0}^{N-1} \gamma^t C(x_t, a_t) \mid x_0, \pi \right]$$

and the risk constraint is defined as

$$\mathcal{D}^\pi(x_0) := \lim_{N \rightarrow \infty} \rho_{0, N-1} \left( D(x_0, a_0), \dots, \gamma^{N-1} D(x_{N-1}, a_{N-1}) \right) \mid x_0, \pi,$$

where for  $N \in \mathbb{N}$ ,  $\rho_{0, N-1}(\cdot) = \underbrace{\rho \circ \rho \circ \dots \circ \rho}_{N}(\cdot)$  is a Markov risk measure (see Section 1.3 for more details).

The problem we wish to solve is then as follows:

**Optimization problem  $\mathcal{OPT}_{\text{RC}}$**  — Given an initial state  $x_0 \in \mathcal{X}$  and a risk threshold  $d_0 \in \mathbb{R}$ , solve

$$\begin{aligned} \min_{\pi \in \Pi_H} \quad & \mathcal{C}^\pi(x_0) \\ \text{subject to} \quad & \mathcal{D}^\pi(x_0) \leq d_0. \end{aligned}$$

If problem  $\mathcal{OPT}_{\text{RC}}$  is not feasible, we say that its value is  $\infty$ . Note that, when the problem is feasible, an optimal policy always exists since the state and control spaces are finite. When  $\rho$  is replaced by an expectation, we recover the usual risk-neutral constrained stochastic optimal control problem studied, e.g., in [40, 104]. In the next section we present a dynamic programming approach to solve problem  $\mathcal{OPT}_{\text{RC}}$ .

To characterize the value function of problem  $\mathcal{OPT}_{\text{RC}}$ , we first define the (non-empty) set of feasible constraint thresholds at state  $x \in \mathcal{X}$  as  $\Phi(x) := [\underline{d}(x), \bar{d}]$ . Here the minimum risk-to-go for each state  $x \in \mathcal{X}$  is given by  $\underline{d}(x) := \min_{\pi \in \Pi_H} \mathcal{D}^\pi(x)$ . Since  $\{\rho_{k, N-1}\}_{k=0}^{N-1}$  is a Markov risk measure for all  $N \in \mathbb{N}$ ,  $\underline{d}(x)$  can be computed by using a dynamic programming recursion (see Theorem 2 in [119]). The function  $\underline{d}(x)$  is clearly the lowest value for a feasible constraint threshold. On the other hand, to characterize the upper bound, let:

$$\rho_{\max} := \max_{(x, a) \in \mathcal{X} \times \mathcal{A}} \rho(D(x, a)).$$

By the monotonicity and translation invariance of Markov risk measures, one can easily show that

$$\max_{\pi \in \Pi_H} \mathcal{D}^\pi(x) \leq \frac{\rho_{\max}}{1 - \gamma} := \bar{d}, \quad \forall x \in \mathcal{X}.$$

Therefore, the value functions are then defined as follows:

- If  $d \in \Phi(x)$ :

$$\begin{aligned} V^*(x, d) &= \min_{\pi \in \Pi_H} \mathcal{C}^\pi(x) \\ &\text{subject to } \mathcal{D}^\pi(x) \leq d; \end{aligned}$$

the minimum is well-defined since the state and control spaces are finite.

- If  $d \notin \Phi(x)$ :  $V^*(x, d) = \infty$ .

For  $(x, d) = (x_0, d_0)$ , we recover the definition of problem  $\mathcal{OPT}_{\text{RC}}$ . Notice that the size of the feasibility region of the above optimization problem is inversely proportional to the constraint threshold  $d$ . One immediate observation to the value function  $V^*(x, d)$  is its non-increasing property in  $d \in \mathbb{R}$  for any given initial state  $x \in \mathcal{X}$ .

### 5.3 A Dynamic Programming Algorithm for Risk-Constrained Multi-Stage Decision-Making

In this section we present the Bellman optimality condition of problem  $\mathcal{OPT}_{\text{RC}}$  and discuss a dynamic programming approach to solve problem  $\mathcal{OPT}_{\text{RC}}$ .

#### 5.3.1 Dynamic Programming Recursion

In this section we prove that the value functions can be computed by dynamic programming. Let  $B(\mathcal{X})$  denote the Borel space of real-valued bounded functions on  $\mathcal{X}$ , and  $B(\mathcal{X} \times \mathbb{R})$  denote the space of real-valued bounded functions on  $\mathcal{X} \times \mathbb{R}$ . Now we define the dynamic programming operator  $\mathbf{T}[V] : B(\mathcal{X} \times \mathbb{R}) \mapsto B(\mathcal{X} \times \mathbb{R})$  according to the equation:

$$\mathbf{T}[V](x, d) := \inf_{(a, d') \in F(x, d)} \left\{ C(x, a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a) V(x', d'(x')) \right\}, \quad (5.1)$$

where  $F \subset \mathbb{R} \times B(\mathcal{X})$  is the set of control/threshold functions:

$$F(x, d) := \left\{ (a, d') \mid a \in \mathcal{A}(x), d'(x') \in \Phi(x') \text{ for all } x' \in \mathcal{X}, \text{ and } D(x, a) + \gamma \rho(d'(x')) \leq d \right\}.$$

If  $F(x, d) = \emptyset$  we set  $\mathbf{T}[V](x, d) = \infty$ . Note that  $d \in \Phi(x)$  implies that  $F(x, d)$  is non-empty; likewise,  $d \notin \Phi(x)$  implies that  $F(x, d)$  is empty (these facts can be easily proven by contradiction).

For a given state and threshold constraint, set  $F$  characterizes the set of feasible pairs of actions and subsequent constraint thresholds. Feasible subsequent constraint thresholds are thresholds which if satisfied at the next stage ensure that the current state satisfies the given constraint threshold. Note that equation (5.1)

involves a functional minimization over the space  $B(\mathcal{X})$ . Indeed, since  $\mathcal{X}$  is finite,  $B(\mathcal{X})$  is isomorphic with  $\mathbb{R}^{|\mathcal{X}|}$ , hence the minimization in equation (5.1) can be re-casted as a regular (although possibly large) optimization problem in the Euclidean space. Computational aspects are further discussed at the end of this section. Also note that the value functions are defined on an *augmented* state space, which combines the original (discrete) states  $x$  with the *real-valued* threshold states  $d$ . We will refer to the MDP problem associated with such augmented state space as augmented MDP (AMDP). We start by stating a number of useful properties for the dynamic programming operator in equation (5.1).

**Lemma 5.3.1.** *Let  $V$  and  $\tilde{V}$  be functions belonging to  $B(\mathcal{X} \times \mathbb{R})$ , and  $\mathbf{T}[V] : B(\mathcal{X} \times \mathbb{R}) \mapsto B(\mathcal{X} \times \mathbb{R})$  be the dynamic programming operator in equation (5.1). Then, the following statements hold:*

1. **Monotonicity:** *For any  $(x, d) \in \mathcal{X} \times \mathbb{R}$ , if  $V \leq \tilde{V}$ , then  $\mathbf{T}[V](x, d) \leq \mathbf{T}[\tilde{V}](x, d)$ .*
2. **Constant shift:** *For any real number  $L$  and  $(x, d) \in \mathcal{X} \times \mathbb{R}$ ,  $\mathbf{T}[V + K](x, d) = \mathbf{T}[V](x, d) + K$ , where  $(V + K)(x, d) := V(x, d) + K$ ,  $\forall (x, d) \in \mathcal{X} \times \mathbb{R}$ .*
3. **Contraction:** *For all  $V, \tilde{V} \in B(\mathcal{X} \times \mathbb{R})$ ,  $\|\mathbf{T}[V] - \mathbf{T}[\tilde{V}]\|_\infty \leq \gamma \|V - \tilde{V}\|_\infty$ , where  $\|\cdot\|_\infty$  denotes the infinity norm.*

The proof of these properties is standard in the dynamic programming literature and we refer interesting readers to [17] for more details. We are now in a position to state the first main result of this chapter on Bellman's equation.

**Theorem 5.3.2** (Bellman's equation with risk constraints). *Assume that, when the optimization problem in equation (5.1) is feasible (i.e.,  $F(x, d) \neq \emptyset$ ), the infimum is attained. Then, the value function  $V^*$  is the unique solution of the Bellman's equation:*

$$V(x, d) = \mathbf{T}[V](x, d), \quad \forall (x, d) \in \mathcal{X} \times \mathbb{R}.$$

**Remark 5.3.3** (On the assumption in Theorem 5.3.2). *In Theorem 5.3.2 we assume that the infimum in equation (5.1) is attained. This is indeed always true in our setup, where, in particular, we assume a finite state space and a finite control space. The proof of this result would be almost identical to the proof of Lemma 5 in [41] and is omitted in the interest of brevity.*

**Remark 5.3.4** (On alternative solution approaches). *In principle, problem  $\text{OPT}_{\text{RC}}$  could also be solved by transforming it into an unconstrained optimization problem via, for example, logarithmic barrier functions. However, the cost function in the unconstrained problem would not have any obvious “compositional” structure, and its minimization would be particularly challenging (e.g., a direct dynamic programming approach would not be, in general, applicable).*

**Remark 5.3.5** (Computational issues). *In our approach, the solution of problem  $\text{OPT}_{\text{RC}}$  entails the solution of two dynamic programming problems, the first one to find the lower bound for the set of feasible constraint*

thresholds (i.e., the function  $\underline{d}(x)$ ), and the second one to compute the value functions  $V(x, d)$ . The latter problem is the most challenging one since it involves a functional minimization. However, as already noted, since  $\mathcal{X}$  is finite,  $B(\mathcal{X})$  is isomorphic with  $|\mathcal{X}|$ , and the functional minimization in the Bellman's operator (5.1) can be re-casted as an optimization problem in the Euclidean space.

### 5.3.2 Construction of optimal policies

In this section we present a procedure to construct optimal policies. Under the assumptions of Theorem 5.3.2, for any given  $x \in \mathcal{X}$  and  $d \in \Phi(x)$  (which implies that  $F(x, d)$  is non-empty), let  $u^*(x, d)$  and  $d'^*(x, d)(\cdot)$  be the minimizers in equation (5.1). By letting  $(x, d)$  as an augmented state (in state space  $\mathcal{X} \times \mathbb{R}$ ), here we notice that  $u^*$  is an *augmented Markovian stationary policy*. Furthermore  $d'^*(x, d)(\cdot)$  is the "optimal" constraint threshold in the next stage (starting at state  $x$  with constraint threshold  $d$ ), and is therefore denoted as the *risk-to-go*. Next theorem shows how to construct optimal policies.

**Theorem 5.3.6** (Optimal policies). *Let  $\pi_H^* = \{\mu_0, \mu_1, \dots\} \in \Pi_H$  be a history-dependent policy recursively defined as:*

$$\mu_k(h_k) = u^*(x_k, d_k), \quad \forall k \geq 0, \quad (5.2)$$

*with initial conditions  $x_0$  and  $y_0 = \alpha$ , and state transitions*

$$\begin{aligned} x_k &\sim P(\cdot \mid x_{k-1}, u^*(x_{k-1}, d_{k-1})), \\ d_k &= d'^*(x_{k-1}, d_{k-1})(x_k), \quad \forall k \geq 1, \end{aligned} \quad (5.3)$$

*Then,  $\pi_H^*$  is an optimal policy for problem  $\mathcal{OPT}_{RC}$  with initial state  $x_0$  and constraint threshold  $d_0 \in \Phi(x_0)$ .*

Interestingly, if one views the constraint thresholds as state variables (whose dynamics are given in the statement of Theorem 5.3.6), the optimal (history-dependant) policies of problem  $\mathcal{OPT}_{RC}$  have a Markovian structure with respect to the augmented control problem.

## 5.4 Discretization/Interpolation Algorithms for AMDP

According to Theorem 5.3.2, problem  $\mathcal{OPT}$  can be (formally) solved using value iteration on an augmented state space. However, the "threshold state"  $d$  appearing in the value function  $V(x, d)$  is a continuous, real-valued variable. This requires the design of discretization/sampling algorithms in order to carry out such value iteration in practice. Our approach is to extend the uniform-grid discretization approximation developed in [45] and the linear interpolation approach developed in Section 2.4.

### 5.4.1 Discretization Algorithm

For any state  $x \in \mathcal{X}$ , we partition  $\Phi(x)$  with a discretization step  $\Delta$  into  $\Theta + 1$  intervals using  $\Theta$  grid points  $\{d^{(1)}, \dots, d^{(\Theta)}\}$  (clearly,  $\Theta$  depends on  $x$ , we omit this dependency for notational simplicity). For

$\theta \in \{0, \dots, \Theta\}$ , define the discretized region  $\Phi^{(\theta)}(x) = [d^{(\theta)}, d^{(\theta+1)})$ , where  $d^{(0)} = \underline{R}(x)$  and  $d^{(\Theta+1)} = \bar{d} + \epsilon$ , for arbitrarily small  $\epsilon > 0$ . We also define  $\bar{\Phi}(x) := \{d^{(0)}, \dots, d^{(\Theta+1)}\}$ . Let  $\theta \in \{0, \dots, \Theta\}$  such that  $d \in \Phi^{(\theta)}(x)$ . Now, define the approximation operator  $\mathbf{T}_{\mathcal{D}}$  for  $x \in \mathcal{X}$ ,  $d \in \Phi^{(\theta)}(x)$  according to:

$$\mathbf{T}_{\mathcal{D}}[V](x, d) := \mathcal{T}_{\mathcal{D}}[V](x, d^{(\theta)}), \quad (5.4)$$

where

$$\mathcal{T}_{\mathcal{D}}[V](x, d) := \min_{(a, d'_{\mathcal{D}}) \in F_{\mathcal{D}}(x, d)} \left\{ C(x, a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a) V(x', d'_{\mathcal{D}}(x')) \right\}, \quad (5.5)$$

and where  $F_{\mathcal{D}}$  is the set of control/threshold functions:

$$F_{\mathcal{D}}(x, d) := \left\{ (a, d'_{\mathcal{D}}) \mid a \in \mathcal{A}(x), d'_{\mathcal{D}}(x') \in \bar{\Phi}(x') \text{ for all } x' \in \mathcal{X}, \text{ and } D(x, a) + \gamma \rho(d'_{\mathcal{D}}(x')) \leq d \right\}.$$

If  $F_{\mathcal{D}}(x, d) = \emptyset$ , then  $\mathcal{T}_{\mathcal{D}}[V^*](x, d) = \infty$ .

By construction, any optimal solution of  $\mathbf{T}_{\mathcal{D}}[V](x, d)$  is a feasible solution for the dynamic programming equation in  $\mathbf{T}[V](x, d)$  (since  $F_{\mathcal{D}}(x, d) \subseteq F(x, d)$  and  $d^{(\theta)} \leq d$ ). Because  $F_{\mathcal{D}}(x, d)$  is a finite set, the minimization in  $\mathbf{T}_{\mathcal{D}}[V](x, d)$  is always attained. One can also readily show that the dynamic programming operator  $\mathbf{T}_{\mathcal{D}}[V]$  also satisfies the properties in Lemma 5.3.1. In the next subsection we will derive a bound for  $\|\mathbf{T}[V](x, d) - \mathbf{T}_{\mathcal{D}}[V](x, d)\|_{\infty}$ ; in particular, we will show that this bound converges to zero as the discretization step converges to zero, and that the convergence is linear in the step size.

#### 5.4.1.1 Error bound analysis

The error bound analysis for the above discretization algorithm relies on two Lipschitz-like assumptions.

**Assumption 5.4.1.** For any  $x \in \mathcal{X}$ ,  $a, \tilde{a} \in \mathcal{A}(x)$ , there exists  $M_C, M_D > 0$  such that

$$|C(x, a) - C(x, \tilde{a})| \leq M_C |a - \tilde{a}|, \quad |D(x, a) - D(x, \tilde{a})| \leq M_D |a - \tilde{a}|.$$

**Assumption 5.4.2.** For any  $a, \tilde{a} \in \mathcal{A}(x)$ , there exists  $M_P > 0$  such that

$$\sum_{x' \in \mathcal{X}} |P(x'|x, a) - P(x'|x, \tilde{a})| \leq M_P |a - \tilde{a}|.$$

The first assumption is rather mild, while the second assumption is more restrictive. Note, however, that this is a typical “regularity” assumption for discretization algorithms for stochastic optimal control [45].

First, we have the following technical lemma on the Lipschitz-ness of set-valued mapping  $F(x, d)$ .

**Lemma 5.4.3.** For every given  $x \in \mathcal{X}$  and  $\tilde{d}, d \in \Phi(x)$ , suppose Assumptions (7.3.2) to (5.4.2) hold. Also, define  $\underline{d}' := \{d'(x')\}_{x' \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}$  and  $\underline{\tilde{d}}' := \{\tilde{d}'(x')\}_{x' \in \mathcal{X}} \in \mathbb{R}^{|\mathcal{X}|}$ . If  $F(x, d)$  and  $F(x, \tilde{d})$  are non-empty

sets, then for any  $(a, \underline{d}') \in F(x, d)$ , there exists  $(\hat{a}, \hat{\underline{d}}') \in F(x, \tilde{d})$  such that for some  $M_R > 0$ ,

$$|a - \hat{a}| + \sum_{x' \in \mathcal{X}} |d'(x') - \hat{d}'(x')| \leq M_R |d - \tilde{d}|. \quad (5.6)$$

The following theorem is the main result of this paper. It provides an error bound between the value function  $V^*(x, d)$  and the discretized value function  $V_{\mathcal{D}}^*(x, d)$ , defined as the unique fixed point solution of discretized bellman's equation:

$$V_{\mathcal{D}}^*(x, d) := \mathbf{T}_{\mathcal{D}}[V_{\mathcal{D}}^*](x, d), \quad \forall x, d.$$

**Theorem 5.4.4.** *Suppose Assumptions (7.3.2) and (5.4.2) hold. Then,*

$$\|V_{\mathcal{D}}^* - V^*\|_{\infty} \leq \frac{1 + \gamma}{1 - \gamma} \left( \frac{M_C}{1 - \gamma} + \frac{M_P C_{\max}}{(1 - \gamma)^2} \right) M_R \Delta,$$

where  $M_R$  is the constant defined in inequality (5.6) and  $\Delta$  is the discretization step size.

Clearly, Theorem 5.4.4 implies that, as the step size  $\Delta \rightarrow 0$ , for any  $x \in \mathcal{X}$  and  $d \in \Phi(x)$  one has  $V_{\mathcal{D}}^*(x, d) \rightarrow V^*(x, d)$ . Note that the convergence is linear in the step size, which is the same convergence rate for discretization algorithms for unconstrained dynamic programming operators [45].

**Remark 5.4.5.** *Similar to the multi-grid discretization approaches discussed in [45, 151, 58], the discretization algorithm in this paper suffers from the curse of dimensionality. In fact, suppose the number of discretization intervals is  $T$ . For each time horizon, the size of the state space for AMDP is  $|\mathcal{X}|T$ , and the size of the action space is  $|\mathcal{A}|T^{|\mathcal{X}|}$ , which is exponential in the size of the original state space. To alleviate this issue, one could use methods such as Branch and Bound or rollout algorithms to find the minimizers at each step, if upper/lower bounds for the value functions can be efficiently calculated.*

## 5.4.2 Interpolation Algorithm

In the last section we explored a discretization approach that approximates the dynamic programming algorithm given in Theorem 5.3.2. While this method is simple and intuitive, Theorem 5.4.4 shows that, in order to obtain an accurate estimate of the value function, one requires a high resolution grid space for the risk-to-go state. This potentially leads to an exponential increase in the size of state space (of the AMDP), and results in computational intractability. To circumvent this issue, in this section we propose an interpolation approach to approximate the dynamic programming algorithm. Indeed it can be showed that approximation by discretization is a special case of the approximation by linear interpolation with an integral constraint. Although theoretically the approximation error still grows linearly with the decrease of grid points, we later show that numerically the interpolation approach is way more efficient compared to the discretization approach.

Formally, let  $N(x)$  denote the number of interpolation points. For every  $x \in \mathcal{X}$ , denote by  $\bar{\Phi}(x)$  the set of interpolation points. We denote by  $\mathcal{I}_x[V](d)$  the linear interpolation of the function  $V(x, d)$  on these points,



**Algorithm 4** Value Iteration with Linear Interpolation1: **Given:**

- Set of interpolation points  $\bar{\Phi}(x)$  at every state  $x \in \mathcal{X}$ .
- Initial value function  $V_0(x, d)$ .

2: For  $k = 1, 2, \dots$ 

- For each  $x \in \mathcal{X}$  and each  $d \in \bar{\Phi}(x)$ , update the value function estimate as follows:

$$V_{k+1}(x, d) = \mathbf{T}_{\mathcal{I}}[V_k](x, d), \quad \forall k \geq 0$$

3: Set the converged value iteration estimate as  $V_{\mathcal{I}}^*(x, d)$ , for any  $x \in \mathcal{X}$ , and  $d \in \bar{\Phi}(x)$ .

i.e.,

$$\mathcal{I}_x[V](d) = V(x, d^{(\theta)}) + \frac{V(x, d^{(\theta+1)}) - V(x, d^{(\theta)})}{d^{(\theta+1)} - d^{(\theta)}}(d - d^{(\theta)}),$$

where  $d^{(\theta)} = \max \{d' \in \bar{\Phi}(x) : d' \leq d\}$  and  $d^{(\theta+1)}$  is the closest interpolation point such that  $d \in [d^{(\theta)}, d^{(\theta+1)}]$ .

We now define the *interpolated* Bellman operator  $\mathbf{T}_{\mathcal{I}}$  as follows:

$$\mathbf{T}_{\mathcal{I}}[V](x, y) = \min_{(a, d_{\mathcal{I}}) \in F(x, d)} \left\{ C(x, a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a) \mathcal{I}_{x'}[V](d'_{\mathcal{I}}(x')) \right\}, \quad (5.7)$$

Algorithm 1 presents value iteration with linear interpolation for problem  $\mathcal{OPT}_{\text{RC}}$ . The only difference between this algorithm and standard value iteration  $V_{k+1}(x, d) = \mathbf{T}[V_k](x, d)$  is the linear interpolation procedure described above. In the following, we show that Algorithm 4 converges by first showing that the useful properties such as contraction also hold for interpolated Bellman operator  $\mathbf{T}_{\mathcal{I}}$ .

**Lemma 5.4.6** (Properties of interpolated Bellman operator).  *$\mathbf{T}_{\mathcal{I}}[V]$  has the same properties of  $\mathbf{T}[V]$  as in Lemma 2.3.2, namely 1) monotonicity, 2) constant shift, and 3) contraction.*

The proof of this technical result is straightforward, and thus we refer interested readers to similar arguments in the proof of Lemma 5.3.1 for more details. Consequently, the contraction property in Lemma 5.4.6 guarantees that Algorithm 4 converges, i.e., there exists a value function  $V_{\mathcal{I}}^* \in \mathbb{R}^{|\mathcal{X}| \times \mathbb{R}}$  such that  $\lim_{N \rightarrow \infty} \mathbf{T}_{\mathcal{I}}^N[V_0](x, d) = V_{\mathcal{I}}^*(x, d)$ ,  $\forall d \in \bar{\Phi}(x)$ . In addition, the convergence rate is geometric and equals to  $\gamma$ .

The following theorem provides an error bound between approximate value iteration and exact value iteration problem  $\mathcal{OPT}_{\text{RC}}$  in terms of the interpolation resolution.

**Theorem 5.4.7** (Convergence and Error Bound). *Suppose Assumptions (7.3.2) and (5.4.2) hold. Then,*

$$V_{\mathcal{I}}^*(x, d) \leq V_{\mathcal{D}}^*(x, d), \quad \forall d \in \bar{\Phi}(x), \quad x \in \mathcal{X}, \quad \text{and} \quad \|V_{\mathcal{I}}^* - V^*\|_{\infty} \leq \frac{1 + \gamma}{1 - \gamma} \left( \frac{M_C}{1 - \gamma} + \frac{M_P C_{\max}}{(1 - \gamma)^2} \right) M_R \Delta,$$

where  $M_R$  is the constant defined in inequality (5.6) and  $\Delta$  is the discretization step size.

Theorem 5.4.7 shows that 1) the interpolation procedure is *consistent*, i.e., when the  $\Delta$  is the discretization step size is arbitrarily small, the approximation error tends to zero; and 2) the approximation error converges linearly with  $\Delta$ , which is similar to error of the discretization approach in Section 5.4.1. While we currently have no further proofs showing that the interpolation approach converges faster than the discretization approach, its superiority in error convergence rate will be later illustrated in the experiments.

## 5.5 Experiments

### 5.5.1 Curse of Dimensionality with Discretization Approach

Consider an MDP with 3 states ( $x \in \{1, 2, 3\}$ ), 2 available actions ( $u \in \{1, 2\}$ ), and discounting factor  $\gamma = 0.667$ . The costs for the objective and constraint functions are respectively given by:

$$\begin{bmatrix} C(1, 1) & C(1, 2) \\ C(2, 1) & C(2, 2) \\ C(3, 1) & C(3, 2) \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \\ 5 & 6 \end{bmatrix}, \quad \begin{bmatrix} D(1, 1) & D(1, 2) \\ D(2, 1) & D(2, 2) \\ D(3, 1) & D(3, 2) \end{bmatrix} = \frac{1}{10} \begin{bmatrix} 5 & 4 \\ 6 & 3 \\ 5 & 1 \end{bmatrix}.$$

The transition probabilities are given by:

$$P(x'|x, 1) = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.4 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}, \quad P(x'|x, 2) = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.2 & 0.3 & 0.5 \\ 0.3 & 0.4 & 0.3 \end{bmatrix}.$$

For any  $x_0 \in \mathcal{X}$  and  $d_0 \in \Phi(x_0)$ , the risk-constrained stochastic optimal control problem we wish to solve is as follows:

$$\begin{aligned} & \min_{\pi \in \Pi_H} \quad \lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=0}^{N-1} \gamma^t C(x_t, a_t) \right] \\ & \text{subject to} \quad \lim_{N \rightarrow \infty} \rho_{0, N-1} \left( D(x_0, a_0), \gamma D(x_1, a_1), \dots, \gamma^{N-1} D(x_{N-1}, a_{N-1}) \right) \leq d_0, \end{aligned}$$

where  $u_k = \pi_k(h_{0,k})$  for  $k \in \mathbb{N}$ ,  $\rho_{0, N-1}(Z_0, Z_1, \dots, Z_{N-1}) = Z_0 + \rho(Z_1 + \rho(Z_2 + \dots + \rho(Z_{N-1})))$ , and the one-step conditional risk measures are given by the mean upper semi-deviation risk metric:

$$\rho(V) = \mathbb{E}[V] + 0.2 \left( \mathbb{E}[[V - \mathbb{E}[V]]_+^2] \right)^{1/2}.$$

As discussed in Section 5.3, this problem can be cast as an AMDP. In light of Theorem 5.4.4, one can use equations (5.4) and (5.5) to approximate the value functions via value iteration. In this example, we discretize the risk threshold sets (that, for simplicity, are suitably adjusted to have the same length - this is always possible given their definition) into  $M$  regions, where

Table 5.1: Computation Times with Different Discretization Step Sizes.

M	Computation time for each horizon (in seconds)
5	20.5542
10	58.4712
20	104.0001
40	353.9901
60	900.5084
80	1751.7630
100	3306.3002
150	14572.6631

$$M = \{5, 10, 20, 40, 60, 80, 100, 150\}.$$

For the different step sizes, we obtain approximations of the value functions with various degrees of accuracies. Figure 5.1 shows the approximations of the value functions for the different step sizes. As expected (Therem 5.4.4), when the number of  $M$  increases, the approximated value functions converge towards the “true” value functions. Table 5.1 provides the computation times for our numerical experiments; one can note the exponential increase of computation time with respect to the discretization step size.

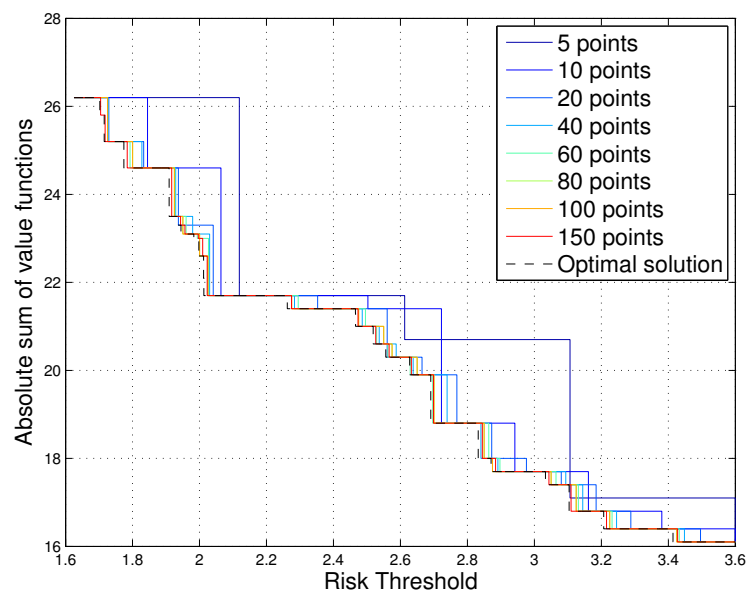


Figure 5.1: Convergence of Approximated Value Functions using Different Discretization Step Sizes.

## 5.6 Conclusion

In this chapter we have presented and analyzed both the uniform-grid discretization algorithm and the approximate value iteration algorithm for the solution of stochastic optimal control problems with dynamic, time-consistent risk constraints. Although the current uniform-grid discretization algorithm suffers from curse of dimensionality, it is the one of the simplest algorithms to practically solve this class of problems. On the other hand, based on approximate value-iteration on an augmented state space, we presented a tractable algorithm for solving the aforementioned risk-constrained control problem, whose convergence analysis and finite-time error bounds were explicitly studied. Furthermore, we concluded the algorithmic analysis of this section by drawing a connection between the uniform-grid discretization algorithm and approximate value iteration, where the former algorithm is indeed a special case (step-wise interpolation) of the general interpolation based method.

In the next chapter, we will summarize all the algorithms developed in this thesis and will discuss several interesting future research directions.

## Chapter 6

# Conclusion

In this thesis, we studied four important aspects regarding to the control of MDPs where risk modeling of the inherent uncertainty and model uncertainty was taken into account.

In the first part of this thesis, we investigated the well-known CVaR MDP problem and proposed a scalable approximate value-iteration algorithm on an augmented state space. We also established convergence guarantees and finite time error bounds, which led to a robust algorithmic stopping criterion with a specific error threshold. In addition, we discovered an interesting relationship between the CVaR risk of total cost and the worst-case expected cost under adversarial model perturbations. In this formulation, the perturbations were correlated in time, and led to a robustness framework significantly less conservative than the popular robust-MDP framework, where the uncertainty was temporally independent. Together, our work provided crucial theoretical underpinnings in CVaR MDPs that guaranteed efficient computation of robust control policies, with respect to cost stochasticity and model perturbations. Furthermore, to increase the practicality of our methodologies, we also proposed extensions such as the sampling based CVaR  $Q$ -learning algorithm and approximate value iteration for Mean-CVaR MDPs.

In the second part of this thesis, we studied the CMDP problem formulation whose constraints were modeled using percentile risks, such as CVaR and tail probabilities. We proposed novel policy gradient and actor-critic algorithms for CVaR-constrained and chance-constrained optimization in MDPs, and proved their convergence. These methodologies circumvented the curse-of-dimensionality issue and made real time computations of large scale risk sensitive CMDPs (whose state and action spaces are large or even continuous) possible. Using an optimal stopping problem and a personalized ad-recommendation problem, we showed that our algorithms resulted in policies whose cost distributions have lower right-tail compared to their risk-neutral counterparts. This was extremely important for a risk-averse decision-maker, especially if the right-tail contained catastrophic costs. Furthermore we also provided insights to the convergence of our AC algorithms when the samples were generated by following the policy and not from its discounted visiting distribution, and extensions via importance sampling [8, 146] where gradient estimates in the right-tail of the cost distribution (worst-case events that are observed with low probability) could significantly be improved.

In the third part of this thesis, we presented a framework for risk-averse MPC. Here the risk metric is chosen to be the time consistent, Markov risk. Advantages of this framework include: (1) it is axiomatically justified; (2) it is amenable to dynamic and convex optimization; and (3) it is general, in that it captures a full gamut of risk assessments from risk-neutral to worst case (given the generality of Markov polytopic risk metrics). For the solution algorithm, we have shown that the risk averse MPC framework can be posed as a hybrid (offline-online) convex optimization problem, which can be implemented using semi-definite programming techniques. We have also rigorously derived the performance of this MPC algorithm, in terms of sub-optimality gap, and numerically illustrated its superiority over its risk neutral counterpart given in [15].

In the final part of this thesis, we presented a dynamic programming approach to stochastic optimal control problems with dynamic, time-consistent (in particular Markov) risk constraints. In particular we showed that the optimal cost functions could be computed by value iteration and that the optimal control policies could be constructed recursively. Furthermore we proposed and analyzed a uniform-grid discretization algorithm for the solution of this stochastic optimal control problem. Although the current algorithm suffered from curse of dimensionality, it was by far the simplest approximation algorithm that practically solved this class of problems. As an improvement to the present approach, we also derived an interpolation based approximate dynamic programming algorithm which further simplified the risk-to-go estimation and led to lower approximation errors.

We hereby summarize the major conclusions of our work in this thesis, and discuss several important future extensions.

## 6.1 Inherent Uncertainty Versus Model Uncertainty

In this thesis, we have studied sequential planning problems in the presence of inherent uncertainty and model uncertainty and proposed algorithms to tackle scenarios with different sources of uncertainty. This was mainly motivated by the representation theorem of Markov coherent risk [119] that drew an equivalence between the dynamic risk measure and conditional worst-case expectation over model-perturbations. Earlier work that addressed model uncertainties in the context of risk-sensitive sequential decision making could be found in [95], in which the author related dynamic Markov coherent-risk to model uncertainties. In Chapter 2, we further extended this equivalence to connect model uncertainty to static CVaR risk, and we showed that such a risk metric represented a particular robust MDP with a coupled uncertainty structure (that could not be solved by conventional robust dynamic programming approaches). This suggested that many risk-sensitive reinforcement learning algorithms developed in this thesis were also applicable to tackle a more general class of robust MDP problems with coupled uncertainty structures.

## 6.2 Time Consistency in Risk-aware Planning

In this thesis, we have investigated the important aspect of time-consistency in risk-aware sequential decision making. While a common strategy to include risk-aversion is to have the objective function or constraints where a static risk metric is applied to the future stream of costs, it is also well-known that using static, single-period risk metrics in multi-period decision processes can lead to an inconsistent behavior where risk preferences change in a seemingly irrational fashion between consecutive assessment periods [62]. On the other hand, it has been shown in [119] that in order to guarantee *time-consistent* risk assessment in multi-period decision processes, risk must be compounded over time. Therefore, one major contribution of this thesis was to integrate time-consistent risk measures in sequential decision making. By exploiting the Markovian structure in the dynamic risk metrics, we successfully developed dynamic programming approaches for risk-sensitive and risk-constrained decision making problems. Not only did this dynamic programming approach effectively analyze risk-assessment across multiple decision-making time steps, it also led to the development of data-driven algorithms that effectively solves for (Markov) optimal control policies in large-scale problems.

## 6.3 Risk-shaping

In this thesis we have showed that by extending the studies of risk-neutral MDPs to include other risk-sensitive objective functions, one might effectively utilize the principle of dynamic programming to accomplish risk-sensitive planning. Furthermore by combining the aforementioned framework with data-driven decision making theories, we have also derived a family of risk-sensitive reinforcement learning algorithms that balances exploration, exploitation and safety. The main objectives of our work were to provide a theoretically sound formulation for risk-aware sequential decision making and to derive a set of computationally tractable tools for solving these MDP problems. Recall that the algorithms studied in this thesis were based on the class of (static and dynamic) coherent risk measures. This feature provided decision-makers with great flexibility to select objective functions that served multiple purposes in cost variability management. Evidently, this important phenomenon of selecting appropriate problem-specific risk measures to achieve effective risk aversion was corroborated by several numerical experiments in this thesis.

Moreover, through the connections of inherent uncertainty and model uncertainty, the decision-maker can easily handle uncertainties in model mis-specification by specifying the risk objective function in the MDP problem of interest. This suggests the potential application of *risk-shaping* —a technique that customizes risk measures to control variability —to model both systematic and cost uncertainties. While risk-shaping was not our major focus, the primary contribution of this thesis was to show that with appropriately chosen risk measures, the resultant risk-sensitive planning problem could be handled efficiently.

## 6.4 Future Work

We finally conclude this thesis with several additional important future research directions.

### 6.4.1 Exploration Versus Exploitation in Risk-sensitive Reinforcement Learning

In this thesis we have not explicitly studied the issue of balancing exploration and exploitation in reinforcement learning (RL) with a risk-sensitive objective. In risk-sensitive  $Q$ -learning and in CVaR constrained policy-gradient, exploration was inherently defined in the policy definition, which was not explicitly adapted during learning. On the other hand, in risk-sensitive MPC one assumed that the stochastic model was explicitly provided, in which MPC merely solved a planning problem without exploration. In general, the issue of balancing exploration-exploitation in RL has been the focus of many studies. In the literature of multi-armed bandit problems, several studies have considered exploration-exploitation tradeoffs with risk-sensitive objectives. However, to the best of our knowledge this line of work has not been detailedly studied in risk-sensitive RL. Interestingly, risk-sensitivity may be used as a metric to quantify the level of exploration in standard risk-neutral MDPs. In particular, in an unknown environment the learner should behave in a risk-seeking fashion in order to retrieve more information from potentially interesting regions. Once the environment is sufficiently explored, one may utilize risk-averse planning to guarantee safety while optimizing the cumulative return. This intuition is formalized in the UCRL2 algorithm [65], in which exploration policy is determined by solving an optimistic-MDP (the risk-seeking MDP counterpart compared to robust MDP). We speculate that the technical results presented in this thesis can be used as the theoretical underpinnings to further develop a principle strategy that trades-off exploration and exploitation in risk-sensitive RL.

### 6.4.2 Risk Sensitive Importance Sampling

Recall that in risk-sensitive policy-gradient one relies on sampled trajectories to estimate the risk-sensitive gradients. However for risk-measures such as CVaR that are sensitive to rare events, this method requires a large number of samples for an effective gradient estimation. To alleviate this issue, importance-sampling based approaches can be applied to estimate the (sampling-based) gradients with lower variance. For example in [141], by assuming the knowledge of the transition probability, the authors proposed an importance-sampling method for CVaR risk-sensitive policy gradient method and showed that this method significantly improved the sampling efficiency of the policy gradient algorithm. Therefore we believe that similar importance sampling based approaches can be applied to a wider range of risk-sensitive RL problems (for example to risk-sensitive RL problems whose objective functions are characterized by other coherent risk measures besides CVaR) to bolster their performances. Furthermore, designing an effective importance-sampling approach that is model-free is another important direction for future research.



### 6.4.3 Relationship to Safe Policy Improvement

Most data-driven risk-sensitive sequential decision making (risk-sensitive RL) algorithms presented in this thesis balance exploration versus exploitation and guarantee safety at the same time. These characteristics can also be found in the recently popular work on safe policy improvement [126, 66, 117]. Either by formulating the objective function with a regret metric (instead of a expected cumulative return) or by imposing extra constraints using the Kullback-Leibler divergence metric, in these studies the corresponding reinforcement learning algorithms always return *safe* policies, i.e., policies that are guaranteed to outperform certain baselines. Since providing safety guarantees in data-driven policy optimization is crucial to many real-world applications such as robotic path planning [57] and online marketing [149], we conjecture that by constructing appropriate risk metrics via shaping techniques (analogous to reward-shaping in reinforcement learning [90]), one can equivalently transform the risk-sensitive RL algorithms in this thesis to incorporate safe policy improvement.

## Chapter 7

# Supplementary Materials

### 7.1 Technical Results in Chapter 2

In this section we present the detailed proofs to the technical results in Chapter 2.

#### 7.1.1 Proof of Proposition 2.2.1

By definition, we have that

$$\begin{aligned} & \mathbb{E}_{\hat{P}} \left[ \sum_{t=0}^T \gamma^t C(x_t, a_t) \right] \\ &= \sum_{(x_0, a_0, \dots, x_T)} P_0(x_0) \delta_1(x_1|x_0, a_0) \cdots P_T(x_T|x_{T-1}, a_{T-1}) \delta_T(x_T|x_{T-1}, a_{T-1}) \sum_{t=0}^T \gamma^t C(x_t, a_t) \\ &= \sum_{(x_0, a_0, \dots, x_T)} P(x_0, a_0, \dots, x_T) \delta_1(x_1|x_0, a_0) \delta_2(x_2|x_1, a_1) \cdots \delta_T(x_T|x_{T-1}, a_{T-1}) \sum_{t=0}^T \gamma^t C(x_t, a_t) \\ &\doteq \sum_{(x_0, a_0, \dots, x_T)} P(x_0, a_0, \dots, x_T) \delta(x_0, a_0, \dots, x_T) \sum_{t=0}^T \gamma^t C(x_t, a_t). \end{aligned}$$

Note that by definition of the set  $\Delta$ , for any  $(\delta_1, \dots, \delta_T) \in \Delta$  we have that

$$P(x_0, a_0, \dots, x_T) > 0 \rightarrow \delta(x_0, a_0, \dots, x_T) \geq 0,$$

and

$$\mathbb{E} [\delta(x_0, a_0, \dots, x_T)] \doteq \sum_{(x_0, a_0, \dots, x_T)} P(x_0, a_0, \dots, x_T) \delta(x_0, a_0, \dots, x_T) = 1.$$

Thus,

$$\begin{aligned}
& \sup_{(\delta_1, \dots, \delta_T) \in \Delta_\eta} \mathbb{E}_{\hat{P}} \left[ \sum_{t=0}^T \gamma^t C(x_t, a_t) \right] \\
&= \sup_{\substack{0 \leq \delta(x_0, a_0, \dots, x_T) \leq \eta, \\ \mathbb{E}[\delta(x_0, a_0, \dots, x_T)] = 1}} \sum_{(x_0, a_0, \dots, x_T)} P(x_0, a_0, \dots, x_T) \delta(x_0, a_0, \dots, x_T) \sum_{t=0}^T \gamma^t C(x_t, a_t) \\
&= \text{CVaR}_{\frac{1}{\eta}} \left( \sum_{t=0}^T \gamma^t C(x_t, a_t) \right),
\end{aligned}$$

where the last equality is by the representation theorem for CVaR [132].

### 7.1.2 Proof of Lemma 2.3.2

The proof of monotonicity and constant shift properties follow directly from the definitions of the Bellman operator, by noting that  $\xi(x')P(x'|x, a)$  is non-negative and

$$\sum_{x' \in \mathcal{X}} \xi(x')P(x'|x, a) = 1$$

for any  $\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))$ . For the contraction property, denote  $c = \|V_1 - V_2\|_\infty$ . Since

$$V_2(x, y) - \|V_1 - V_2\|_\infty \leq V_1(x, y) \leq V_2(x, y) + \|V_1 - V_2\|_\infty, \quad \forall x \in \mathcal{X}, y \in \mathcal{Y},$$

by monotonicity and constant shift property,

$$\mathbf{T}[V_2](x, y) - \gamma\|V_1 - V_2\|_\infty \leq \mathbf{T}[V_1](x, y) \leq \mathbf{T}[V_2](x, y) + \gamma\|V_1 - V_2\|_\infty \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}.$$

This further implies that

$$|\mathbf{T}[V_1](x, y) - \mathbf{T}[V_2](x, y)| \leq \gamma\|V_1 - V_2\|_\infty \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

and the contraction property follows.

Now, we prove the concavity preserving property. Assume that  $yV(x, y)$  is concave in  $y$  for any  $x \in \mathcal{X}$ .

Let  $y_1, y_2 \in \mathcal{Y}$ , and  $\lambda \in [0, 1]$ , and define  $y_\lambda = (1 - \lambda)y_1 + \lambda y_2$ . We have

$$\begin{aligned}
& (1 - \lambda)y_1 \mathbf{T}[V](x, y_1) + \lambda y_2 \mathbf{T}[V](x, y_2) \\
&= (1 - \lambda)y_1 \min_{a_1 \in \mathcal{A}} \left[ C(x, a_1) + \gamma \max_{\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x, a_1))} \sum_{x' \in \mathcal{X}} \xi_1(x') V(x', y_1 \xi_1(x')) P(x'|x, a_1) \right] \\
&\quad + \lambda y_2 \min_{a_2 \in \mathcal{A}} \left[ C(x, a_2) + \gamma \max_{\xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x, a_2))} \sum_{x' \in \mathcal{X}} \xi_2(x') V(x', y_2 \xi_2(x')) P(x'|x, a_2) \right] \\
&= \min_{a_1 \in \mathcal{A}} \left[ (1 - \lambda)y_1 C(x, a_1) + \gamma \max_{\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x, a_1))} \sum_{x' \in \mathcal{X}} \xi_1(x') V(x', y_1 \xi_1(x')) P(x'|x, a_1) (1 - \lambda)y_1 \right] \\
&\quad + \min_{a_2 \in \mathcal{A}} \left[ \lambda y_2 C(x, a_2) + \gamma \max_{\xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x, a_2))} \sum_{x' \in \mathcal{X}} \xi_2(x') V(x', y_2 \xi_2(x')) P(x'|x, a_2) \lambda y_2 \right] \\
&\leq \min_{a \in \mathcal{A}} \left[ y_\lambda C(x, a) + \gamma \max_{\substack{\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x, a)) \\ \xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x, a))}} \sum_{x' \in \mathcal{X}} P(x'|x, a) ((1 - \lambda)y_1 \xi_1(x') V(x', y_1 \xi_1(x')) + \lambda y_2 \xi_2(x') V(x', y_2 \xi_2(x'))) \right] \\
&\leq \min_{a \in \mathcal{A}} \left[ y_\lambda C(x, a) + \gamma \max_{\substack{\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x, a)) \\ \xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x, a))}} \sum_{x' \in \mathcal{X}} P(x'|x, a) ((1 - \lambda)y_1 \xi_1(x') + \lambda y_2 \xi_2(x')) V(x', ((1 - \lambda)y_1 \xi_1(x') + \lambda y_2 \xi_2(x'))) \right]
\end{aligned}$$

where the first inequality is by concavity of the min, and the second is by the concavity assumption. Now, define

$$\xi = \frac{(1 - \lambda)y_1 \xi_1 + \lambda y_2 \xi_2}{y_\lambda}.$$

When  $\xi_1 \in \mathcal{U}_{\text{CVaR}}(y_1, P(\cdot|x, a))$  and  $\xi_2 \in \mathcal{U}_{\text{CVaR}}(y_2, P(\cdot|x, a))$ , we have that  $\xi \in \left[0, \frac{1}{y_\lambda}\right]$  and

$$\sum_{x' \in \mathcal{X}} \xi(x') P(x'|x, a) = 1.$$

We thus have

$$\begin{aligned}
& (1 - \lambda)y_1 \mathbf{T}[V](x, y_1) + \lambda y_2 \mathbf{T}[V](x, y_2) \\
&\leq \min_{a \in \mathcal{A}} \left[ y_\lambda C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_\lambda, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} P(x'|x, a) y_\lambda \xi(x') V(x', y_\lambda \xi(x')) \right] \\
&= y_\lambda \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_\lambda, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} P(x'|x, a) \xi(x') V(x', y_\lambda \xi(x')) \right] = y_\lambda \mathbf{T}[V](x, y_\lambda).
\end{aligned}$$

Finally, to show that the inner problem in (2.4) is a concave maximization, we need to show that

$$\Lambda_{x,y,a}(z) := \begin{cases} zV(x',z)P(x'|x,a)/y & \text{if } y \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

is a concave function in  $z \in \mathbb{R}$  for any given  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$  and  $a \in \mathcal{A}$ . Suppose  $zV(x,z)$  is a concave function in  $z$ . Immediately we can see that  $\Lambda_{x,y,a}(z)$  is concave in  $z$  when  $y = 0$ . Also notice that when  $y \in \mathcal{Y} \setminus \{0\}$ , since the transition probability  $P(x'|x,a)$  is non-negative, we have the result that  $\Lambda_{x,y,a}(z)$  is concave in  $z$ . This further implies

$$\sum_{x' \in \mathcal{X}} \frac{P(x'|x,a)}{y} \Lambda_{x,y,a}(y\xi(x')) = \sum_{x' \in \mathcal{X}} \xi(x')V(x',y\xi(x'))P(x'|x,a)$$

is concave in  $\xi$ . Furthermore by combining the result with the fact that the feasible set of  $\xi$  is a polytope, we complete the proof of this claim.

### 7.1.3 Proof of Theorem 2.3.3

Let  $\mathcal{C}_{0,T} = \sum_{t=0}^T \gamma^t C(x_t, a_t)$  denote the total discounted cost from time 0 up to time  $T$ . The first part of the proof is to show that for any  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ,

$$V_n(x, y) := \mathbf{T}^n[V_0](x, y) = \min_{\mu \in \Pi_S} \text{CVaR}_y(\mathcal{C}_{0,n} + \gamma^n V_0(x_n, y_n) \mid x_0 = x, \mu), \quad (7.1)$$

by induction, where the initial condition is  $(x_0, y_0) = (x, y)$  and control action  $a_t$  is induced by  $\mu(x_t, y_t)$ . For  $n = 1$ , we have that

$$\begin{aligned} V_1(x, y) &= \mathbf{T}[V_0](x, y) \\ &= \min_{\mu \in \Pi_S} C(x_0, a_0) + \gamma \text{CVaR}_y(C(x_1, a_1) + V_0(x_1, y_1) \mid x_0 = x, \mu) \end{aligned}$$

from definition. By induction hypothesis, assume the above expression holds at  $n = k$ . For  $n = k + 1$ ,

$$\begin{aligned} V_{k+1}(x, y) &:= \mathbf{T}^{k+1}[V_0](x, y) = \mathbf{T}[V_k](x, y) \\ &= \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \xi(x') V_k \left( x', \underbrace{y\xi(x')}_{y'} \right) P(x'|x, a) \right] \\ &= \min_{a \in \mathcal{A}} \left[ C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \xi(x') P(x'|x, a) \min_{\mu \in \Pi_S} \text{CVaR}_{y'}(\mathcal{C}_{0,k} + \gamma^k V_0 \mid x_0 = x', \mu) \right] \quad (7.2) \\ &= \min_{a \in \mathcal{A}} \left[ C(x, a) + \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \mathbb{E}_\xi \left[ \min_{\mu \in \Pi_S} \text{CVaR}_{y_1}(\mathcal{C}_{1,k+1} + \gamma^{k+1} V_0 \mid x_1, \mu) \right] \right] \\ &= \min_{\mu \in \Pi_S} \text{CVaR}_y(\mathcal{C}_{0,k+1} + \gamma^{k+1} V_0 \mid x_0 = x, \mu), \end{aligned}$$

where the initial state condition is given by  $(x_0, y_0) = (x, y)$ . Thus, the equality in (7.1) is proved by induction.

The second part of the proof is to show that  $V^*(x_0, y_0) = \min_{\mu \in \Pi_S} \text{CVaR}_{y_0}(\lim_{n \rightarrow \infty} \mathcal{C}_{0,n} \mid x_0, \mu)$ . Recall  $\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))} \mathbb{E}_\xi[V \mid x, y, a]$ . Since  $\mathbf{T}$  is a contraction and  $V_0$  is bounded, one obtains

$$V^*(x, y) = \mathbf{T}[V^*](x, y) = \lim_{n \rightarrow \infty} \mathbf{T}^n[V_0](x, y) = \lim_{n \rightarrow \infty} V_n(x, y)$$

for any  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ . The first and the second equality follow directly from Proposition 2.1 and Proposition 2.2 in [17] and the third equality follows from the definition of  $V_n$ . Furthermore since  $V_0(x, y)$  is bounded for any  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ , the result in (7.2) implies

$$-\lim_{n \rightarrow \infty} \gamma^n \|V_0\|_\infty \leq V^*(x_0, y_0) - \min_{\mu \in \Pi_S} \text{CVaR}_{y_0} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{0,n} \mid x_0, \mu \right) \leq \lim_{n \rightarrow \infty} \gamma^n \|V_0\|_\infty.$$

Therefore, by taking  $n \rightarrow \infty$ , we have just shown that for any  $(x_0, y_0) \in \mathcal{X} \times \mathcal{Y}$ ,

$$V^*(x_0, y_0) = \min_{\mu \in \Pi_S} \text{CVaR}_{y_0} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{0,n} \mid x_0, \mu \right).$$

The third part of the proof is to show that for the initial state  $x_0$  and confidence interval  $y_0$ , we have that

$$V^*(x_0, y_0) = \min_{\pi \in \Pi_H} \text{CVaR}_{y_0} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{0,n} \mid x_0, \pi \right).$$

At any  $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ , we first define the  $t^{\text{th}}$  tail-subproblem of problem  $\mathcal{OPT}_{\text{CM}}$  as follows:

$$\mathbb{V}(x_t, y_t) = \min_{\pi \in \Pi_H} \text{CVaR}_{y_t} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t,n} \mid x_t, \pi \right)$$

where the tail policy sequence is equal to  $\pi = \{\mu_t, \mu_{t+1}, \dots\}$  and the action is given by  $a_j = \mu_j(h_j)$  for  $j \geq t$ . For any history depend policy  $\tilde{\pi} \in \Pi_H$ , we also define the  $\tilde{\pi}$ -induced value function as  $\text{CVaR}_{y_t}(\lim_{n \rightarrow \infty} \mathcal{C}_{t,n} \mid x_t, \tilde{\pi})$  where  $\tilde{\pi} = \{\tilde{\mu}_t, \tilde{\mu}_{t+1}, \dots\}$  and  $a_j = \tilde{\mu}_j(h_j)$  for  $j \geq t$ .

Now let  $\pi^*$  be the optimal policy of the above  $t^{\text{th}}$  tail-subproblem. Clearly, the truncated policy  $\tilde{\pi} = \{\mu_{t+1}^*, \mu_{t+2}^*, \dots\}$  is a feasible policy for the  $(t+1)^{\text{th}}$  tail subproblem at any state  $x_{t+1}$  and confidence interval  $y_{t+1}$ :

$$\min_{\pi \in \Pi_H} \text{CVaR}_{y_{t+1}} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1,n} \mid x_{t+1}, \pi \right).$$

Collecting the above results, for any pair  $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$  and with  $a_t = \mu_t^*(x_t)$  we can write

$$\begin{aligned} \mathbb{V}(x_t, y_t) &= C(x_t, a_t) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, a_t))} \mathbb{E} \left[ \underbrace{\xi(x_{t+1}) \cdot \text{CVaR}_{y_{t+1}} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1, n} \mid x_{t+1}, \tilde{\pi} \right)}_{\mathbb{V}^{\tilde{\mu}}(x_{t+1}, y_{t+1}), y_{t+1} = y_t \xi(x_{t+1})} \right] \\ &\geq C(x_t, a_t) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, a_t))} \mathbb{E}_{\xi} [\mathbb{V}(x_{t+1}, y_t \xi(x_{t+1})) \mid x_t, y_t, a_t] \geq \mathbf{T}[\mathbb{V}](x_t, y_t). \end{aligned}$$

The first equality follows from the definition of  $\mathbb{V}(x_t, y_t)$  and the decomposition of CVaRs (Theorem 2.3.1). The first inequality uses the inequality:  $\mathbb{V}^{\tilde{\pi}}(x, y) \geq \mathbb{V}(x, y)$ ,  $\forall (x, y) \in \mathcal{X} \times \mathcal{Y}$ . The second inequality follows from the definition of Bellman operator  $\mathbf{T}$ .

On the other hand, starting at any state  $x_{t+1}$  and confidence interval  $y_{t+1}$ , let  $\pi^* = \{\mu_{t+1}^*, \mu_{t+2}^*, \dots\} \in \Pi_H$  be an optimal policy for the tail subproblem:

$$\min_{\pi \in \Pi_H} \text{CVaR}_{y_{t+1}} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1, n} \mid x_{t+1}, \pi \right).$$

For a given pair of  $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ , construct the “extended” policy  $\tilde{\pi} = \{\tilde{\mu}_t, \tilde{\mu}_{t+1}, \dots\} \in \Pi_H$  as follows:

$$\tilde{\mu}_t(x_t) = u^*(x_t, y_t), \text{ and } \tilde{\mu}_j(h_j) = \mu_j^*(h_j) \text{ for } j \geq t+1,$$

where  $u^*(x_t, y_t)$  is the minimizer of the fixed-point equation

$$u^*(x_t, y_t) \in \operatorname{argmin}_{a \in \mathcal{A}} C(x_t, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, a))} \mathbb{E}_{\xi} [\mathbb{V}(x_{t+1}, y_t \xi(x_{t+1})) \mid x_t, y_t, a],$$

with  $y_t$  is the given confidence interval to the  $t^{\text{th}}$  tail-subproblem and the transition from  $y_t$  to  $y_{t+1}$  is given by  $y_{t+1} = y_t \xi^*(x_{t+1})$  where

$$\xi^* \in \arg \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, a^*))} \mathbb{E} \left[ \xi(x_{t+1}) \cdot \text{CVaR}_{y_t \xi(x_{t+1})} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1, n} \mid x_{t+1}, \tilde{\pi} \right) \right]$$

Since  $\pi^*$  is an optimal, and a fortiori feasible policy for the tail subproblem (from time  $t+1$ ), the policy  $\tilde{\pi} \in \Pi_H$  is a feasible policy for the tail subproblem (from time  $t$ ):  $\min_{\pi \in \Pi_H} \text{CVaR}_{y_t} (\lim_{n \rightarrow \infty} \mathcal{C}_{t, n} \mid x_t, \pi)$ . Hence, we can write

$$\mathbb{V}(x_t, y_t) \leq C(x_t, \tilde{\mu}_t(x_t)) + \gamma \text{CVaR}_{y_t} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1, n} \mid x_t, \tilde{\pi} \right).$$

Hence from the definition of  $\pi^*$ , one easily obtains:

$$\begin{aligned}
& \mathbb{V}(x_t, y_t) \\
& \leq C(x_t, u^*(x_t, y_t)) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, u^*(x_t, y_t)))} \mathbb{E} \left[ \xi(x_{t+1}) \cdot \text{CVaR}_{y_t \xi(x_{t+1})} \left( \lim_{n \rightarrow \infty} \mathcal{C}_{t+1, n} \mid x_{t+1}, \tilde{\pi} \right) \mid x_t, y_t, u^*(x_t, y_t) \right] \\
& = C(x_t, u^*(x_t, y_t)) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y_t, P(\cdot | x_t, u^*(x_t, y_t)))} \mathbb{E}_\xi [\mathbb{V}(x_{t+1}, y_t \xi(x_{t+1})) \mid x_t, y_t, u^*(x_t, y_t)] \\
& = \mathbf{T}[\mathbb{V}](x_t, y_t).
\end{aligned}$$

Collecting the above results, we have shown that  $\mathbb{V}$  is a fixed-point solution to  $V(x, y) = \mathbf{T}[V](x, y)$  for any  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ . Since the fixed-point solution is unique, combining both of these arguments implies  $V^*(x, y) = \mathbb{V}(x, y)$  for any  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ . Therefore, it follows that with initial state  $(x, y)$ , we have  $V^*(x, y) = \mathbb{V}(x, y) = \min_{\pi \in \Pi_H} \text{CVaR}_y(\lim_{T \rightarrow \infty} \mathcal{C}_{0, T} \mid x_0 = x, \pi)$ .

Combining the above three parts of the proof, the claims of this theorem follows.

#### 7.1.4 Proof of Theorem 2.3.4

Similar to the definition of the optimal Bellman operator  $\mathbf{T}$ , for any augmented stationary Markovian policy  $u : \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{A}$ , we define the policy induced Bellman operator  $\mathbf{T}_u$  as

$$\mathbf{T}_u[V](x, y) = C(x, u(x, y)) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot | x, u(x, y)))} \sum_{x' \in \mathcal{X}} \xi(x') V(x', y \xi(x')) P(x' | x, u(x, y)).$$

Analogous to Theorem 2.3.3, we can easily show that the fixed-point solution to  $\mathbf{T}_u[V](x, y) = V(x, y)$  is unique and the CVaR decomposition theorem (Theorem 2.3.1) further implies this solution is equal to

$$\text{CVaR}_y \left( \lim_{T \rightarrow \infty} \mathcal{C}_{0, T} \mid x_0 = x, \pi_H \right),$$

where the history dependent policy  $\pi_H = \{\mu_0, \mu_1, \dots\}$  is given by  $\mu_k(h_k) = u(x_k, y_k)$  for any  $k \geq 0$ , with initial states  $x_0, y_0 = \alpha$ , state transitions (2.6), but with augmented stationary Markovian policy  $u^*$  replaced by  $u$ .

To complete the proof of this theorem, we need to show that the augmented stationary Markovian policy  $u^*$  is optimal if and only if

$$\mathbf{T}[V^*](x, y) = \mathbf{T}_{u^*}[V^*](x, y), \quad \forall x \in \mathcal{X}, y \in \mathcal{Y}, \quad (7.3)$$

where  $V^*(x, y)$  is the unique fixed-point solution of  $\mathbf{T}[V](x, y) = V(x, y)$ . Here an augmented stationary Markovian policy  $u^*$  is optimal if and only if the induced history dependent policy  $\pi_H^*$  in (2.5) is optimal to problem  $\mathcal{OPT}_{\text{CM}}$ .

First suppose  $u^*$  is an optimal augmented stationary Markovian policy. Then using the definition of  $u^*$



and the result from Theorem 2.3.3 that

$$V^*(x, y) = \min_{\pi \in \Pi_H} \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x_0 = x, \pi \right),$$

we immediately show that  $V^*(x, y) = V_{u^*}(x, y)$ , where  $V_{u^*}$  is the fixed-point solution to  $V(x, y) = \mathbf{T}_{u^*}[V](x, y)$  for any  $x, y$ . By the fixed-point equation  $\mathbf{T}[V^*](x, y) = V^*(x, y)$  and  $\mathbf{T}_{u^*}[V_{u^*}](x, y) = V_{u^*}(x, y)$ , this further implies (7.3) holds.

Second suppose  $u^*$  satisfies the equality in (7.3). Then by the fixed-point equality  $\mathbf{T}[V^*](x, y) = V^*(x, y)$ , we immediately obtain the equation  $V^*(x, y) = \mathbf{T}_{u^*}[V^*](x, y)$  for any  $x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ . since the fixed-point solution to  $\mathbf{T}_{u^*}[V](x, y) = V(x, y)$  is unique, we further show that  $\mathbf{T}[V^*](x, y) = V^*(x, y) = V_{u^*}(x, y)$  and  $V_{u^*}(x, y) = \min_{\pi \in \Pi_H} \text{CVaR}_y (\lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x_0 = x, \pi)$  from Theorem 2.3.3. By using the policy construction formula in (2.5) to obtain the history dependent policy  $\pi_H^*$  and following the above arguments at which the augmented Markovian stationary policy  $u$  is replaced by  $u^*$ , this further implies

$$\min_{\pi \in \Pi_H} \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x_0 = x, \pi \right) = \text{CVaR}_y \left( \lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x_0 = x, \pi_H^* \right),$$

i.e.,  $u^*$  is an optimal augmented stationary Markovian policy.

### 7.1.5 Proof of Lemma 2.4.3

We first proof the monotonicity property. Based on the definition of  $\mathcal{I}_x[V](y)$ , if  $V_1(x, y) \geq V_2(x, y) \forall x \in \mathcal{X}$  and  $y \in \mathcal{Y}$ , we have that

$$\mathcal{I}_x[V_1](y) = \frac{y_{i+1}V_1(x, y_{i+1})(y - y_i) + y_iV_1(x, y_i)(y_{i+1} - y)}{y_{i+1} - y_i}, \text{ if } y \in \mathbf{I}_i(x).$$

Since  $y_i, y_{i+1} \in \mathcal{Y}$  and  $(y_{i+1} - y), (y - y_i) \geq 0$  (because  $y \in \mathbf{I}_i(x)$ ), we can easily see that  $\mathcal{I}_x[V_1](y) \geq \mathcal{I}_x[V_2](y)$ . As  $y \in \mathcal{Y}$  and  $\xi(\cdot)P(\cdot \mid x, a) \geq 0$  for any  $\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))$ , this further implies  $\mathbf{T}_{\mathcal{I}}[V_1](x, y) \geq \mathbf{T}_{\mathcal{I}}[V_2](x, y)$ .

Next we prove the constant shift property. Note from the definition of  $\mathcal{I}_x[V](y)$  that

$$\begin{aligned} & \mathcal{I}_x[V + K](y) \\ &= y_i(V(x, y_i) + K) + \frac{y_{i+1}(V(x, y_{i+1}) + K) - y_i(V(x, y_i) + K)}{y_{i+1} - y_i}(y - y_i), \text{ if } y \in \mathbf{I}_i(x), \\ &= y_iK + y_iV(x, y_i) + \frac{y_{i+1}V(x, y_{i+1}) - y_iV(x, y_i)}{y_{i+1} - y_i}(y - y_i), \text{ if } y \in \mathbf{I}_i(x) \\ &= \mathcal{I}_x[V](y) + yK. \end{aligned}$$

Therefore by definition of  $\mathbf{T}_{\mathcal{I}}[V](x, y)$ , the constant shift property:  $\mathbf{T}_{\mathcal{I}}[V + K](x, y) = \mathbf{T}_{\mathcal{I}}[V](x, y) + \gamma K$  for any  $x \in \mathcal{X}, y \in \mathcal{Y}$ , follows directly from the above arguments.

Equipped with both properties in monotonicity and constant shift, the proof of contraction of  $\mathbf{T}_{\mathcal{I}}$  directly

follows from the analogous proof in Lemma 2.3.2.

Finally we prove the concavity preserving property. Assume  $yV(x, y)$  is concave in  $y \in \mathcal{Y}$  for any  $x \in \mathcal{X}$ . Then for  $y_{i+2} > y_{i+1} > y_i$ ,  $\forall i \in \{1, \dots, N(x) - 2\}$  the following inequality immediately follows from the definition of a concave function:

$$\begin{aligned} \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y \in \mathbf{I}_{i+1}(x)} &= \frac{y_{i+1}V(x, y_{i+1}) - y_iV(x, y_i)}{y_{i+1} - y_i} \\ &\geq \frac{y_{i+2}V(x, y_{i+2}) - y_{i+1}V(x, y_{i+1})}{y_{i+2} - y_{i+1}} = \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y \in \mathbf{I}_{i+2}(x)}. \end{aligned} \quad (7.4)$$

We then show that the following inequality in each of the following cases, whenever the slope exists:

$$\mathcal{I}_x[V](z_1) \leq \mathcal{I}_x[V](z_2) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_1 - z_2), \quad \forall z_1, z_2 \in \mathcal{Y} \setminus \{0\}.$$

(1) There exists  $i \in \{1, \dots, N(x) - 1\}$  such that  $z_1, z_2 \in \mathbf{I}_{i+1}(x)$ . In this case we have that

$$\frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} = \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2},$$

and this further implies

$$\mathcal{I}_x[V](z_1) = \mathcal{I}_x[V](z_2) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_1 - z_2).$$

(2) There exists  $i, j \in \{1, \dots, N(x) - 2\}$ ,  $i + 1 < j$  such that  $z_1 \in \mathbf{I}_{i+1}(x)$  and  $z_2 \in \mathbf{I}_j(x)$ . In this case, without loss of generality we assume  $j = i + 1$ . The proof for case:  $j > i + 2$  is omitted for the sake of brevity, as it can be completed by iteratively applying the same arguments from case:  $j = i + 2$ . Since  $z_1 \in \mathbf{I}_i(x)$ ,  $z_2 \in \mathbf{I}_j(x)$ , we have  $z_2 - z_1 \geq 0$  and

$$\frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} \geq \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2}.$$

Based on the definition of the linear interpolation function, we have that

$$\mathcal{I}_x[V](y_{i+1}) = y_{i+1}V(x, y_{i+1}) = \mathcal{I}_x[V](y_i) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y \in \mathbf{I}_{i+1}(x)} (y_{i+1} - y_i).$$

Furthermore, combining previous arguments with the definitions of  $\mathcal{I}_x[V](z_1)$ ,  $\mathcal{I}_x[V](z_2)$  implies that for

$$(z_2 - y_{i+1}) \geq 0,$$

we have that

$$\begin{aligned}
\mathcal{I}_x[V](z_2) &= \mathcal{I}_x[V](y_{i+1}) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_2 - y_{i+1}) \\
&\leq \mathcal{I}_x[V](y_{i+1}) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} (z_2 - y_{i+1}) \\
&= \mathcal{I}_x[V](y_i) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y \in \mathbf{I}_{i+1}(x)} (z_2 - y_i) \\
&= \mathcal{I}_x[V](z_1) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} (z_2 - z_1).
\end{aligned}$$

(3) There exists  $i, j \in \{1, \dots, N(x) - 2\}$ ,  $i + 1 < j$  such that  $z_2 \in \mathbf{I}_{i+1}(x)$  and  $z_1 \in \mathbf{I}_j(x)$ . In this case, without loss of generality we assume  $j = i + 1$ . The proof for case:  $j > i + 2$  is omitted for the sake of brevity, as it can be completed by iteratively applying the same arguments from case:  $j = i + 2$ . Since  $z_2 \in \mathbf{I}_{i+1}(x)$ ,  $z_1 \in \mathbf{I}_j(x)$ , we have  $z_1 - z_2 \geq 0$  and

$$\frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} \leq \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2}.$$

Similar to the analysis in the previous case, we have that

$$\mathcal{I}_x[V](y_i) = y_i V(x, y_i) = \mathcal{I}_x[V](y_{i+1}) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y \in \mathbf{I}_{i+1}(x)} (y_i - y_{i+1})$$

Furthermore, combining previous arguments with the definitions of  $\mathcal{I}_x[V](z_1)$ ,  $\mathcal{I}_x[V](z_2)$  implies that for  $(z_2 - z_1) \leq 0$ ,

$$\begin{aligned}
\mathcal{I}_x[V](z_2) &= \mathcal{I}_x[V](y_i) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_2 - y_i) \\
&= \mathcal{I}_x[V](y_{i+1}) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_2 - y_{i+1}) \\
&= \mathcal{I}_x[V](z_1) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_2} (z_2 - z_1) \\
&\leq \mathcal{I}_x[V](z_1) + \frac{d\mathcal{I}_x[V](y)}{dy} \Big|_{y=z_1} (z_2 - z_1).
\end{aligned}$$

Thus we have just shown that the first order sufficient condition for concave functions, corresponding to  $\mathcal{I}_x[V](y)$ , holds, i.e.,  $\mathcal{I}_x[V](y)$  is concave in  $y \in \mathcal{Y} \setminus \{0\}$  for any given  $x \in \mathcal{X}$ . Now since  $\mathcal{I}_x[V](y)$  is a continuous piecewise linear function in  $y \in \mathcal{Y}$  and a concave function when the domain is restricted to  $\mathcal{Y} \setminus \{0\}$ . By continuity this immediately implies that  $\mathcal{I}_x[V](y)$  is concave in  $y \in \mathcal{Y}$  as well. Then following the identical arguments in the proof of Lemma 2.3.2 for the concavity preserving property, we can thereby

show that

$$y\mathbf{T}_{\mathcal{I}}[V](x, y) = \min_{a \in \mathcal{A}} \left\{ yC(x, a) + \max_{\xi \in \mathcal{U}_{\text{cVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \mathcal{I}_{x'}[V](y\xi(x'))P(x'|x, a) \right\}$$

is concave in  $y \in \mathcal{Y}$  for any given  $x \in \mathcal{X}$ .

### 7.1.6 Useful Intermediate Results

**Lemma 7.1.1.** *Let  $f(y) : [0, 1] \rightarrow R$  be a concave function, differentiable almost everywhere, with Lipschitz constant  $M$ . Then the linear interpolation  $\mathcal{I}[f](y)$  is also concave, and with Lipschitz constant  $M_I \leq M$ .*

**Proof** For every segment  $[y_j, y_{j+1}]$  in the linear interpolation,  $f(y)$  is concave, and with Lipschitz constant  $M$ , and  $\mathcal{I}[f](y)$  is linear. Also,  $f(y_j) = \mathcal{I}[f](y_j)$ , and  $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$ , by definition of the linear interpolation. Denote by  $c_j$  the magnitude of the slope of  $\mathcal{I}[f](y)$  at  $y \in [y_j, y_{j+1}]$ .

Assume by contradiction that  $c_j > \max_{y \in [y_j, y_{j+1}]} |f'(y)|$  whenever  $f'(y)$  exists. Consider the case when  $f(y_{j+1}) \geq f(y_j)$ . This implies  $c_j$  is the slope of the interpolation function  $\mathcal{I}[f](y)$  at  $y \in [y_j, y_{j+1}]$ . Then by the fundamental theorem of calculus, we have

$$f(y_{j+1}) - f(y_j) = \int_{y_j}^{y_{j+1}} f'(y) dy \leq \int_{y_j}^{y_{j+1}} |f'(y)| dy < \int_{y_j}^{y_{j+1}} c_j dy = (\mathcal{I}[f](y_{j+1}) - \mathcal{I}[f](y_j)),$$

contradicting  $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$  and  $f(y_j) = \mathcal{I}[f](y_j)$ .

On the other hand, consider the case when  $f(y_{j+1}) \leq f(y_j)$ . This implies  $-c_j$  is the slope of the interpolation function  $\mathcal{I}[f](y)$  at  $y \in [y_j, y_{j+1}]$ . Again by fundamental theorem of calculus,

$$0 \leq f(y_{j+1}) - f(y_j) = \int_{y_j}^{y_{j+1}} f'(y) dy \geq \int_{y_j}^{y_{j+1}} -|f'(y)| dy > \int_{y_j}^{y_{j+1}} -c_j dy = \mathcal{I}[f](y_j) - \mathcal{I}[f](y_{j+1}).$$

Since  $f(y_{j+1}) = \mathcal{I}[f](y_{j+1})$  and  $f(y_j) = \mathcal{I}[f](y_j)$ , which implies  $\mathcal{I}[f](y_j) - \mathcal{I}[f](y_{j+1}) \geq 0$ , the above expression clearly leads to a contradiction.

We finally have that  $\max_{y \in [y_j, y_{j+1}]} |f'(y)| \geq c_j$  for segment  $j \in \{1, \dots, N(x) - 1\}$ . As this argument holds for each segment, by maximizing over  $j$  over  $\{1, \dots, N(x) - 1\}$ , we have that

$$M \geq \max_{j \in \{1, \dots, N(x) - 1\}} \max_{y \in [y_j, y_{j+1}]} |f'(y)| \geq \max_{j \in \{1, \dots, N(x) - 1\}} c_j = M_I.$$

The concavity property (thus differentiability almost everywhere) are well-known results of linear interpolation [103].

**Lemma 7.1.2.** *Let  $yV(x, y)$  be Lipschitz with constant  $M$ , concave, and differentiable almost everywhere, for every  $x \in \mathcal{X}$  and  $y \in [0, 1]$ . Then  $y\mathbf{T}[V](x, y)$  is also Lipschitz with constant  $C_{\max} + \gamma M$ .*

**Proof** For any given state-action pair  $x \in \mathcal{X}$ , and  $a \in \mathcal{A}$ , let  $P(x') = P(x'|x, a)$  be the transition kernel. Consider the function

$$H(y) \doteq \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot))} \sum_{x' \in \mathcal{X}} y\xi(x') V(x', y\xi(x')) P(x').$$

Note that, by definition of  $\mathcal{U}_{\text{CVaR}}$ , and a change of variables  $z(x') = y\xi(x')$ , we can write  $H(y)$  as follows:

$$H(y) = \max_{\substack{0 \leq z(x') \leq 1, \\ \sum_{x'} P(x') z(x') = y}} \sum_{x' \in \mathcal{X}} z(x') V(x', z(x')) P(x'). \quad (7.5)$$

The Lagrangian of the above maximization problem is

$$L(z, \lambda; y) = \sum_{x' \in \mathcal{X}} z(x') V(x', z(x')) P(x') - \lambda \left( \sum_{x'} P(x') z(x') - y \right).$$

Since  $yV(x, y)$  is concave, the maximum is attained. By first order optimality condition the following expression holds:

$$\frac{\partial L(z, \lambda; y)}{\partial z(x')} = P(x') \frac{\partial [z(x') V(x', z(x'))]}{\partial z(x')} - \lambda P(x') = 0.$$

Summing the last expression over  $x'$ , we obtain:

$$\sum_{x' \in \mathcal{X}} P(x') \frac{\partial [z(x') V(x', z(x'))]}{\partial z(x')} = \sum_{x' \in \mathcal{X}} \lambda P(x') = \lambda.$$

Now, from the Lipschitz property of  $yV(x, y)$ , we have

$$\left| \sum_{x' \in \mathcal{X}} \lambda P(x') \right| \leq \sum_{x' \in \mathcal{X}} P(x') \left| \frac{\partial [z(x') V(x', z(x'))]}{\partial z(x')} \right| \leq \sum_{x' \in \mathcal{X}} P(x') M = M.$$

Thus,

$$|\lambda| \leq \sum_{x' \in \mathcal{X}} P(x') \left| \frac{\partial [z(x') V(x', z(x'))]}{\partial z(x')} \right| \leq M.$$

Note that the objective in (7.5) does not depend on  $y$ . From the envelope theorem [86], it follows that

$$\frac{dH(y)}{dy} = \lambda,$$

therefore,  $H(y)$  is Lipschitz, with constant  $M$ .

Now, by definition,

$$y\mathbf{T}[V](x, y) = \min_{a \in \mathcal{A}} \left[ yC(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} y\xi(x') V(x', y\xi(x')) P(x'|x, a) \right].$$

Using our Lipschitz result for  $H(y)$ , we have that for any  $a \in \mathcal{A}$ , the function

$$yC(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} y\xi(x') V(x', y\xi(x')) P(x'|x, a)$$

is Lipschitz in  $y$ , with constant  $C(x, a) + \gamma M$ . Using again the envelope theorem [86], we obtain that  $y\mathbf{T}[V](x, y)$  is Lipschitz, with constant  $C_{\max} + \gamma M$ .

**Lemma 7.1.3.** *Consider Algorithm 1. Assume that for any  $x \in \mathcal{X}$ , the initial value function satisfies that  $yV_0(x, y)$  is Lipschitz (in  $y$ ), with uniform constant  $M_0$ . We have that for any  $t \in \{0, 1, \dots\}$ , the function  $yV_t(x, y)$  is Lipschitz in  $y$  for any  $x \in \mathcal{X}$ , with Lipschitz constant*

$$M_t = \frac{1 - \gamma^t}{1 - \gamma} C_{\max} + \gamma^t M_0 \leq \frac{C_{\max}}{1 - \gamma} + M_0, \quad \forall t.$$

**Proof** Let  $\mathbf{T}_{\mathcal{I}}[V]$  denote the application of the Bellman operator  $\mathbf{T}$  to the linearly-interpolated version of  $yV(x, y)$ . We have, by definition, that

$$V_1(x, y) = \mathbf{T}_{\mathcal{I}}[V_0](x, y).$$

Using Lemma 7.1.1 and Lemma 7.1.2, we have that  $V_1(x, y)$  is Lipschitz, with  $M_1 \leq C_{\max} + \gamma M_0$ .

Note now, that  $V_2(x, y) = \mathbf{T}_{\mathcal{I}}[V_1](x, y)$ . Thus, by induction, we have

$$M_t \leq \frac{1 - \gamma^t}{1 - \gamma} C_{\max} + \gamma^t M_0,$$

and the result follows.

### 7.1.7 Proof of Theorem 2.4.4

The proof of this theorem is split into three parts. In the first part, we bound the difference  $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$  at each state  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  using the previous technical lemmas and Lipschitz property.

In the second part, we bound the difference of  $\mathbf{T}_{\mathcal{I}}[V_t](x, y) - \mathbf{T}[V_t](x, y)$ .

In the third part we bound the interpolation error using contraction properties of Bellman recursions.

First we analyze the bounds for  $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$  in the following four cases. Notice that from Lemma 7.1.3, we have that  $|d\mathcal{I}_x[V_t](y)/dy| \leq M := C_{\max}/(1 - \gamma) + M_0$ .

(1) When  $y = 0$  (for which  $y \in \mathbf{I}_1(x)$ ).

Using previous analysis and L'Hospital's rule we have that  $\lim_{y \rightarrow 0} \mathcal{I}_x[V_t](y)/y = V_t(x, 0)$ . This further implies  $\lim_{y \rightarrow 0} \mathcal{I}_x[V_t](y)/y - V_t(x, 0) = 0$ .

(2) When  $y \in \mathbf{I}_{i+1}(x)$ ,  $2 \leq i < N(x) - 1$ .

Similar to the inequality in (7.4), by concavity of  $yV_t(x, y)$  in  $y \in \mathcal{Y}$ , we have that

$$\left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+1}(x)} = \frac{y_{i+1}V_t(x, y_{i+1}) - y_iV_t(x, y_i)}{y_{i+1} - y_i} \leq \frac{yV_t(x, y) - y_iV_t(x, y_i)}{y - y_i},$$

and

$$\left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+2}(x)} = \frac{y_{i+2}V_t(x, y_{i+2}) - y_{i+1}V_t(x, y_{i+1})}{y_{i+2} - y_{i+1}} \leq \frac{y_{i+1}V_t(x, y_{i+1}) - yV_t(x, y)}{y_{i+1} - y}.$$

From the first inequality, for each  $(x, y) \in \mathcal{X} \times \mathcal{Y}$  we get,

$$\frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y) \leq \frac{1}{y} \left( y_iV_t(x, y_i) + \frac{y_{i+1}V_t(x, y_{i+1}) - y_iV_t(x, y_i)}{y_{i+1} - y_i} (y - y_i) - yV_t(x, y) \right) \leq 0. \quad (7.6)$$

On the other hand, rearranging the second inequality gives

$$\begin{aligned} & \frac{1}{y} (\mathcal{I}_x[V_t](y) - yV_t(x, y)) \\ & \geq \frac{1}{y} \left( y_iV_t(x, y_i) + \left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+1}(x)} (y - y_i) - y_{i+1}V_t(x, y_{i+1}) - \left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+2}(x)} (y - y_{i+1}) \right) \\ & = \left( \left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+1}(x)} - \left. \frac{d\mathcal{I}_x[V_t](y)}{dy} \right|_{y \in \mathbf{I}_{i+2}(x)} \right) \frac{y - y_{i+1}}{y} \geq -2M \left( \frac{y_{i+1}}{y} - 1 \right). \end{aligned} \quad (7.7)$$

Furthermore by the Lipschitz property, we also have the following inequality as well:

$$\begin{aligned} & \frac{1}{y} (\mathcal{I}_x[V_t](y) - yV_t(x, y)) \\ & = \frac{y_{i+1}V_t(x, y_{i+1})(y - y_i) + y_iV_t(x, y_i)(y_{i+1} - y)}{(y_{i+1} - y_i)y} - V_t(x, y) \\ & \geq \frac{y_iV_t(x, y_i)(y - y_i) + y_iV_t(x, y_i)(y_{i+1} - y) - M(y_{i+1} - y_i)(y - y_i)}{(y_{i+1} - y_i)y} - V_t(x, y) \\ & = \frac{y_iV_t(x, y_i) - M(y - y_i)}{y} - V_t(x, y) \geq -2M \left( 1 - \frac{y_i}{y} \right). \end{aligned} \quad (7.8)$$

Combining the inequalities (7.7) and (7.8), the following lower bound for  $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$  holds:

$$\frac{1}{y} (\mathcal{I}_x[V_t](y) - yV_t(x, y)) \geq \delta := -2M \min \left\{ 1 - \frac{y_i}{y}, \frac{y_{i+1}}{y} - 1 \right\}, \quad \forall y \in \mathbf{I}_{i+1}(x), \quad i \geq 2.$$

From the above definition, when  $y_i \leq y \leq (y_i + y_{i+1})/2$ , the lower bound becomes  $\delta = -2M(1 - y_i/y)$  and when  $(y_i + y_{i+1})/2 \leq y \leq y_{i+1}$ , the corresponding lower bound is  $\delta = -2M(y_{i+1}/y - 1)$ . In both cases,  $\delta$  is minimized when  $y = (y_i + y_{i+1})/2$ . Therefore, the above analysis implies the following lower

bound:

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \geq -2M \frac{y_{i+1} - y_i}{y_{i+1} + y_i}, \forall y \in \mathbf{I}_{i+1}(x), i \geq 2.$$

When  $y_{i+1} = \theta y_i$  for  $i \in \{2, \dots, N(x) - 1\}$  for some constant  $\theta \geq 1$ , this further implies that

$$\frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \geq -2M \frac{\theta - 1}{\theta + 1} \geq -M(\theta - 1), \forall y \in \mathcal{Y} \setminus [0, \epsilon].$$

Then combining the results, here we get the following bound for  $\mathcal{I}_x[V_t](y)/y - V_t(x, y)$ :

$$-M(\theta - 1) \leq \frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y) \leq 0, \forall y \in \mathbf{I}_{i+1}(x), i \geq 2.$$

(3) When  $y \in \mathbf{I}_{N(x)}(x)$ , i.e.,  $y \in (y_{N(x)-1}, 1]$ .

Similar to the proof of case (2), we can show that for any  $x \in \mathcal{X}$  and  $y \in \mathbf{I}_{N(x)}(x)$ , the same lines of arguments in inequality (7.6) and (7.8) hold, which implies

$$-2M(1 - y_{N(x)-1}) \leq -2M \left(1 - \frac{y_{N(x)-1}}{y}\right) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \leq 0.$$

When  $y_{N(x)} = 1 = \theta y_{N(x)-1}$ , this further shows that

$$-2M y_{N(x)-1}(\theta - 1) = -2M(y_{N(x)} - y_{N(x)-1}) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \leq 0,$$

and

$$-2M(\theta - 1) \leq -\frac{2}{\theta}M(\theta - 1) \leq \frac{1}{y}(\mathcal{I}_x[V_t](y) - yV_t(x, y)) \leq 0.$$

(4) When  $y \in \mathbf{I}_2(x)$ , i.e.,  $y \in (0, y_2]$ .

From inequality (7.6), the definition of  $\mathcal{I}_x[V_t](y)$ , we have that

$$0 \geq \frac{\mathcal{I}_x[V_t](y) - yV_t(x, y)}{y} = \frac{y(V_t(x, y_2) - V_t(x, y))}{y} = V_t(x, y_2) - V_t(x, y) \geq V_t(x, y_2) - V_t(x, 0).$$

The first inequality is due to the fact that  $yV_t(x, y)$  is concave in  $y \in \mathcal{Y}$  for any  $x \in \mathcal{X}$ , thus the first order condition implies

$$\frac{y_2 V_n(x, y_2) - y_1 V_n(x, y_1)}{y_2 - y_1} \leq \frac{y V_n(x, y) - y_1 V_n(x, y_1)}{y - y_1}, \forall y \in \mathbf{I}_2(x),$$

and the last inequality is due to the similar fact that

$$V_t(x, w) = \frac{wV_t(x, w) - 0 \cdot V_t(x, 0)}{w - 0} \leq \frac{zV_t(x, z) - 0 \cdot V_t(x, 0)}{z - 0} = V_t(x, z), \forall z, w \in \mathcal{Y}, z \leq w.$$



Therefore the condition of this theorem implies

$$0 \geq \frac{\mathcal{I}_x[V_t](y) - yV_t(x, y)}{y} \geq -\epsilon, \quad \forall t \geq 0, x \in \mathcal{X}, y \in \mathcal{Y}.$$

Combining the above four cases, we have that for each state  $(x, y) \in \mathcal{X} \times \mathcal{Y}$ ,

$$0 \geq \frac{\mathcal{I}_x[V_t](y)}{y} - V_t(x, y) \geq -2M(\theta - 1) - \epsilon, \quad \forall t.$$

Second, we bound the difference of  $\mathbf{T}_{\mathcal{I}}[V_t](x, y) - \mathbf{T}[V_t](x, y)$ . By recalling that  $\xi(\cdot)P(\cdot|x, a)$  is a probability distribution for any  $\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))$ , we then combine all previous arguments and show that at any  $t \in \{0, 1, \dots\}$  and any  $x \in \mathcal{X}, a \in \mathcal{A}, y \in \mathcal{Y}(x)$ ,

$$\max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}, \xi(x') \neq 0} \left( \frac{\mathcal{I}_{x'}[V_t](y\xi(x'))}{y\xi(x')} - V_t(x', y\xi(x')) \right) \xi(x')P(x'|x, a) \geq -2M(\theta - 1) - \epsilon.$$

This further implies

$$\mathbf{T}[V_t](x, y) - \gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_t](x, y) \leq \mathbf{T}[V_t](x, y). \quad (7.9)$$

Third, we prove the error bound of interpolated value iteration using the above properties. By putting  $t = 0$  in (7.9), we have that

$$-\gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_0](x, y) - \mathbf{T}[V_0](x, y) \leq 0.$$

Applying the Bellman operator  $\mathbf{T}$  on all sides of the above inequality and noting that  $\mathbf{T}$  is a translational invariant mapping, the above expression implies

$$\mathbf{T}^2[V_0](x, y) - \gamma^2(2M(\theta - 1) + \epsilon) \leq \mathbf{T}[\mathbf{T}_{\mathcal{I}}[V_0]](x, y) = \mathbf{T}[V_1](x, y) \leq \mathbf{T}^2[V_0](x, y).$$

By adding the inequality:  $-\gamma(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_1](x, y) - \mathbf{T}[V_1](x, y) \leq 0$  to the above expression, this further implies the following expression:

$$\mathbf{T}^2[V_0](x, y) - \gamma(1 + \gamma)(2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}[V_1](x, y) = \mathbf{T}_{\mathcal{I}}^2[V_0](x, y) \leq \mathbf{T}^2[V_0](x, y).$$

Then, by repeating this process, we can show that for any  $n \in \mathbb{N}$ , the following inequality holds:

$$\mathbf{T}^n[V_0](x, y) - \gamma \frac{1 - \gamma^n}{1 - \gamma} (2M(\theta - 1) + \epsilon) \leq \mathbf{T}_{\mathcal{I}}^n[V_0](x, y) \leq \mathbf{T}^n[V_0](x, y).$$

Note that when  $n \rightarrow \infty$ , we have that  $\gamma^n$  converges to 0,  $\mathbf{T}^n[V_0](x, y)$  converges to  $\min_{\pi \in \Pi_H} \text{CVaR}_y(\lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x, \mu)$  (follow from Theorem 2.3.3) and  $\mathbf{T}_{\mathcal{I}}^n[V_0](x, y)$  converges to  $\hat{V}^*(x, y)$  (follow from the contraction property

in Lemma 2.4.3).

Furthermore, from Proposition 1.6.4 in [17], the contraction property of Bellman operator  $\mathbf{T}$  implies that for any  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ , the following expression holds:

$$|\mathbf{T}^n[V_0](x, y) - V^*(x, y)| \leq \frac{\gamma^n}{1 - \gamma} (C_{\max} + \|Z\|_{\infty})$$

where  $Z$  is the bounded random variable of the initial value function  $V_0(x, y) = \text{CVaR}_y(Z \mid x_0 = x)$  such that  $\|V_0\|_{\infty} \leq \|Z\|_{\infty}$ , and  $V^*(x, y) = \min_{\pi \in \Pi_H} \text{CVaR}_y(\lim_{T \rightarrow \infty} \mathcal{C}_{0,T} \mid x, \mu)$ . This further implies for any  $x \in \mathcal{X}$ ,  $y \in \mathcal{Y}$ ,

$$|\mathbf{T}_{\mathcal{I}}^n[V_0](x, y) - V^*(x, y)| \leq \gamma \frac{1 - \gamma^n}{1 - \gamma} (2M(\theta - 1) + \epsilon) + \frac{\gamma^n}{1 - \gamma} (C_{\max} + \|Z\|_{\infty}).$$

Then, by combining all the above arguments, we prove the claim of this theorem.

### 7.1.8 Proof of Theorem 2.5.1

The convergence proof of  $Q$ -learning is mainly based on stochastic approximation theories. Recall that the state-action Bellman operator  $\mathbf{F}$  is given as follows:

$$\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x' \mid x, a).$$

where  $V(x, y) = \min_{a \in \mathcal{A}} Q(x, y, a)$ . Therefore, the  $Q$ -update can be re-written as

$$\begin{aligned} Q_{k+1}(x, y, a) = & (1 - \zeta_k(x, y, a))Q_k(x, y, a) \\ & + \gamma \zeta_k(x, y, a) \left( \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_k](y\xi(x'))}{y} P(x' \mid x, a) + M_k(x, y, a) \right), \end{aligned}$$

where the noise term is given by

$$M_k(x, y, a) = \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P_{N_k}(\cdot \mid x, a))} \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\mathcal{I}_{x',i}[V_k](y\xi(x',i))}{y} - \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot \mid x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_k](y\xi(x'))}{y} P(x' \mid x, a), \quad (7.10)$$

for which  $M_k(x, y, a) \rightarrow 0$  almost surely as  $k \rightarrow \infty$  (consistency property whose proof follows from Section 7.1 D of [146] using the law of large numbers) and for any  $k \in \mathbb{N}$ ,

$$M_k^2(x, y, a) \leq \left| \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{\mathcal{I}_{x',i}[V_k](y\xi(x',i))}{y} - \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_k](y\xi(x'))}{y} \right|^2 \leq 2 \max_{x, y, a} Q_k^2(x, y, a).$$

Then the assumptions in Proposition 4.5 in [19] on the noise term  $M_k(x, y, a)$  are verified. Furthermore, following the same analysis from Lemma 2.4.3 that  $\mathbf{T}_{\mathcal{I}}$  is a contraction operator with respect to the  $\infty$ -norm, for any two state-action value functions  $Q_1(x, y, a)$  and  $Q_2(x, y, a)$  such that  $V_1(x, y) = \min_{a \in \mathcal{A}} Q_1(x, y, a)$  and  $V_2(x, y) = \min_{a \in \mathcal{A}} Q_2(x, y, a)$ , we have that

$$\begin{aligned}
& \left| C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_1](y\xi(x'))}{y} P(x'|x, a) - C(x, a) \right. \\
& \quad \left. - \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_2](y\xi(x'))}{y} P(x'|x, a) \right| \\
& \leq \gamma \left| \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_1](y\xi(x'))}{y} P(x'|x, a) - \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_2](y\xi(x'))}{y} P(x'|x, a) \right| \\
& \leq \gamma \left| \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \left[ \frac{\mathcal{I}_{x'}[V_1](y\xi(x'))}{y} - \frac{\mathcal{I}_{x'}[V_2](y\xi(x'))}{y} \right] P(x'|x, a) \right| \\
& \leq \gamma \max_{x, y} |V_1(x, y) - V_2(x, y)| \leq \gamma \|Q_1 - Q_2\|_{\infty}.
\end{aligned} \tag{7.11}$$

The second inequality follows from sub-additivity of max-operator:

$$\begin{aligned}
& \left| \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_1](y\xi(x'))}{y} P(x'|x, a) - \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_2](y\xi(x'))}{y} P(x'|x, a) \right| \\
& \leq \left| \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \left[ \frac{\mathcal{I}_{x'}[V_1](y\xi(x'))}{y} - \frac{\mathcal{I}_{x'}[V_2](y\xi(x'))}{y} \right] P(x'|x, a) \right|.
\end{aligned}$$

The third inequality is due to the same lines of arguments in Lemma 2.4.3. Therefore the above expression implies that  $\|\mathbf{F}_{\mathcal{I}}[Q_1] - \mathbf{F}_{\mathcal{I}}[Q_2]\|_{\infty} \leq \gamma \|Q_1 - Q_2\|_{\infty}$ , i.e.,  $\mathbf{F}_{\mathcal{I}}$  is a contraction mapping.

By combining these arguments, all assumptions in Proposition 4.5 in [19] are justified. This in turns implies the convergence of  $\{Q_k(x, y, a)\}_{k \in \mathbb{N}}$  to  $Q^*(x, y, a)$  component-wise, where  $Q^*$  is the unique fixed-point solution of  $\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$ .

### 7.1.9 Proof of Theorem 2.5.2

Similar to the proof of Theorem 2.5.1, the  $Q$ -update in asynchronous  $Q$ -learning can be written as:

$$Q_{k+1}(x, y, a) = (1 - \zeta_k(x, y, a))Q_k(x, y, a) + \zeta_k(x, y, a)(\Theta_k(x, y, a) + \Psi_k(x, y, a)),$$

where

$$\Theta_k(x, y, a) = \begin{cases} C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V_k](y\xi(x'))}{y} P(x'|x, a) & \text{if } (x, y, a) = (x_k(y), y, a_k(y)) \\ Q_k(x, y, a) & \text{otherwise} \end{cases}$$

with  $V_k(x, y) = \min_{a \in \mathcal{A}} Q_k(x, y, a)$  and the noise term is given by

$$\Psi_k(x, y, a) = \begin{cases} M_k(x, y, a) & \text{if } (x, y, a) = (x_k(y), y, a_k(y)) \\ 0 & \text{otherwise} \end{cases}$$

with  $M_k$  defined in (7.10). Since  $M_k(x, y, a) \rightarrow 0$  as  $k \rightarrow \infty$ , it can also be seen that  $\Psi_k(x, y, a) \rightarrow 0$  as  $k \rightarrow \infty$ . Furthermore, for any  $k \in \mathbb{N}$ , we also have that  $\Psi_k^2(x, y, a) \leq M_k^2(x, y, a) \leq 2 \max_{x, y, a} Q_k^2(x, y, a)$ . Then the assumptions in Proposition 4.5 in [19] on the noise term  $M_k(x, y, a)$  are verified. Now we define the asynchronous Bellman operator

$$\widehat{\mathbf{F}}_{\mathcal{I}}[Q](x, y, a) = \begin{cases} C(x, a) + \gamma \max_{\xi \in \mathcal{U}_{\text{CVaR}}(y, P(\cdot|x, a))} \sum_{x' \in \mathcal{X}} \frac{\mathcal{I}_{x'}[V](y\xi(x'))}{y} P(x'|x, a) & \text{if } (x, y, a) = (x_k(y), y, a_k(y)) \\ Q(x, y, a) & \text{otherwise} \end{cases},$$

where  $V(x, y) = \min_{a \in \mathcal{A}} Q(x, y, a)$ . It can easily checked that the fixed-point solution of  $\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$ , i.e.,  $Q^*$ , is also a fixed-point solution of  $\widehat{\mathbf{F}}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$ . Next we want to show that  $\widehat{\mathbf{F}}_{\mathcal{I}}[Q]$  is a contraction operator with respect to  $\infty$ -norm. Let  $\{\ell_k\}$  be a strictly increasing sequence ( $\ell_k < \ell_{k+1}$  for all  $k$ ) such that  $\ell_0 = 0$ , and at every CVaR confidence level  $y \in \mathbf{Y}$  every state-action pair  $(x, a)$  in  $\mathcal{X} \times \mathcal{A}$  is being updated at least once during this time period. Since every state action pair is visited infinitely often, Borel-Cantelli lemma [116] implies that for each finite  $k$ , both  $\ell_k$  and  $\ell_{k+1}$  are finite. For any  $\ell \in [\ell_k, \ell_{k+1}]$ , the result in (7.11) implies the following expression:

$$\begin{aligned} |\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell+1}[Q](x, y, a) - Q^*(x, y, a)| &\leq \gamma \left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{\ell}[Q] - Q^* \right\|_{\infty} & \text{if } (x, y, a) = (x_k(y), y, a_k(y)) \\ |\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell+1}[Q](x, y, a) - Q^*(x, y, a)| &= |\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell}[Q](x, y, a) - Q^*(x, y, a)| & \text{otherwise} \end{aligned}$$

From this result, one can first conclude that  $\widehat{\mathbf{F}}_{\mathcal{I}}[Q]$  is a non-expansive operator, i.e.,

$$|\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell+1}[Q](x, y, a) - Q^*(x, y, a)| \leq \left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{\ell}[Q] - Q^* \right\|_{\infty}.$$

Let  $l(x, y, a)$  be the last index strictly between  $\ell_k$  and  $\ell_{k+1}$  where the state-action pair  $(x, y, a)$  is updated.

Then

$$|\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_{k+1}}[Q](x, y, a) - Q^*(x, y, a)| \leq \gamma \left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{l(x, y, a)}[Q] - Q^* \right\|_{\infty}$$

From the definition of  $\ell_{k+1}$ , it is obvious that  $\ell_k < \max_{x, y, a} l(x, y, a) < \ell_{k+1}$ . The non-expansive property of  $\widehat{\mathbf{F}}_{\mathcal{I}}$  also implies that

$$\left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{l(x, y, a)}[Q] - Q^* \right\|_{\infty} \leq \left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_k}[Q] - Q^* \right\|_{\infty}.$$

Therefore we have that

$$|\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_{k+1}}[Q](x, y, a) - Q^*(x, y, a)| \leq \gamma \left\| \widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_k}[Q] - Q^* \right\|_{\infty}.$$

Combining these arguments implies that  $\|\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_{k+1}}[Q] - Q^*\|_{\infty} \leq \gamma \|\widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_k}[Q] - Q^*\|_{\infty}$ . Thus for  $\delta_k = \ell_{k+1} - \ell_k > 1$  and  $Q_k(x, y, a) = \widehat{\mathbf{F}}_{\mathcal{I}}^{\ell_k}[Q](x, y, a)$ , the following contraction property holds:

$$\|\widehat{\mathbf{F}}_{\mathcal{I}}^{\delta_k}[Q_k] - Q^*\|_{\infty} \leq \gamma \|Q_k - Q^*\|_{\infty}, \quad (7.12)$$

where the following fixed-point equation holds:  $\widehat{\mathbf{F}}_{\mathcal{I}}^{\delta_k}[Q^*](x, u) = Q^*(x, u)$ . Then by Proposition 4.5 in [19], the sequence  $\{Q_k(x, y, a)\}_{k \in \mathbb{N}}$  converges to  $Q^*(x, y, a)$  component-wise, where  $Q^*$  is the unique fixed-point solution of both  $\mathbf{F}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$  and  $\widehat{\mathbf{F}}_{\mathcal{I}}[Q](x, y, a) = Q(x, y, a)$ .

### 7.1.10 Proof of Theorem 2.6.1

First we want to show that

$$\rho_{\delta, \beta}(Z \mid h_t, \pi) \geq \max_{\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot | x_t, a_t))} \mathbb{E}[\xi(x_{t+1}) \cdot \rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) \mid h_t, \pi], \quad (7.13)$$

where the risk envelop is given by

$$\mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot | x_t, a_t)) = \left\{ \xi : \xi_i(x_{t+1}) \in [\delta, \beta^{-1}], \sum_{x_{t+1} \in \mathcal{X}} \xi_i(x_{t+1}) P(x_{t+1} | x_t, a_t) = 1, \forall i \right\}.$$

Recall that  $a_t$  is the control input induced by  $\mu_t(h_t)$  for any  $t \geq 0$ . From the dual representation theorem of coherent risk [132], one obtains

$$\rho_{\delta\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) = \max_{\xi' \in \mathcal{U}_{\text{Mix}}(\delta/\xi(x_{t+1}), \beta\xi(x_{t+1}), P(\cdot | x_{t+1}, a_{t+1}))} \mathbb{E}_P[\xi' Z \mid h_{t+1}, \pi]. \quad (7.14)$$

For any feasible  $\xi'$  in  $\mathcal{U}_{\text{Mix}}(\delta/\xi(x_{t+1}), \beta\xi(x_{t+1}), P(\cdot | x_{t+1}, a_{t+1}))$  where  $a_{t+1}$  is the control input induced by  $\mu_{t+1}(h_{t+1})$  and any feasible  $\xi$  in  $\mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot | x_t, a_t))$ , we have that

$$\delta \leq \xi'(x_{t+2})\xi(x_{t+1}) \leq \frac{1}{\beta}, \quad \forall x_{t+2} \in \mathcal{X}, \quad \sum_{x_{t+2} \in \mathcal{X}} \xi'_j(x_{t+2}) P(x_{t+2} | x_{t+1}, a_{t+1}) = 1.$$

Now since the following equality holds:

$$\mathbb{E}[\xi\xi' | h_t, \pi] = \sum_{x_{t+1} \in \mathcal{X}} \xi(x_{t+1}) \sum_{x_{t+2} \in \mathcal{X}} \xi'(x_{t+2}) P(x_{t+2} | x_{t+1}, a_{t+1}) P(x_{t+1} | x_t, a_t) = 1,$$

we can then show that  $\xi\xi'$  is in  $\mathcal{U}_{\text{Mix}}(\delta, \beta, \mathbb{P}(\cdot | x_t, a_t))$ . Furthermore based on the dual representation theorem of  $\rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi)$  from (7.14), for any  $\epsilon > 0$  there exists

$$\tilde{\xi}' \in \mathcal{U}_{\text{Mix}}(\delta/\xi(x_{t+1}), \beta\xi(x_{t+1}), P(\cdot | x_{t+1}, a_{t+1}))$$

such that

$$\rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) \leq \mathbb{E}_P[\tilde{\xi}'Z \mid h_{t+1}, \pi] + \epsilon.$$

This immediately implies the following inequality for any  $\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))$ :

$$\begin{aligned} \max_{\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))} \mathbb{E}_{\mathbb{P}}[\xi Z \mid h_t, \pi] &\geq \mathbb{E}_{\mathbb{P}}[\xi \tilde{\xi}'Z \mid h_t, \pi] \\ &\geq \mathbb{E}[\xi(x_{t+1}) \cdot \rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) \mid h_t, \pi] - \epsilon. \end{aligned}$$

By taking the supremum on the right side of the above inequality over  $\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))$ , using the dual representation theorem (7.14) and letting  $\epsilon \rightarrow 0$ , one obtains the inequality in (7.13).

Second we want to show that

$$\rho_{\delta, \beta}(Z \mid h_t, \pi) \leq \max_{\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))} \mathbb{E}[\xi(x_{t+1}) \cdot \rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) \mid h_t, \pi]. \quad (7.15)$$

We first choose some  $\xi^* \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))$  such that  $\mathbb{E}[\xi^*Z \mid h_t, \pi] = \rho_{\delta, \beta}(Z \mid h_t, \pi)$ . Furthermore define  $\tilde{\xi}_i(x_{t+1}) = \sum_{x_{t+2} \in \mathcal{X}} \xi_i^*(x_{t+1}, x_{t+2})P(x_{t+2} \mid x_{t+1}, a_{t+1})$ , where one immediately sees that the following properties hold:

$$\delta \leq \tilde{\xi}_i(x_{t+1}) \leq \frac{1}{\beta}, \quad \sum_{x_{t+1} \in \mathcal{X}} \tilde{\xi}_i(x_{t+1})P(x_{t+1} \mid x_t, a_t) = 1.$$

On the other hand, by defining

$$\xi(x_{t+1}, x_{t+2}) = \begin{cases} \frac{\xi^*(x_{t+1}, x_{t+2})}{\tilde{\xi}(x_{t+1})} & \text{if } \tilde{\xi}(x_{t+1}) > 0 \\ 1 & \text{otherwise} \end{cases},$$

the following properties hold as well:

$$\frac{\delta}{\tilde{\xi}(x_{t+1})} \leq \xi(x_{t+1}, x_{t+2}) \leq \frac{1}{\beta\tilde{\xi}(x_{t+1})}, \quad \forall i, \quad \sum_{x_{t+2} \in \mathcal{X}} \xi_i(x_{t+1}, x_{t+2})P(x_{t+2} \mid x_{t+1}, a_{t+1}) = 1.$$

Utilizing the above construction of  $\tilde{\xi}$ , we have the following chain of inequalities:

$$\begin{aligned} \rho_{\delta, \beta}(Z \mid h_t, \pi) &= \mathbb{E}[\xi^*Z \mid h_t, \pi] = \mathbb{E}_{\mathbb{P}}[\tilde{\xi}\xi'Z \mid h_t, \pi] \\ &= \mathbb{E}[\xi(x_{t+1})\mathbb{E}[\xi'Z \mid \mathcal{F}_{t+1}] \mid h_t, \pi] \\ &\leq \mathbb{E}[\tilde{\xi}(x_{t+1}) \cdot \rho_{\delta/\tilde{\xi}(x_{t+1}), \beta\tilde{\xi}(x_{t+1})}(Z \mid h_{t+1}, \pi) \mid h_t, \pi] \\ &\leq \max_{\xi \in \mathcal{U}_{\text{Mix}}(\delta, \beta, P(\cdot|x_t, a_t))} \mathbb{E}[\xi(x_{t+1}) \cdot \rho_{\delta/\xi(x_{t+1}), \beta\xi(x_{t+1})}(Z \mid h_{t+1}, \pi) \mid h_t, \pi]. \end{aligned}$$

Therefore the claim of this theorem is concluded by combining both arguments in (7.13) and (7.15).

## 7.2 Technical Results in Chapter 3: Policy Gradient Methods

In this section we present the convergence proof to the risk constrained policy gradient method.

### 7.2.1 Computing the Gradients

**i)  $\nabla_\theta L(\nu, \theta, \lambda)$ : Gradient of  $L(\nu, \theta, \lambda)$  w.r.t.  $\theta$**  By expanding the expectations in the definition of the objective function  $L(\nu, \theta, \lambda)$  in (3.5), we obtain

$$L(\nu, \theta, \lambda) = \sum_{\xi} \mathbb{P}_{\theta}(\xi) \mathcal{C}(\xi) + \lambda \nu + \frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) (\mathcal{D}(\xi) - \nu)^+ - \lambda \beta.$$

By taking the gradient with respect to  $\theta$ , we have

$$\nabla_{\theta} L(\nu, \theta, \lambda) = \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) \mathcal{C}(\xi) + \frac{\lambda}{1-\alpha} \sum_{\xi} \nabla_{\theta} \mathbb{P}_{\theta}(\xi) (\mathcal{D}(\xi) - \nu)^+.$$

This gradient can be rewritten as

$$\nabla_{\theta} L(\nu, \theta, \lambda) = \sum_{\xi: \mathbb{P}_{\theta}(\xi) \neq 0} \mathbb{P}_{\theta}(\xi) \cdot \nabla_{\theta} \log \mathbb{P}_{\theta}(\xi) \left( \mathcal{C}(\xi) + \frac{\lambda}{1-\alpha} (\mathcal{D}(\xi) - \nu) \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\} \right), \quad (7.16)$$

where in the case of  $\mathbb{P}_{\theta}(\xi) \neq 0$ , the term  $\nabla_{\theta} \log \mathbb{P}_{\theta}(\xi)$  is given by:

$$\begin{aligned} \nabla_{\theta} \log \mathbb{P}_{\theta}(\xi) &= \nabla_{\theta} \left\{ \sum_{k=0}^{T-1} \log P(x_{k+1}|x_k, a_k) + \log \mu(a_k|x_k; \theta) + \log \mathbf{1}\{x_0 = x^0\} \right\} \\ &= \sum_{k=0}^{T-1} \nabla_{\theta} \log \mu(a_k|x_k; \theta) = \sum_{k=0}^{T-1} \frac{1}{\mu(a_k|x_k; \theta)} \nabla_{\theta} \mu(a_k|x_k; \theta). \end{aligned}$$

**ii)  $\partial_{\nu} L(\nu, \theta, \lambda)$ : Sub-differential of  $L(\nu, \theta, \lambda)$  w.r.t.  $\nu$**  From the definition of  $L(\nu, \theta, \lambda)$ , we can easily see that  $L(\nu, \theta, \lambda)$  is a convex function in  $\nu$  for any fixed  $\theta \in \Theta$ . Note that for every fixed  $\nu$  and any  $\nu'$ , we have

$$(\mathcal{D}(\xi) - \nu')^+ - (\mathcal{D}(\xi) - \nu)^+ \geq g \cdot (\nu' - \nu),$$

where  $g$  is any element in the set of sub-derivatives:

$$g \in \partial_{\nu} (\mathcal{D}(\xi) - \nu)^+ := \begin{cases} -1 & \text{if } \nu < \mathcal{D}(\xi), \\ -q : q \in [0, 1] & \text{if } \nu = \mathcal{D}(\xi), \\ 0 & \text{otherwise.} \end{cases}$$

Since  $L(\nu, \theta, \lambda)$  is finite-valued for any  $\nu \in \mathbb{R}$ , by the additive rule of sub-derivatives, we have

$$\partial_\nu L(\nu, \theta, \lambda) = \left\{ -\frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) \mathbf{1}\{\mathcal{D}(\xi) > \nu\} - \frac{\lambda q}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) \mathbf{1}\{\mathcal{D}(\xi) = \nu\} + \lambda \mid q \in [0, 1] \right\}. \quad (7.17)$$

In particular for  $q = 1$ , we may write the sub-gradient of  $L(\nu, \theta, \lambda)$  w.r.t.  $\nu$  as

$$\partial_\nu L(\nu, \theta, \lambda)|_{q=1} = \lambda - \frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) \cdot \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\}$$

or

$$\lambda - \frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) \cdot \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\} \in \partial_\nu L(\nu, \theta, \lambda).$$

**iii)  $\nabla_\lambda L(\nu, \theta, \lambda)$ : Gradient of  $L(\nu, \theta, \lambda)$  w.r.t.  $\lambda$**  Since  $L(\nu, \theta, \lambda)$  is a linear function in  $\lambda$ , one can express the gradient of  $L(\nu, \theta, \lambda)$  w.r.t.  $\lambda$  as follows:

$$\nabla_\lambda L(\nu, \theta, \lambda) = \nu - \beta + \frac{1}{1-\alpha} \sum_{\xi} \mathbb{P}_{\theta}(\xi) \cdot (\mathcal{D}(\xi) - \nu) \mathbf{1}\{\mathcal{D}(\xi) \geq \nu\}. \quad (7.18)$$

## 7.2.2 Proof of Convergence of the Policy Gradient Algorithm

In this section, we prove the convergence of the policy gradient algorithm (Algorithm 2).

Since  $\nu$  converges on the faster timescale than  $\theta$  and  $\lambda$ , the  $\nu$ -update can be rewritten by assuming  $(\theta, \lambda)$  as invariant quantities, i.e.,

$$\nu_{i+1} = \Gamma_N \left[ \nu_i - \zeta_3(i) \left( \lambda - \frac{\lambda}{(1-\alpha)N} \sum_{j=1}^N \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} \right) \right]. \quad (7.19)$$

Consider the continuous time dynamics of  $\nu$  defined using differential inclusion

$$\dot{\nu} \in \Upsilon_\nu [-g(\nu)], \quad \forall g(\nu) \in \partial_\nu L(\nu, \theta, \lambda), \quad (7.20)$$

where

$$\Upsilon_\nu [K(\nu)] := \lim_{0 < \eta \rightarrow 0} \frac{\Gamma_N(\nu + \eta K(\nu)) - \Gamma_N(\nu)}{\eta}.$$

Here  $\Upsilon_\nu [K(\nu)]$  is the left directional derivative of the function  $\Gamma_N(\nu)$  in the direction of  $K(\nu)$ . By using the left directional derivative  $\Upsilon_\nu [-g(\nu)]$  in the sub-gradient descent algorithm for  $\nu$ , the gradient will point in the descent direction along the boundary of  $\nu$  whenever the  $\nu$ -update hits its boundary.

Furthermore, since  $\nu$  converges on a faster timescale than  $\theta$ , and  $\lambda$  is on the slowest time-scale, the  $\theta$ -update can be rewritten using the converged  $\nu^*(\theta)$ , assuming  $\lambda$  as an invariant quantity, i.e.,



$$\begin{aligned} \theta_{i+1} = & \Gamma_{\Theta} \left[ \theta_i - \zeta_2(i) \left( \frac{1}{N} \sum_{j=1}^N \nabla_{\theta} \log \mathbb{P}_{\theta}(\xi_{j,i})|_{\theta=\theta_i} \mathcal{C}(\xi_{j,i}) \right. \right. \\ & \left. \left. + \frac{\lambda}{(1-\alpha)N} \sum_{j=1}^N \nabla_{\theta} \log \mathbb{P}_{\theta}(\xi_{j,i})|_{\theta=\theta_i} (\mathcal{D}(\xi_{j,i}) - \nu) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu^*(\theta_i)\} \right) \right]. \end{aligned}$$

Consider the continuous time dynamics of  $\theta \in \Theta$ :

$$\dot{\theta} = \Upsilon_{\theta} [-\nabla_{\theta} L(\nu, \theta, \lambda)]|_{\nu=\nu^*(\theta)}, \quad (7.21)$$

where

$$\Upsilon_{\theta}[K(\theta)] := \lim_{0 < \eta \rightarrow 0} \frac{\Gamma_{\Theta}(\theta + \eta K(\theta)) - \Gamma_{\Theta}(\theta)}{\eta}.$$

Similar to the analysis of  $\nu$ ,  $\Upsilon_{\theta}[K(\theta)]$  is the left directional derivative of the function  $\Gamma_{\Theta}(\theta)$  in the direction of  $K(\theta)$ . By using the left directional derivative  $\Upsilon_{\theta} [-\nabla_{\theta} L(\nu, \theta, \lambda)]$  in the gradient descent algorithm for  $\theta$ , the gradient will point in the descent direction along the boundary of  $\Theta$  whenever the  $\theta$ -update hits its boundary.

Finally, since the  $\lambda$ -update converges in the slowest time-scale, the  $\lambda$ -update can be rewritten using the converged  $\theta^*(\lambda)$  and  $\nu^*(\lambda)$ , i.e.,

$$\lambda_{i+1} = \Gamma_{\Lambda} \left( \lambda_i + \zeta_1(i) \left( \nu^*(\lambda_i) + \frac{1}{1-\alpha} \frac{1}{N} \sum_{j=1}^N (\mathcal{D}(\xi_{j,i}) - \nu^*(\lambda_i))^+ - \beta \right) \right). \quad (7.22)$$

Consider the continuous time system

$$\dot{\lambda}(t) = \Upsilon_{\lambda} \left[ \nabla_{\lambda} L(\nu, \theta, \lambda) \right]_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)}, \quad \lambda(t) \geq 0, \quad (7.23)$$

where

$$\Upsilon_{\lambda}[K(\lambda)] := \lim_{0 < \eta \rightarrow 0} \frac{\Gamma_{\Lambda}(\lambda + \eta K(\lambda)) - \Gamma_{\Lambda}(\lambda)}{\eta}.$$

Again, similar to the analysis of  $(\nu, \theta)$ ,  $\Upsilon_{\lambda}[K(\lambda)]$  is the left directional derivative of the function  $\Gamma_{\Lambda}(\lambda)$  in the direction of  $K(\lambda)$ . By using the left directional derivative  $\Upsilon_{\lambda} [\nabla_{\lambda} L(\nu, \theta, \lambda)]$  in the gradient ascent algorithm for  $\lambda$ , the gradient will point in the ascent direction along the boundary of  $[0, \lambda_{\max}]$  whenever the  $\lambda$ -update hits its boundary.

Define

$$L^*(\lambda) = L(\nu^*(\lambda), \theta^*(\lambda), \lambda),$$

for  $\lambda \geq 0$  where  $(\theta^*(\lambda), \nu^*(\lambda)) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  is a local minimum of  $L(\nu, \theta, \lambda)$  for fixed  $\lambda \geq 0$ , i.e.,

$L(\nu, \theta, \lambda) \geq L(\nu^*(\lambda), \theta^*(\lambda), \lambda)$  for any  $(\theta, \nu) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] \cap \mathcal{B}_{(\theta^*(\lambda), \nu^*(\lambda))}(r)$  for some  $r > 0$ .

Next, we want to show that the ODE (7.23) is actually a gradient ascent of the Lagrangian function using the envelope theorem from mathematical economics [86]. The envelope theorem describes sufficient conditions for the derivative of  $L^*$  with respect to  $\lambda$  to equal the partial derivative of the objective function  $L$  with respect to  $\lambda$ , holding  $(\theta, \nu)$  at its local optimum  $(\theta, \nu) = (\theta^*(\lambda), \nu^*(\lambda))$ . We will show that  $\nabla_\lambda L^*(\lambda)$  coincides with  $\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)}$  as follows.

**Theorem 7.2.1.** *The value function  $L^*$  is absolutely continuous. Furthermore,*

$$L^*(\lambda) = L^*(0) + \int_0^\lambda \nabla_{\lambda'} L(\nu, \theta, \lambda') \Big|_{\theta=\theta^*(s), \nu=\nu^*(s), \lambda'=s} ds, \quad \lambda \geq 0. \quad (7.24)$$

*Proof.* The proof follows from analogous arguments to Lemma 4.3 in [34]. From the definition of  $L^*$ , observe that for any  $\lambda', \lambda'' \geq 0$  with  $\lambda' < \lambda''$ ,

$$\begin{aligned} |L^*(\lambda'') - L^*(\lambda')| &\leq \sup_{\theta \in \Theta, \nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]} |L(\nu, \theta, \lambda'') - L(\nu, \theta, \lambda')| \\ &= \sup_{\theta \in \Theta, \nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]} \left| \int_{\lambda'}^{\lambda''} \nabla_\lambda L(\nu, \theta, s) ds \right| \\ &\leq \int_{\lambda'}^{\lambda''} \sup_{\theta \in \Theta, \nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]} |\nabla_\lambda L(\nu, \theta, s)| ds \leq \frac{3D_{\max}}{(1-\alpha)(1-\gamma)} (\lambda'' - \lambda'). \end{aligned}$$

This implies that  $L^*$  is absolutely continuous. Therefore,  $L^*$  is continuous everywhere and differentiable almost everywhere.

By the Milgrom–Segal envelope theorem in mathematical economics (Theorem 1 of [86]), one concludes that the derivative of  $L^*(\lambda)$  coincides with the derivative of  $L(\nu, \theta, \lambda)$  at the point of differentiability  $\lambda$  and  $\theta = \theta^*(\lambda)$ ,  $\nu = \nu^*(\lambda)$ . Also since  $L^*$  is absolutely continuous, the limit of  $(L^*(\lambda) - L^*(\lambda'))/(\lambda - \lambda')$  at  $\lambda \uparrow \lambda'$  (or  $\lambda \downarrow \lambda'$ ) coincides with the lower/upper directional derivatives if  $\lambda'$  is a point of non-differentiability. Thus, there is only a countable number of non-differentiable points in  $L^*$  and the set of non-differentiable points of  $L^*$  has measure zero. Therefore, expression (7.24) holds and one concludes that  $\nabla_\lambda L^*(\lambda)$  coincides with  $\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)}$ .  $\square$

Before getting into the main result, we have the following technical proposition whose proof directly follows from the definition of  $\log \mathbb{P}_\theta(\xi)$  and Assumption 3.2.3 that  $\nabla_\theta \mu(a_k | x_k; \theta)$  is Lipschitz in  $\theta$ .

**Proposition 7.2.2.**  *$\nabla_\theta L(\nu, \theta, \lambda)$  is Lipschitz in  $\theta$ .*

**Remark 7.2.3.** *The fact that  $\nabla_\theta L(\nu, \theta, \lambda)$  is Lipschitz in  $\theta$  implies that  $\|\nabla_\theta L(\nu, \theta, \lambda)\|^2 \leq 2(\|\nabla_\theta L(\nu, \theta_0, \lambda)\| + \|\theta_0\|)^2 + 2\|\theta\|^2$  which further implies that*

$$\|\nabla_\theta L(\nu, \theta, \lambda)\|^2 \leq K_1(1 + \|\theta\|^2).$$

for  $K_1 = 2 \max(1, (\|\nabla_\theta L(\nu, \theta_0, \lambda)\| + \|\theta_0\|)^2) > 0$ . Similarly, the fact that  $\nabla_\theta \log \mathbb{P}_\theta(\xi)$  is Lipschitz implies that

$$\|\nabla_\theta \log \mathbb{P}_\theta(\xi)\|^2 \leq K_2(\xi)(1 + \|\theta\|^2)$$

for a positive random variable  $K_2(\xi)$ . Furthermore, since  $T < \infty$  w.p. 1,  $\mu(a_k|x_k; \theta) \in (0, 1]$  and  $\nabla_\theta \mu(a_k|x_k; \theta)$  is Lipschitz for any  $k < T$ ,  $K_2(\xi) < \infty$  w.p. 1.

**Remark 7.2.4.** For any given  $\theta \in \Theta$ ,  $\lambda \geq 0$ , and  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ , we have

$$|g(\nu)| \leq 3\lambda(1 + |\nu|)/(1 - \alpha). \quad (7.25)$$

To see this, recall that the set of  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$  can be parameterized by  $q \in [0, 1]$  as

$$g(\nu; q) = -\frac{\lambda}{(1 - \alpha)} \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) > \nu\} - \frac{\lambda q}{1 - \alpha} \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) = \nu\} + \lambda.$$

It is obvious that  $|\mathbf{1}\{\mathcal{D}(\xi) = \nu\}|, |\mathbf{1}\{\mathcal{D}(\xi) > \nu\}| \leq 1 + |\nu|$ . Thus,  $\left| \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) > \nu\} \right| \leq \sup_{\xi} |\mathbf{1}\{\mathcal{D}(\xi) > \nu\}| \leq 1 + |\nu|$ , and  $\left| \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) = \nu\} \right| \leq 1 + |\nu|$ . Recalling that  $0 < (1 - q)$ ,  $(1 - \alpha) < 1$ , these arguments imply the claim of (7.25).

We are now in a position to prove the convergence analysis of Theorem 3.3.2.

*Proof of Theorem 3.3.2.* We split the proof into the following four steps:

**Step 1 (Convergence of  $\nu$ -update)** Since  $\nu$  converges on a faster time scale than  $\theta$  and  $\lambda$ , one can take both  $\theta$  and  $\lambda$  as fixed quantities in the  $\nu$ -update, i.e.,

$$\nu_{i+1} = \Gamma_N \left( \nu_i + \zeta_3(i) \left( \frac{\lambda}{(1 - \alpha)N} \sum_{j=1}^N \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} - \lambda + \delta\nu_{i+1} \right) \right), \quad (7.26)$$

and the Martingale difference term with respect to  $\nu$  is given by

$$\delta\nu_{i+1} = \frac{\lambda}{1 - \alpha} \left( -\frac{1}{N} \sum_{j=1}^N \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu_i\} + \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) \geq \nu_i\} \right). \quad (7.27)$$

First, one can show that  $\delta\nu_{i+1}$  is square integrable, i.e.,

$$\mathbb{E}[\|\delta\nu_{i+1}\|^2 \mid \mathcal{F}_{\nu,i}] \leq 4 \left( \frac{\lambda_{\max}}{1 - \alpha} \right)^2$$

where  $\mathcal{F}_{\nu,i} = \sigma(\nu_m, \delta\nu_m, m \leq i)$  is the filtration of  $\nu_i$  generated by different independent trajectories.

Second, since the history trajectories are generated based on the sampling probability mass function  $\mathbb{P}_\theta(\xi)$ , expression (7.17) implies that  $\mathbb{E}[\delta\nu_{i+1} \mid \mathcal{F}_{\nu,i}] = 0$ . Therefore, the  $\nu$ -update is a stochastic approximation of the ODE (7.20) with a Martingale difference error term, i.e.,

$$\frac{\lambda}{1-\alpha} \sum_{\xi} \mathbb{P}_\theta(\xi) \mathbf{1}\{\mathcal{D}(\xi) \geq \nu_i\} - \lambda \in -\partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_i}.$$

Then one can invoke Corollary 4 in Chapter 5 of [35] (stochastic approximation theory for non-differentiable systems) to show that the sequence  $\{\nu_i\}$ ,  $\nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  converges almost surely to a fixed point  $\nu^* \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  of the differential inclusion (7.20), where

$$\nu^* \in N_c := \left\{ \nu \in \left[ -\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma} \right] : \Upsilon_\nu[-g(\nu)] = 0, g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \right\}.$$

To justify the assumptions of this corollary, 1) from Remark 7.2.4, the Lipschitz property is satisfied, i.e.,  $\sup_{g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)} |g(\nu)| \leq 3\lambda(1 + |\nu|)/(1 - \alpha)$ , 2)  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  and  $\partial_\nu L(\nu, \theta, \lambda)$  are convex compact sets by definition, which implies  $\{(\nu, g(\nu)) \mid g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)\}$  is a closed set, and further implies  $\partial_\nu L(\nu, \theta, \lambda)$  is an upper semi-continuous set valued mapping, 3) the step-size rule follows from Assumption 3.3.1, 4) the Martingale difference assumption follows from (7.27), and 5)  $\nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ ,  $\forall i$  implies that  $\sup_i \|\nu_i\| < \infty$  almost surely.

Consider the ODE for  $\nu \in \mathbb{R}$  in (7.20), we define the set-valued derivative of  $L$  as follows:

$$D_t L(\nu, \theta, \lambda) = \{g(\nu) \Upsilon_\nu[-g(\nu)] \mid \forall g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)\}.$$

One can conclude that

$$\max_{g(\nu)} D_t L(\nu, \theta, \lambda) = \max \{g(\nu) \Upsilon_\nu[-g(\nu)] \mid g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)\}.$$

We now show that  $\max_{g(\nu)} D_t L(\nu, \theta, \lambda) \leq 0$  and this quantity is non-zero if  $\Upsilon_\nu[-g(\nu)] \neq 0$  for every  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$  by considering three cases. To distinguish the latter two cases, we need to define,

$$\mathcal{G}(\nu) := \left\{ g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \mid \forall \eta_0 > 0, \exists \eta \in (0, \eta_0] \text{ such that } \theta - \eta g(\nu) \notin \left[ -\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma} \right] \right\}.$$

*Case 1:*  $\nu \in (-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma})$ .

For every  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ , there exists a sufficiently small  $\eta_0 > 0$  such that  $\nu - \eta_0 g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  and

$$\Gamma_N(\theta - \eta_0 g(\nu)) - \theta = -\eta_0 g(\nu).$$

Therefore, the definition of  $\Upsilon_\theta[-g(\nu)]$  implies

$$\max_{g(\nu)} D_t L(\nu, \theta, \lambda) = \max \{ -g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \} \leq 0. \quad (7.28)$$

The maximum is attained because  $\partial_\nu L(\nu, \theta, \lambda)$  is a convex compact set and  $g(\nu)\Upsilon_\nu[-g(\nu)]$  is a continuous function. At the same time, we have  $\max_{g(\nu)} D_t L(\nu, \theta, \lambda) < 0$  whenever  $0 \notin \partial_\nu L(\nu, \theta, \lambda)$ .

*Case 2:*  $\nu \in \{-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\}$  and  $\mathcal{G}(\nu)$  is empty.

The condition  $\nu - \eta g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  implies that

$$\Upsilon_\nu[-g(\nu)] = -g(\nu).$$

Then we obtain

$$\max_{g(\nu)} D_t L(\nu, \theta, \lambda) = \max \{ -g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \} \leq 0. \quad (7.29)$$

Furthermore, we have  $\max_{g(\nu)} D_t L(\nu, \theta, \lambda) < 0$  whenever  $0 \notin \partial_\nu L(\nu, \theta, \lambda)$ .

*Case 3:*  $\nu \in \{-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\}$  and  $\mathcal{G}(\nu)$  is nonempty.

First, consider any  $g(\nu) \in \mathcal{G}(\nu)$ . For any  $\eta > 0$ , define  $\nu_\eta := \nu - \eta g(\nu)$ . The above condition implies that when  $0 < \eta \rightarrow 0$ ,  $\Gamma_N[\nu_\eta]$  is the projection of  $\nu_\eta$  to the tangent space of  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ . For any element  $\hat{\nu} \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ , since the set  $\{\nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] : \|\nu - \nu_\eta\|_2 \leq \|\hat{\nu} - \nu_\eta\|_2\}$  is compact, the projection of  $\nu_\eta$  on  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  exists. Furthermore, since  $f(\nu) := \frac{1}{2}(\nu - \nu_\eta)^2$  is a strongly convex function and  $\nabla f(\nu) = \nu - \nu_\eta$ , by the first order optimality condition, one obtains

$$\nabla f(\nu_\eta^*)(\nu - \nu_\eta^*) = (\nu_\eta^* - \nu_\eta)(\nu - \nu_\eta^*) \geq 0, \quad \forall \nu \in \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right]$$

where  $\nu_\eta^*$  is the unique projection of  $\nu_\eta$  (the projection is unique because  $f(\nu)$  is strongly convex and  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  is a convex compact set). Since the projection (minimizer) is unique, the above equality holds if and only if  $\nu = \nu_\eta^*$ .

Therefore, for any  $\nu \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  and  $\eta > 0$ ,

$$\begin{aligned} g(\nu)\Upsilon_\nu[-g(\nu)] &= g(\nu) \left( \lim_{0 < \eta \rightarrow 0} \frac{\nu_\eta^* - \nu}{\eta} \right) \\ &= \left( \lim_{0 < \eta \rightarrow 0} \frac{\nu - \nu_\eta}{\eta} \right) \left( \lim_{0 < \eta \rightarrow 0} \frac{\nu_\eta^* - \nu}{\eta} \right) = \lim_{0 < \eta \rightarrow 0} \frac{-\|\nu_\eta^* - \nu\|^2}{\eta^2} + \lim_{0 < \eta \rightarrow 0} (\nu_\eta^* - \nu_\eta) \left( \frac{\nu_\eta^* - \nu}{\eta^2} \right) \leq 0. \end{aligned}$$

Second, for any  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \cap \mathcal{G}(\nu)^c$ , one obtains  $\nu - \eta g(\nu) \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ , for any  $\eta \in (0, \eta_0]$  and some  $\eta_0 > 0$ . In this case, the arguments follow from case 2 and the following expression holds:  $\Upsilon_\nu[-g(\nu)] = -g(\nu)$ .

Combining these arguments, one concludes that

$$\begin{aligned} & \max_{g(\nu)} D_t L(\nu, \theta, \lambda) \\ & \leq \max \left\{ \max \left\{ g(\nu) \Upsilon_\nu[-g(\nu)] \mid g(\nu) \in \mathcal{G}(\nu) \right\}, \max \left\{ -g^2(\nu) \mid g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \cap \mathcal{G}(\nu)^c \right\} \right\} \leq 0. \end{aligned} \quad (7.30)$$

This quantity is non-zero whenever  $0 \notin \{g(\nu) \Upsilon_\nu[-g(\nu)] \mid \forall g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)\}$  (this is because, for any  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda) \cap \mathcal{G}(\nu)^c$ , one obtains  $g(\nu) \Upsilon_\nu[-g(\nu)] = -g(\nu)^2$ ). Thus, by similar arguments one may conclude that  $\max_{g(\nu)} D_t L(\nu, \theta, \lambda) \leq 0$  and it is non-zero if  $\Upsilon_\nu[-g(\nu)] \neq 0$  for every  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ .

Now for any given  $\theta$  and  $\lambda$ , define the following Lyapunov function

$$\mathcal{L}_{\theta, \lambda}(\nu) = L(\nu, \theta, \lambda) - L(\nu^*, \theta, \lambda)$$

where  $\nu^*$  is a minimum point (for any given  $(\theta, \lambda)$ ,  $L$  is a convex function in  $\nu$ ). Then  $\mathcal{L}_{\theta, \lambda}(\nu)$  is a positive definite function, i.e.,  $\mathcal{L}_{\theta, \lambda}(\nu) \geq 0$ . On the other hand, by the definition of a minimum point, one easily obtains  $0 \in \{g(\nu^*) \Upsilon_\nu[-g(\nu^*)] \mid \nu = \nu^* \mid \forall g(\nu^*) \in \partial_\nu L(\nu, \theta, \lambda) \mid \nu = \nu^*\}$  which means that  $\nu^*$  is also a stationary point, i.e.,  $\nu^* \in N_c$ .

Note that  $\max_{g(\nu)} D_t \mathcal{L}_{\theta, \lambda}(\nu) = \max_{g(\nu)} D_t L(\nu, \theta, \lambda) \leq 0$  and this quantity is non-zero if  $\Upsilon_\nu[-g(\nu)] \neq 0$  for every  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$ . Therefore, by the Lyapunov theory for asymptotically stable differential inclusions (see Theorem 3.10 and Corollary 3.11 in [14], where the Lyapunov function  $\mathcal{L}_{\theta, \lambda}(\nu)$  satisfies Hypothesis 3.1 and the property in (7.30) is equivalent to Hypothesis 3.9 in the reference), the above arguments imply that with any initial condition  $\nu(0)$ , the state trajectory  $\nu(t)$  of (7.20) converges to  $\nu^*$ , i.e.,  $L(\nu^*, \theta, \lambda) \leq L(\nu(t), \theta, \lambda) \leq L(\nu(0), \theta, \lambda)$  for any  $t \geq 0$ .

As stated earlier, the sequence  $\{\nu_i\}$ ,  $\nu_i \in [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$  constitutes a stochastic approximation to the differential inclusion (7.20), and thus converges almost surely its solution [35], which further converges almost surely to  $\nu^* \in N_c$ . Also, it can be easily seen that  $N_c$  is a closed subset of the compact set  $[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ , and therefore a compact set itself.

**Step 2 (Convergence of  $\theta$ -update)** Since  $\theta$  converges on a faster time scale than  $\lambda$  and  $\nu$  converges faster than  $\theta$ , one can take  $\lambda$  as a fixed quantity and  $\nu$  as a converged quantity  $\nu^*(\theta)$  in the  $\theta$ -update. The  $\theta$ -update can be rewritten as a stochastic approximation, i.e.,

$$\theta_{i+1} = \Gamma_\Theta \left( \theta_i + \zeta_2(i) \left( -\nabla_\theta L(\nu, \theta, \lambda) \big|_{\theta=\theta_i, \nu=\nu^*(\theta_i)} + \delta\theta_{i+1} \right) \right), \quad (7.31)$$

where

$$\begin{aligned} \delta\theta_{i+1} = & \nabla_\theta L(\nu, \theta, \lambda) \big|_{\theta=\theta_i, \nu=\nu^*(\theta_i)} - \frac{1}{N} \sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i}) \big|_{\theta=\theta_i} \mathcal{C}(\xi_{j,i}) \\ & - \frac{\lambda}{(1-\alpha)N} \sum_{j=1}^N \nabla_\theta \log \mathbb{P}_\theta(\xi_{j,i}) \big|_{\theta=\theta_i} (\mathcal{D}(\xi_{j,i}) - \nu^*(\theta_i)) \mathbf{1}\{\mathcal{D}(\xi_{j,i}) \geq \nu^*(\theta_i)\}. \end{aligned} \quad (7.32)$$

First, one can show that  $\delta\theta_{i+1}$  is square integrable, i.e.,  $\mathbb{E}[\|\delta\theta_{i+1}\|^2 \mid \mathcal{F}_{\theta,i}] \leq K_i(1 + \|\theta_i\|^2)$  for some  $K_i > 0$ , where  $\mathcal{F}_{\theta,i} = \sigma(\theta_m, \delta\theta_m, m \leq i)$  is the filtration of  $\theta_i$  generated by different independent trajectories. To see this, notice that

$$\begin{aligned}
& \|\delta\theta_{i+1}\|^2 \\
& \leq 2 \left( \nabla_{\theta} L(\nu, \theta, \lambda) \Big|_{\theta=\theta_i, \nu=\nu^*(\theta_i)} \right)^2 + \frac{2}{N^2} \left( \frac{C_{\max}}{1-\gamma} + \frac{2\lambda D_{\max}}{(1-\alpha)(1-\gamma)} \right)^2 \left( \sum_{j=1}^N \nabla_{\theta} \log \mathbb{P}_{\theta}(\xi_{j,i}) \Big|_{\theta=\theta_i} \right)^2 \\
& \leq 2K_{1,i}(1 + \|\theta_i\|^2) + \frac{2^N}{N^2} \left( \frac{C_{\max}}{1-\gamma} + \frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)} \right)^2 \left( \sum_{j=1}^N \|\nabla_{\theta} \log \mathbb{P}_{\theta}(\xi_{j,i}) \Big|_{\theta=\theta_i}\|^2 \right) \\
& \leq 2K_{1,i}(1 + \|\theta_i\|^2) + \frac{2^N}{N^2} \left( \frac{C_{\max}}{1-\gamma} + \frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)} \right)^2 \left( \sum_{j=1}^N K_2(\xi_{j,i})(1 + \|\theta_i\|^2) \right) \\
& \leq 2 \left( K_{1,i} + \frac{2^{N-1}}{N} \left( \frac{C_{\max}}{1-\gamma} + \frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)} \right)^2 \max_{1 \leq j \leq N} K_2(\xi_{j,i}) \right) (1 + \|\theta_i\|^2).
\end{aligned}$$

The Lipschitz upper bounds are due to the results in Remark 7.2.3. Since  $K_2(\xi_{j,i}) < \infty$  w.p. 1, there exists  $K_{2,i} < \infty$  such that  $\max_{1 \leq j \leq N} K_2(\xi_{j,i}) \leq K_{2,i}$ . By combining these results, one concludes that  $\mathbb{E}[\|\delta\theta_{i+1}\|^2 \mid \mathcal{F}_{\theta,i}] \leq K_i(1 + \|\theta_i\|^2)$  where

$$K_i = 2 \left( K_{1,i} + \frac{2^{N-1} K_{2,i}}{N} \left( \frac{C_{\max}}{1-\gamma} + \frac{2\lambda_{\max} D_{\max}}{(1-\alpha)(1-\gamma)} \right)^2 \right) < \infty.$$

Second, since the history trajectories are generated based on the sampling probability mass function  $\mathbb{P}_{\theta_i}(\xi)$ , expression (7.16) implies that  $\mathbb{E}[\delta\theta_{i+1} \mid \mathcal{F}_{\theta,i}] = 0$ . Therefore, the  $\theta$ -update is a stochastic approximation of the ODE (7.21) with a Martingale difference error term. In addition, from the convergence analysis of the  $\nu$ -update,  $\nu^*(\theta)$  is an asymptotically stable equilibrium point for the sequence  $\{\nu_i\}$ . From (7.17),  $\partial_{\nu} L(\nu, \theta, \lambda)$  is a Lipschitz set-valued mapping in  $\theta$  (since  $\mathbb{P}_{\theta}(\xi)$  is Lipschitz in  $\theta$ ), and thus it can be easily seen that  $\nu^*(\theta)$  is a Lipschitz continuous mapping of  $\theta$ .

Now consider the continuous time dynamics for  $\theta \in \Theta$ , given in (7.21). We may write

$$\frac{dL(\nu, \theta, \lambda)}{dt} \Big|_{\nu=\nu^*(\theta)} = \left( \nabla_{\theta} L(\nu, \theta, \lambda) \Big|_{\nu=\nu^*(\theta)} \right)^{\top} \Upsilon_{\theta} \left[ -\nabla_{\theta} L(\nu, \theta, \lambda) \Big|_{\nu=\nu^*(\theta)} \right]. \quad (7.33)$$

By considering the following cases, we now show that  $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} \leq 0$  and this quantity is non-zero whenever  $\|\Upsilon_{\theta} [-\nabla_{\theta} L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}]\| \neq 0$ .

*Case 1: When  $\theta \in \Theta^{\circ} = \Theta \setminus \partial\Theta$ .*

Since  $\Theta^{\circ}$  is the interior of the set  $\Theta$  and  $\Theta$  is a convex compact set, there exists a sufficiently small  $\eta_0 > 0$

such that  $\theta - \eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$  and

$$\Gamma_\Theta(\theta - \eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}) - \theta = -\eta_0 \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}.$$

Therefore, the definition of  $\Upsilon_\theta[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}]$  implies

$$\left. \frac{dL(\nu, \theta, \lambda)}{dt} \right|_{\nu=\nu^*(\theta)} = -\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\|^2 \leq 0. \quad (7.34)$$

At the same time, we have  $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} < 0$  whenever  $\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\| \neq 0$ .

*Case 2: When  $\theta \in \partial\Theta$  and  $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$  for any  $\eta \in (0, \eta_0]$  and some  $\eta_0 > 0$ .*

The condition  $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \in \Theta$  implies that

$$\Upsilon_\theta[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}] = -\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}.$$

Then we obtain

$$\left. \frac{dL(\nu, \theta, \lambda)}{dt} \right|_{\nu=\nu^*(\theta)} = -\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\|^2 \leq 0. \quad (7.35)$$

Furthermore,  $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} < 0$  when  $\|\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}\| \neq 0$ .

*Case 3: When  $\theta \in \partial\Theta$  and  $\theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)} \notin \Theta$  for some  $\eta \in (0, \eta_0]$  and any  $\eta_0 > 0$ .*

For any  $\eta > 0$ , define  $\theta_\eta := \theta - \eta \nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}$ . The above condition implies that when  $0 < \eta \rightarrow 0$ ,  $\Gamma_\Theta[\theta_\eta]$  is the projection of  $\theta_\eta$  to the tangent space of  $\Theta$ . For any element  $\hat{\theta} \in \Theta$ , since the set  $\{\theta \in \Theta : \|\theta - \theta_\eta\|_2 \leq \|\hat{\theta} - \theta_\eta\|_2\}$  is compact, the projection of  $\theta_\eta$  on  $\Theta$  exists. Furthermore, since  $f(\theta) := \frac{1}{2}\|\theta - \theta_\eta\|_2^2$  is a strongly convex function and  $\nabla f(\theta) = \theta - \theta_\eta$ , by the first order optimality condition, one obtains

$$\nabla f(\theta_\eta^*)^\top (\theta - \theta_\eta^*) = (\theta_\eta^* - \theta_\eta)^\top (\theta - \theta_\eta^*) \geq 0, \quad \forall \theta \in \Theta,$$

where  $\theta_\eta^*$  is the unique projection of  $\theta_\eta$  (the projection is unique because  $f(\theta)$  is strongly convex and  $\Theta$  is a convex compact set). Since the projection (minimizer) is unique, the above equality holds if and only if  $\theta = \theta_\eta^*$ .

Therefore, for any  $\theta \in \Theta$  and  $\eta > 0$ ,

$$\begin{aligned} & (\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)})^\top \Upsilon_\theta[-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}] = (\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)})^\top \left( \lim_{0 < \eta \rightarrow 0} \frac{\theta_\eta^* - \theta}{\eta} \right) \\ & = \left( \lim_{0 < \eta \rightarrow 0} \frac{\theta - \theta_\eta}{\eta} \right)^\top \left( \lim_{0 < \eta \rightarrow 0} \frac{\theta_\eta^* - \theta}{\eta} \right) = \lim_{0 < \eta \rightarrow 0} \frac{-\|\theta_\eta^* - \theta\|^2}{\eta^2} + \lim_{0 < \eta \rightarrow 0} (\theta_\eta^* - \theta)^\top \left( \frac{\theta_\eta^* - \theta}{\eta^2} \right) \leq 0. \end{aligned}$$



By combining these arguments, one concludes that  $dL(\nu, \theta, \lambda)/dt|_{\nu=\nu^*(\theta)} \leq 0$  and this quantity is non-zero whenever  $\|\Upsilon_\theta [-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}]\| \neq 0$ .

Now, for any given  $\lambda$ , define the Lyapunov function

$$\mathcal{L}_\lambda(\theta) = L(\nu^*(\theta), \theta, \lambda) - L(\nu^*(\theta^*), \theta^*, \lambda),$$

where  $\theta^*$  is a local minimum point. Then there exists a ball centered at  $\theta^*$  with radius  $r$  such that for any  $\theta \in \mathcal{B}_{\theta^*}(r)$ ,  $\mathcal{L}_\lambda(\theta)$  is a locally positive definite function, i.e.,  $\mathcal{L}_\lambda(\theta) \geq 0$ . On the other hand, by the definition of a local minimum point, one obtains  $\Upsilon_\theta [-\nabla_\theta L(\theta^*, \nu, \lambda)|_{\nu=\nu^*(\theta^*)}]|_{\theta=\theta^*} = 0$  which means that  $\theta^*$  is a stationary point, i.e.,  $\theta^* \in \Theta_c$ .

Note that  $d\mathcal{L}_\lambda(\theta(t))/dt = dL(\theta(t), \nu^*(\theta(t)), \lambda)/dt \leq 0$  and the time-derivative is non-zero whenever

$$\|\Upsilon_\theta [-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}]\| \neq 0.$$

Therefore, by the Lyapunov theory for asymptotically stable systems [68], the above arguments imply that with any initial condition  $\theta(0) \in \mathcal{B}_{\theta^*}(r)$ , the state trajectory  $\theta(t)$  of (7.21) converges to  $\theta^*$ , i.e.,

$$L(\theta^*, \nu^*(\theta^*), \lambda) \leq L(\theta(t), \nu^*(\theta(t)), \lambda) \leq L(\theta(0), \nu^*(\theta(0)), \lambda)$$

for any  $t \geq 0$ .

Based on the above properties and noting that 1) from Proposition 7.2.2,  $\nabla_\theta L(\nu, \theta, \lambda)$  is a Lipschitz function in  $\theta$ , 2) the step-size rule follows from Assumption 3.3.1, 3) expression (7.37) implies that  $\delta\theta_{i+1}$  is a square integrable Martingale difference, and 4)  $\theta_i \in \Theta, \forall i$  implies that  $\sup_i \|\theta_i\| < \infty$  almost surely, one can invoke Theorem 2 in Chapter 6 of [35] (multi-time scale stochastic approximation theory) to show that the sequence  $\{\theta_i\}$ ,  $\theta_i \in \Theta$  converges almost surely to the solution of the ODE (7.21), which further converges almost surely to  $\theta^* \in \Theta$ .

**Step 3 (Local Minimum)** Now, we want to show that the sequence  $\{\theta_i, \nu_i\}$  converges to a local minimum of  $L(\nu, \theta, \lambda)$  for any fixed  $\lambda$ . Recall that  $\{\theta_i, \nu_i\}$  converges to  $(\theta^*, \nu^*) := (\theta^*, \nu^*(\theta^*))$ . Previous arguments on the  $(\nu, \theta)$ -convergence imply that with any initial condition  $(\theta(0), \nu(0))$ , the state trajectories  $\theta(t)$  and  $\nu(t)$  of (7.20) and (7.21) converge to the set of stationary points  $(\theta^*, \nu^*)$  in the positive invariant set  $\Theta_c \times N_c$  and

$$L(\theta^*, \nu^*, \lambda) \leq L(\theta(t), \nu^*(\theta(t)), \lambda) \leq L(\theta(0), \nu^*(\theta(0)), \lambda) \leq L(\theta(0), \nu(t), \lambda) \leq L(\theta(0), \nu(0), \lambda)$$

for any  $t \geq 0$ .

By contradiction, suppose  $(\theta^*, \nu^*)$  is not a local minimum. Then there exists  $(\bar{\theta}, \bar{\nu}) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] \cap$

$\mathcal{B}_{(\theta^*, \nu^*)}(r)$  such that

$$L(\bar{\theta}, \bar{\nu}, \lambda) = \min_{(\theta, \nu) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r)} L(\nu, \theta, \lambda).$$

The minimum is attained by the Weierstrass extreme value theorem. By putting  $\theta(0) = \bar{\theta}$ , the above arguments imply that

$$L(\bar{\theta}, \bar{\nu}, \lambda) = \min_{(\theta, \nu) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r)} L(\nu, \theta, \lambda) < L(\theta^*, \nu^*, \lambda) \leq L(\bar{\theta}, \bar{\nu}, \lambda)$$

which is a contradiction. Therefore, the stationary point  $(\theta^*, \nu^*)$  is a local minimum of  $L(\nu, \theta, \lambda)$  as well.

**Step 4 (Convergence of  $\lambda$ -update)** Since the  $\lambda$ -update converges in the slowest time scale, it can be rewritten using the converged  $\theta^*(\lambda) = \theta^*(\nu^*(\lambda), \lambda)$  and  $\nu^*(\lambda)$ , i.e.,

$$\lambda_{i+1} = \Gamma_{\Lambda} \left( \lambda_i + \zeta_1(i) \left( \nabla_{\lambda} L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda_i), \nu=\nu^*(\lambda_i), \lambda=\lambda_i} + \delta\lambda_{i+1} \right) \right) \quad (7.36)$$

where

$$\delta\lambda_{i+1} = -\nabla_{\lambda} L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda_i} + \left( \nu^*(\lambda_i) + \frac{1}{1-\alpha} \frac{1}{N} \sum_{j=1}^N (\mathcal{D}(\xi_{j,i}) - \nu^*(\lambda_i))^+ - \beta \right). \quad (7.37)$$

From (7.18), we see that  $\nabla_{\lambda} L(\nu, \theta, \lambda)$  is a constant function of  $\lambda$ . Similar to the  $\theta$ -update, one can easily show that  $\delta\lambda_{i+1}$  is square integrable, i.e.,

$$\mathbb{E}[\|\delta\lambda_{i+1}\|^2 \mid \mathcal{F}_{\lambda,i}] \leq 2 \left( \beta + \frac{3D_{\max}}{(1-\gamma)(1-\alpha)} \right)^2,$$

where  $\mathcal{F}_{\lambda,i} = \sigma(\lambda_m, \delta\lambda_m, m \leq i)$  is the filtration of  $\lambda$  generated by different independent trajectories. Furthermore, expression (7.18) implies that  $\mathbb{E}[\delta\lambda_{i+1} \mid \mathcal{F}_{\lambda,i}] = 0$ . Therefore, the  $\lambda$ -update is a stochastic approximation of the ODE (7.23) with a Martingale difference error term. In addition, from the convergence analysis of the  $(\theta, \nu)$ -update,  $(\theta^*(\lambda), \nu^*(\lambda))$  is an asymptotically stable equilibrium point for the sequence  $\{\theta_i, \nu_i\}$ . From (7.16),  $\nabla_{\theta} L(\nu, \theta, \lambda)$  is a linear mapping in  $\lambda$ , and  $(\theta^*(\lambda), \nu^*(\lambda))$  is a Lipschitz continuous mapping of  $\lambda$ .

Consider the ODE for  $\lambda \in [0, \lambda_{\max}]$  in (7.23). Analogous to the arguments for the  $\theta$ -update, we can write

$$\frac{d(-L(\nu, \theta, \lambda))}{dt} \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)} = -\nabla_{\lambda} L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)} \Upsilon_{\lambda} \left[ \nabla_{\lambda} L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)} \right],$$

and show that  $-dL(\nu, \theta, \lambda)/dt|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)} \leq 0$ . This quantity is non-zero whenever

$$\|\Upsilon_\lambda [dL(\nu, \theta, \lambda)/d\lambda|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)}]\| \neq 0.$$

Consider the Lyapunov function

$$\mathcal{L}(\lambda) = -L(\theta^*(\lambda), \nu^*(\lambda), \lambda) + L(\theta^*(\lambda^*), \nu^*(\lambda^*), \lambda^*)$$

where  $\lambda^*$  is a local maximum point. Then there exists a ball centered at  $\lambda^*$  with radius  $r$  such that for any  $\lambda \in \mathcal{B}_{\lambda^*}(r)$ ,  $\mathcal{L}(\lambda)$  is a locally positive definite function, i.e.,  $\mathcal{L}(\lambda) \geq 0$ . On the other hand, by the definition of a local maximum point, one obtains

$$\Upsilon_\lambda [dL(\nu, \theta, \lambda)/d\lambda|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}]|_{\lambda=\lambda^*} = 0$$

which means that  $\lambda^*$  is also a stationary point, i.e.,  $\lambda^* \in \Lambda_c$ . Since

$$\frac{d\mathcal{L}(\lambda(t))}{dt} = -\frac{dL(\theta^*(\lambda(t)), \nu^*(\lambda(t)), \lambda(t))}{dt} \leq 0$$

and the time-derivative is non-zero whenever  $\|\Upsilon_\lambda [\nabla_\lambda L(\nu, \theta, \lambda)|_{\nu=\nu^*(\lambda), \theta=\theta^*(\lambda)}]\| \neq 0$ , the Lyapunov theory for asymptotically stable systems implies that  $\lambda(t)$  converges to  $\lambda^*$ .

Given the above results and noting that the step size rule is selected according to Assumption 3.3.1, one can apply the multi-time scale stochastic approximation theory (Theorem 2 in Chapter 6 of [35]) to show that the sequence  $\{\lambda_i\}$  converges almost surely to the solution of the ODE (7.23), which further converges almost surely to  $\lambda^* \in [0, \lambda_{\max}]$ . Since  $[0, \lambda_{\max}]$  is a compact set, following the same lines of arguments and recalling the envelope theorem (Theorem 7.2.1) for local optima, one further concludes that  $\lambda^*$  is a local maximum of  $L(\theta^*(\lambda), \nu^*(\lambda), \lambda) = L^*(\lambda)$ .

**Step 5 (Local Optima)** By letting  $\theta^* = \theta^*(\nu^*(\lambda^*), \lambda^*)$  and  $\nu^* = \nu^*(\lambda^*)$ , we will show that  $\theta^*$  is a locally optimal policy for the CVaR-constrained optimization problem, which constitutes a (local) saddle point  $(\theta^*, \nu^*, \lambda^*)$  of the Lagrangian function  $L(\nu, \theta, \lambda)$  if  $\lambda^* \in [0, \lambda_{\max}]$ .

Suppose the sequence  $\{\lambda_i\}$  generated from (7.36) converges to a stationary point  $\lambda^* \in [0, \lambda_{\max}]$ . Since step 3 implies that  $(\theta^*, \nu^*)$  is a local minimum of  $L(\nu, \theta, \lambda^*)$  over the feasible set  $(\theta, \nu) \in \Theta \times [-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}]$ , there exists a  $r > 0$  such that

$$L(\theta^*, \nu^*, \lambda^*) \leq L(\nu, \theta, \lambda^*), \quad \forall (\theta, \nu) \in \Theta \times \left[-\frac{D_{\max}}{1-\gamma}, \frac{D_{\max}}{1-\gamma}\right] \cap \mathcal{B}_{(\theta^*, \nu^*)}(r).$$

In order to complete the proof, we must show

$$\nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] \leq \beta, \quad (7.38)$$

and

$$\lambda^* \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] - \beta \right) = 0. \quad (7.39)$$

These two equations imply

$$\begin{aligned} L(\theta^*, \nu^*, \lambda^*) &= V^{\theta^*}(x^0) + \lambda^* \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] - \beta \right) \\ &= V^{\theta^*}(x^0) \\ &\geq V^{\theta^*}(x^0) + \lambda \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] - \beta \right) = L(\theta^*, \nu^*, \lambda), \end{aligned}$$

which further implies that  $(\theta^*, \nu^*, \lambda^*)$  is a saddle point of  $L(\nu, \theta, \lambda)$ . We now show that (7.38) and (7.39) hold.

Recall that

$$\Upsilon_\lambda \left[ \nabla_\lambda L(\nu, \theta, \lambda) |_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*} \right] |_{\lambda=\lambda^*} = 0.$$

We show (7.38) by contradiction. Suppose

$$\nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] > \beta.$$

This implies that for  $\lambda^* \in [0, \lambda_{\max})$ , we have

$$\Gamma_\Lambda \left( \lambda^* - \eta \left( \beta - \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] \right) \right) \right) = \lambda^* - \eta \left( \beta - \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] \right) \right)$$

for any  $\eta \in (0, \eta_{\max}]$ , for some sufficiently small  $\eta_{\max} > 0$ . Therefore,

$$\Upsilon_\lambda \left[ \nabla_\lambda L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*} \right] \Big|_{\lambda=\lambda^*} = \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] - \beta > 0.$$

This is in contradiction with the fact that  $\Upsilon_\lambda \left[ \nabla_\lambda L(\nu, \theta, \lambda) |_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*} \right] |_{\lambda=\lambda^*} = 0$ . Therefore, (7.38) holds.

To show that (7.39) holds, we only need to show that  $\lambda^* = 0$  if

$$\nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] < \beta.$$

Suppose  $\lambda^* \in (0, \lambda_{\max})$ , then there exists a sufficiently small  $\eta_0 > 0$  such that

$$\begin{aligned} &\frac{1}{\eta_0} \left( \Gamma_\Lambda \left( \lambda^* - \eta_0 \left( \beta - \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] \right) \right) \right) - \Gamma_\Lambda(\lambda^*) \right) \\ &= \nu^* + \frac{1}{1-\alpha} \mathbb{E} \left[ (\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+ \right] - \beta < 0. \end{aligned}$$

This again contradicts the assumption  $\Upsilon_\lambda [\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}]|_{\lambda=\lambda^*} = 0$ . Therefore (7.39) holds.

When  $\lambda^* = \lambda_{\max}$  and  $\nu^* + \frac{1}{1-\alpha} \mathbb{E}[(\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+] > \beta$ ,

$$\Gamma_\Lambda \left( \lambda^* - \eta \left( \beta - \left( \nu^* + \frac{1}{1-\alpha} \mathbb{E}[(\mathcal{D}^{\theta^*}(x^0) - \nu^*)^+] \right) \right) \right) = \lambda_{\max}$$

for any  $\eta > 0$  and

$$\Upsilon_\lambda [\nabla_\lambda L(\nu, \theta, \lambda)|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda^*}]|_{\lambda=\lambda^*} = 0.$$

In this case one cannot guarantee feasibility using the above analysis, and  $(\theta^*, \nu^*, \lambda^*)$  is not a local saddle point. Such a  $\lambda^*$  is referred to as a spurious fixed point [73]. Notice that  $\lambda^*$  is bounded (otherwise we can conclude that the problem is infeasible), so that by incrementally increasing  $\lambda_{\max}$  in Algorithm 2, we can always prevent ourselves from obtaining a spurious fixed point solution.

Combining the above arguments, we finally conclude that  $\theta^*$  is a locally optimal policy.  $\square$

### 7.3 Technical Results in Chapter 3: Actor-Critic Algorithms

In this section we present the convergence proof to the risk constrained actor-critic method. Recall from Assumption 3.3.1 that the SPSA step size  $\{\Delta_k\}$  satisfies  $\Delta_k \rightarrow 0$  as  $k \rightarrow \infty$  and  $\sum_k (\zeta_2(k)/\Delta_k)^2 < \infty$ .

#### 7.3.1 Gradient with Respect to $\lambda$ (Proof of Lemma 3.4.4)

By taking the gradient of  $V^\theta(x^0, \nu)$  w.r.t.  $\lambda$  (recall that both  $V$  and  $Q$  depend on  $\lambda$  through the cost function  $\bar{C}$  of the augmented MDP  $\bar{\mathcal{M}}$ ), we obtain

$$\begin{aligned}
 \nabla_\lambda V^\theta(x^0, \nu) &= \sum_{a \in \bar{\mathcal{A}}} \mu(a|x^0, \nu; \theta) \nabla_\lambda Q^\theta(x^0, \nu, a) \\
 &= \sum_{a \in \bar{\mathcal{A}}} \mu(a|x^0, \nu; \theta) \nabla_\lambda \left[ \bar{C}(x^0, \nu, a) + \sum_{(x', s') \in \bar{\mathcal{X}}} \gamma \bar{P}(x', s'|x^0, \nu, a) V^\theta(x', s') \right] \\
 &= \underbrace{\sum_a \mu(a|x^0, \nu; \theta) \nabla_\lambda \bar{C}(x^0, \nu, a)}_{h(x^0, \nu)} + \gamma \sum_{a, x', s'} \mu(a|x^0, \nu; \theta) \bar{P}(x', s'|x^0, \nu, a) \nabla_\lambda V^\theta(x', s') \\
 &= h(x^0, \nu) + \gamma \sum_{a, x', s'} \mu(a|x^0, \nu; \theta) \bar{P}(x', s'|x^0, \nu, a) \nabla_\lambda V^\theta(x', s') \\
 &= h(x^0, \nu) + \gamma \sum_{a, x', s'} \mu(a|x^0, \nu; \theta) \bar{P}(x', s'|x^0, \nu, a) \left[ h(x', s') \right. \\
 &\quad \left. + \gamma \sum_{a', x'', s''} \mu(a'|x', s'; \theta) \bar{P}(x'', s''|x', s', a') \nabla_\lambda V^\theta(x'', s'') \right].
 \end{aligned} \tag{7.40}$$

By unrolling the last equation using the definition of  $\nabla_\lambda V^\theta(x, s)$  from (7.40), we obtain

$$\begin{aligned}
 \nabla_\lambda V^\theta(x^0, \nu) &= \sum_{k=0}^{\infty} \gamma^k \sum_{x, s} \mathbb{P}(x_k = x, s_k = s \mid x_0 = x^0, s_0 = \nu; \theta) h(x, s) \\
 &= \frac{1}{1-\gamma} \sum_{x, s} d_\gamma^\theta(x, s|x^0, \nu) h(x, s) = \frac{1}{1-\gamma} \sum_{x, s, a} d_\gamma^\theta(x, s|x^0, \nu) \mu(a|x, s) \nabla_\lambda \bar{C}(x, s, a) \\
 &= \frac{1}{1-\gamma} \sum_{x, s, a} \pi_\gamma^\theta(x, s, a|x^0, \nu) \nabla_\lambda \bar{C}(x, s, a) \\
 &= \frac{1}{1-\gamma} \sum_{x, s, a} \pi_\gamma^\theta(x, s, a|x^0, \nu) \frac{1}{1-\alpha} \mathbf{1}\{x = x_{\text{Tar}}\} (-s)^+.
 \end{aligned}$$

This completes the proof.

## 7.3.2 Proof of Convergence of the Actor-Critic Algorithms

### 7.3.2.1 Proof of Theorem 3.4.3: Critic Update ( $v$ -update)

By the step size conditions, one notices that  $\{v_k\}$  converges on a faster time scale than  $\{\nu_k\}$ ,  $\{\theta_k\}$ , and  $\{\lambda_k\}$ . Thus, one can take  $(\nu, \theta, \lambda)$  in the  $v$ -update as fixed quantities. The critic update can be re-written as follows:

$$v_{k+1} = v_k + \zeta_4(k) \phi(x_k, s_k) \delta_k(v_k), \quad (7.41)$$

where the scalar

$$\delta_k(v_k) = -v_k^\top \phi(x_k, s_k) + \gamma v_k^\top \phi(x_{k+1}, s_{k+1}) + \bar{C}_\lambda(x_k, s_k, a_k)$$

is the temporal difference (TD) from (3.18). Define

$$A := \sum_{y, a', s'} \pi_\gamma^\theta(y, s', a' | x, s) \phi(y, s') \left( \phi^\top(y, s') - \gamma \sum_{z, s''} \bar{P}(z, s'' | y, s', a) \phi^\top(z, s'') \right), \quad (7.42)$$

and

$$b := \sum_{y, a', s'} \pi_\gamma^\theta(y, s', a' | x, s) \phi(y, s') \bar{C}_\lambda(y, s', a'). \quad (7.43)$$

It is easy to see that the critic update  $v_k$  in (7.41) can be re-written as the following stochastic approximation scheme:

$$v_{k+1} = v_k + \zeta_4(k) (b - Av_k + \delta A_{k+1}), \quad (7.44)$$

where the noise term  $\delta A_{k+1}$  is a square integrable Martingale difference, i.e.,  $\mathbb{E}[\delta A_{k+1} | \mathcal{F}_k] = 0$  if the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$  is used to generate samples of  $(x_k, s_k, a_k)$ —with  $\mathcal{F}_k$  being the filtration generated by different independent trajectories. By writing

$$\delta A_{k+1} = -(b - Av_k) + \phi(x_k, s_k) \delta_k(v_k)$$

and noting  $\mathbb{E}_{\pi_\gamma^\theta}[\phi(x_k, s_k) \delta_k(v_k) | \mathcal{F}_k] = -Av_k + b$ , one can easily verify that the stochastic approximation scheme in (7.44) is equivalent to the critic iterates in (7.41) and  $\delta A_{k+1}$  is a Martingale difference, i.e.,  $\mathbb{E}_{\pi_\gamma^\theta}[\delta A_{k+1} | \mathcal{F}_k] = 0$ . Let

$$h(v) := -Av + b.$$

Before getting into the convergence analysis, we present a technical lemma whose proof can be found in [20, Lemma 6.10].

**Lemma 7.3.1.** *Every eigenvalue of the matrix  $A$  has positive real part.*

We now turn to the analysis of the critic iteration. Note that the following properties hold for the critic update scheme in (7.41): 1)  $h(v)$  is Lipschitz, 2) the step size satisfies the properties in Assumption 3.4.1, 3) the noise term  $\delta A_{k+1}$  is a square integrable Martingale difference, 4) the function  $h_c(v) := h(cv)/c$ ,  $c \geq 1$

converges uniformly to a continuous function  $h_\infty(v)$  for any  $v$  in a compact set, i.e.,  $h_c(v) \rightarrow h_\infty(v)$  as  $c \rightarrow \infty$ , and 5) the ordinary differential equation (ODE)  $\dot{v} = h_\infty(v)$  has the origin as its unique globally asymptotically stable equilibrium. The fourth property can be easily verified from the fact that the magnitude of  $b$  is finite and  $h_\infty(v) = -Av$ . The fifth property follows directly from the facts that  $h_\infty(v) = -Av$  and all eigenvalues of  $A$  have positive real parts.

By Theorem 3.1 in [35], these five properties imply:

The critic iterates  $\{v_k\}$  are bounded almost surely, i.e.,  $\sup_k \|v_k\| < \infty$  almost surely.

The convergence of the critic iterates in (7.41) can be related to the asymptotic behavior of the ODE

$$\dot{v} = h(v) = b - Av. \quad (7.45)$$

Specifically, Theorem 2 in Chapter 2 of [35] and the above conditions imply  $v_k \rightarrow v^*$  with probability 1, where the limit  $v^*$  depends on  $(\nu, \theta, \lambda)$  and is the unique solution satisfying  $h(v^*) = 0$ , i.e.,  $Av^* = b$ . Therefore, the critic iterates converge to the unique fixed point  $v^*$  almost surely, as  $k \rightarrow \infty$ .

### 7.3.2.2 Proof of Theorem 3.4.5

**Step 1 (Convergence of  $v$ -update)** The proof of convergence for the critic parameter follows directly from Theorem 3.4.3.

**Step 2 (Convergence of SPSA based  $\nu$ -update)** In this section, we analyze the  $\nu$ -update for the incremental actor-critic method. This update is based on the SPSA perturbation method. The idea of this method is to estimate the sub-gradient  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$  using two simulated value functions corresponding to  $\nu^- = \nu - \Delta$  and  $\nu^+ = \nu + \Delta$ . Here  $\Delta \geq 0$  is a positive random perturbation that vanishes asymptotically. The SPSA-based estimate for a sub-gradient  $g(\nu) \in \partial_\nu L(\nu, \theta, \lambda)$  is given by

$$g(\nu) \approx \lambda + \frac{1}{2\Delta} (\phi^\top(x^0, \nu + \Delta) - \phi^\top(x^0, \nu - \Delta)) v.$$

We turn to the convergence analysis of the sub-gradient estimation and  $\nu$ -update. Since  $v$  converges faster than  $\nu$ , and  $\nu$  converges faster than  $\theta$  and  $\lambda$ , the  $\nu$ -update in (3.20) can be rewritten using the converged critic parameter  $v^*(\nu)$ , i.e.,

$$\nu_{k+1} = \Gamma_N \left( \nu_k - \zeta_3(k) \left( \lambda + \frac{1}{2\Delta_k} (\phi^\top(x^0, \nu_k + \Delta_k) - \phi^\top(x^0, \nu_k - \Delta_k)) v^*(\nu_k) \right) \right), \quad (7.46)$$

where  $(\theta, \lambda)$  in this expression are viewed as constant quantities.



First, we consider the following assumption on the feature functions in order to prove that the SPSA approximation is asymptotically unbiased.

**Assumption 7.3.2.** *For any  $v \in \mathbb{R}^{\kappa_1}$ , the feature functions satisfy the following conditions*

$$|\phi_V^\top(x^0, \nu + \Delta) v - \phi_V^\top(x^0, \nu - \Delta) v| \leq K_1(v)(1 + \Delta).$$

Furthermore, the Lipschitz constants are uniformly bounded, i.e.,  $\sup_{v \in \mathbb{R}^{\kappa_1}} K_1^2(v) < \infty$ .

This assumption is mild as the expected utility objective function implies that  $L(\nu, \theta, \lambda)$  is Lipschitz in  $\nu$ , and  $\phi_V^\top(x^0, \nu) v$  is just a linear function approximation of  $V^\theta(x^0, \nu)$ .

Next, we establish the bias and convergence of the stochastic sub-gradient estimate. Let

$$\bar{g}(\nu_k) \in \arg \max \{g : g \in \partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_k}\},$$

and

$$\begin{aligned} \Lambda_{1,k+1} &= \left( \frac{(\phi^\top(x^0, \nu_k + \Delta_k) - \phi^\top(x^0, \nu_k - \Delta_k)) v^*(\nu_k)}{2\Delta_k} - E_M(k) \right), \\ \Lambda_{2,k} &= \lambda_k + E_M^L(k) - \bar{g}(\nu_k), \\ \Lambda_{3,k} &= E_M(k) - E_M^L(k), \end{aligned}$$

where

$$\begin{aligned} E_M(k) &:= \mathbb{E} \left[ \frac{1}{2\Delta_k} (\phi^\top(x^0, \nu_k + \Delta_k) - \phi^\top(x^0, \nu_k - \Delta_k)) v^*(\nu_k) \mid \Delta_k \right], \\ E_M^L(k) &:= \mathbb{E} \left[ \frac{1}{2\Delta_k} (V^\theta(x^0, \nu_k + \Delta_k) - V^\theta(x^0, \nu_k - \Delta_k)) \mid \Delta_k \right]. \end{aligned}$$

Note that (7.46) is equivalent to

$$\nu_{k+1} = \Gamma_N(\nu_k - \zeta_3(k)(\bar{g}(\nu_k) + \Lambda_{1,k+1} + \Lambda_{2,k} + \Lambda_{3,k})). \quad (7.47)$$

First, it is clear that  $\Lambda_{1,k+1}$  is a Martingale difference as  $\mathbb{E}[\Lambda_{1,k+1} \mid \mathcal{F}_k] = 0$ , which implies that

$$M_{k+1} = \sum_{j=0}^k \zeta_3(j) \Lambda_{1,j+1}$$

is a Martingale w.r.t. the filtration  $\mathcal{F}_k$ . By the Martingale convergence theorem, we can show that if  $\sup_{k \geq 0} \mathbb{E}[M_k^2] < \infty$ , when  $k \rightarrow \infty$ ,  $M_k$  converges almost surely and  $\zeta_3(k) \Lambda_{1,k+1} \rightarrow 0$  almost surely. To show that

$\sup_{k \geq 0} \mathbb{E}[M_k^2] < \infty$ , for any  $t \geq 0$  one observes that

$$\begin{aligned} \mathbb{E}[M_{k+1}^2] &= \sum_{j=0}^k (\zeta_3(j))^2 \mathbb{E}[\mathbb{E}[\Lambda_{1,j+1}^2 \mid \Delta_j]] \\ &\leq 2 \sum_{j=0}^k \mathbb{E} \left[ \left( \frac{\zeta_3(j)}{2\Delta_j} \right)^2 \left\{ \mathbb{E} \left[ \left( (\phi^\top(x^0, \nu_j + \Delta_j) - \phi^\top(x^0, \nu_j - \Delta_j)) v^*(\nu_j) \right)^2 \mid \Delta_j \right] \right. \right. \\ &\quad \left. \left. + \mathbb{E} \left[ (\phi^\top(x^0, \nu_j + \Delta_j) - \phi^\top(x^0, \nu_j - \Delta_j)) v^*(\nu_j) \mid \Delta_j \right]^2 \right\} \right]. \end{aligned}$$

Now based on Assumption 7.3.2, the above expression implies

$$\mathbb{E}[M_{k+1}^2] \leq 2 \sum_{j=0}^k \mathbb{E} \left[ \left( \frac{\zeta_3(j)}{2\Delta_j} \right)^2 2K_1^2(1 + \Delta_j)^2 \right].$$

Combining the above results with the step size conditions, there exists  $K = 4K_1^2 > 0$  such that

$$\sup_{k \geq 0} \mathbb{E}[M_{k+1}^2] \leq K \sum_{j=0}^{\infty} \mathbb{E} \left[ \left( \frac{\zeta_3(j)}{2\Delta_j} \right)^2 \right] + (\zeta_3(j))^2 < \infty.$$

Second, by the Min Common/Max Crossing theorem in [18], one can show that  $\partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_k}$  is a non-empty, convex, and compact set. Therefore, by duality of directional derivatives and sub-differentials, i.e.,

$$\max \{g : g \in \partial_\nu L(\nu, \theta, \lambda)|_{\nu=\nu_k}\} = \lim_{\xi \downarrow 0} \frac{L(\nu_k + \xi, \theta, \lambda) - L(\nu_k - \xi, \theta, \lambda)}{2\xi},$$

one concludes that for  $\lambda_k = \lambda$  (we can treat  $\lambda_k$  as a constant because it converges on a slower time scale than  $\nu_k$ ),

$$\lambda + E_M^L(k) = \bar{g}(\nu_k) + O(\Delta_k),$$

almost surely. This further implies that

$$\Lambda_{2,k} = O(\Delta_k), \quad \text{i.e., } \Lambda_{2,k} \rightarrow 0 \text{ as } k \rightarrow \infty,$$

almost surely.

Third, since  $d_\gamma^\theta(x^0, \nu | x^0, \nu) = 1$ , from the definition of  $\epsilon_\theta(v^*(\nu_k))$ ,

$$|\Lambda_{3,k}| \leq 2\epsilon_\theta(v^*(\nu_k))/\Delta_k.$$

As  $t$  goes to infinity,  $\epsilon_\theta(v^*(\nu_k))/\Delta_k \rightarrow 0$  by assumption and  $\Lambda_{3,k} \rightarrow 0$ .

Finally, since  $\zeta_3(k)\Lambda_{1,k+1} \rightarrow 0$ ,  $\Lambda_{2,k} \rightarrow 0$ , and  $\Lambda_{3,k} \rightarrow 0$  almost surely, the  $\nu$ -update in (7.47) is a noisy sub-gradient descent update with vanishing disturbance bias. Thus, the  $\nu$ -update in (3.20) can be viewed as

an Euler discretization of an element of the following differential inclusion,

$$\dot{\nu} \in \Upsilon_{\nu}[-g(\nu)], \quad \forall g(\nu) \in \partial_{\nu} L(\nu, \theta, \lambda), \quad (7.48)$$

and the  $\nu$ -convergence analysis is analogous to Step 1 of the proof of Theorem 3.3.2.

**Step 2' (Convergence of semi-trajectory  $\nu$ -update)** Since  $\nu$  converges on a faster timescale than  $\theta$  and  $\lambda$ , the  $\nu$ -update in (3.23) can be rewritten using a fixed pair  $(\theta, \lambda)$ , i.e.,

$$\nu_{k+1} = \Gamma_N \left( \nu_i - \zeta_3(k) \left( \lambda - \frac{\lambda}{1-\alpha} \left( \mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_k, \mu) + \delta\nu_{M,k+1} \right) \right) \right), \quad (7.49)$$

where

$$\delta\nu_{M,k+1} = -\mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_i, \mu) + \mathbf{1}\{x_k = x_{\text{Tar}}, s_k \leq 0\} \quad (7.50)$$

is a square integrable stochastic term, specifically,

$$\mathbb{E}[(\delta\nu_{M,k+1})^2 \mid \mathcal{F}_{\nu,k}] \leq 2,$$

where  $\mathcal{F}_{\nu,k} = \sigma(\nu_m, \delta\nu_m, m \leq k)$  is the filtration generated by  $\nu$ . Since  $\mathbb{E}[\delta\nu_{M,k+1} \mid \mathcal{F}_{\nu,k}] = 0$ ,  $\delta\nu_{M,k+1}$  is a Martingale difference and the  $\nu$ -update in (7.49) is a stochastic approximation of an element of the differential inclusion

$$\frac{\lambda}{1-\alpha} \mathbb{P}(s_{\text{Tar}} \leq 0 \mid x_0 = x^0, s_0 = \nu_k, \mu) - \lambda \in -\partial_{\nu} L(\nu, \theta, \lambda)|_{\nu=\nu_k}.$$

Thus, the  $\nu$ -update in (3.23) can be viewed as an Euler discretization of the differential inclusion in (7.48), and the  $\nu$ -convergence analysis is analogous to Step 1 of the proof of Theorem 3.3.2.

**Step 3 (Convergence of  $\theta$ -update)** We first analyze the actor update ( $\theta$ -update). Since  $\theta$  converges on a faster time scale than  $\lambda$ , one can take  $\lambda$  in the  $\theta$ -update as a fixed quantity. Furthermore, since  $v$  and  $\nu$  converge on a faster scale than  $\theta$ , one can also replace  $v$  and  $\nu$  with their limits  $v^*(\theta)$  and  $\nu^*(\theta)$  in the convergence analysis. In the following analysis, we assume that the initial state  $x^0 \in \mathcal{X}$  is given. Then the  $\theta$ -update in (3.21) can be rewritten as follows:

$$\theta_{k+1} = \Gamma_{\Theta} \left( \theta_k - \zeta_2(k) \left( \nabla_{\theta} \log \mu(a_k \mid x_k, s_k; \theta) \Big|_{\theta=\theta_k} \frac{\delta_k(v^*(\theta_k))}{1-\gamma} \right) \right). \quad (7.51)$$

Consider the case in which the value function for a fixed policy  $\mu$  is approximated by a learned function approximator,  $\phi^{\top}(x, s)v^*$ . If the approximation is sufficiently good, we might hope to use it in place of  $V^{\theta}(x, s)$  and still point roughly in the direction of the true gradient. Recall the temporal difference error

(random variable) for a given pair  $(x_k, s_k) \in \mathcal{X} \times \mathbb{R}$ :

$$\delta_k(v) = -v^\top \phi(x_k, s_k) + \gamma v^\top \phi(x_{k+1}, s_{k+1}) + \bar{C}_\lambda(x_k, s_k, a_k).$$

Define the  $v$ -dependent approximated advantage function

$$\tilde{A}^{\theta, v}(x, s, a) := \tilde{Q}^{\theta, v}(x, s, a) - v^\top \phi(x, s),$$

where

$$\tilde{Q}^{\theta, v}(x, s, a) = \gamma \sum_{x', s'} \bar{P}(x', s' | x, s, a) v^\top \phi(x', s') + \bar{C}_\lambda(x, s, a).$$

The following lemma, whose proof follows from the proof of Lemma 3 in [27], shows that  $\delta_k(v)$  is an unbiased estimator of  $\tilde{A}^{\theta, v}$ .

**Lemma 7.3.3.** *For any given policy  $\mu$  and  $v \in \mathbb{R}^{\kappa_1}$ , we have*

$$\tilde{A}^{\theta, v}(x, s, a) = \mathbb{E}[\delta_k(v) \mid x_k = x, s_k = s, a_k = a].$$

Define

$$\nabla_\theta \tilde{L}_v(\nu, \theta, \lambda) := \frac{1}{1 - \gamma} \sum_{x, a, s} \pi_\gamma^\theta(x, s, a | x_0 = x^0, s_0 = \nu) \nabla_\theta \log \mu(a | x, s; \theta) \tilde{A}^{\theta, v}(x, s, a)$$

as the linear function approximation of  $\nabla_\theta \tilde{L}(\nu, \theta, \lambda)$ . Similar to Proposition 7.2.2, we present the following technical lemma on the Lipschitz property of  $\nabla_\theta \tilde{L}_v(\nu, \theta, \lambda)$ .

**Proposition 7.3.4.**  *$\nabla_\theta \tilde{L}_v(\nu, \theta, \lambda)$  is a Lipschitz function in  $\theta$ .*

*Proof.* Consider the feature vector  $v$ . Recall that the feature vector satisfies the linear equation  $Av = b$ , where  $A$  and  $b$  are given by (7.42) and (7.43), respectively. from Lemma 1 in [25], by exploiting the inverse of  $A$  using Cramer's rule, one may show that  $v$  is continuously differentiable in  $\theta$ . Now consider the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$ . By applying Theorem 2 in [5] (or Theorem 3.1 in [134]), it can be seen that the occupation measure  $\pi_\gamma^\theta$  of the process  $(x_k, s_k)$  is continuously differentiable in  $\theta$ . Recall from Assumption 3.2.3 in Section 3.2.2 that  $\nabla_\theta \mu(a_k | x_k, s_k; \theta)$  is a Lipschitz function in  $\theta$  for any  $a \in \mathcal{A}$  and  $k \in \{0, \dots, T-1\}$ , and  $\mu(a_k | x_k, s_k; \theta)$  is differentiable in  $\theta$ . By combining these arguments and noting that the sum of products of Lipschitz functions is Lipschitz, one concludes that  $\nabla_\theta \tilde{L}_v(\nu, \theta, \lambda)$  is Lipschitz in  $\theta$ . ■

We turn to the convergence proof of  $\theta$ .

**Theorem 7.3.5.** *The sequence of  $\theta$ -updates in (3.21) converges almost surely to an equilibrium point  $\hat{\theta}^*$  that satisfies  $\Upsilon_\theta \left[ -\nabla_\theta \tilde{L}_{v^*(\theta)}(\nu^*(\theta), \theta, \lambda) \right] = 0$ , for a given  $\lambda \in [0, \lambda_{\max}]$ . Furthermore, if the function approximation error  $\epsilon_\theta(v_k)$  vanishes as the feature vector  $v_k$  converges to  $v^*$ , then the sequence of  $\theta$ -updates converges to  $\theta^*$  almost surely, where  $\theta^*$  is a local minimum point of  $L(\nu^*(\theta), \theta, \lambda)$  for a given  $\lambda \in [0, \lambda_{\max}]$ .*

*Proof.* We will mainly focus on proving the convergence of  $\theta_k \rightarrow \theta^*$  (second part of the theorem). Since we just showed in Proposition 7.3.4 that  $\nabla_{\theta} \tilde{L}_{v^*(\theta)}(\nu^*(\theta), \theta, \lambda)$  is Lipschitz in  $\theta$ , the convergence proof of  $\theta_k \rightarrow \hat{\theta}^*$  (first part of the theorem) follows from identical arguments.

Note that the  $\theta$ -update in (7.51) can be rewritten as:

$$\theta_{k+1} = \Gamma_{\Theta} \left( \theta_k + \zeta_2(k) \left( -\nabla_{\theta} L(\nu, \theta, \lambda) \big|_{\nu=\nu^*(\theta), \theta=\theta_k} + \delta\theta_{k+1} + \delta\theta_{\epsilon} \right) \right),$$

where

$$\begin{aligned} \delta\theta_{k+1} = & \sum_{x', a', s'} \pi_{\gamma}^{\theta_k}(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta_k)) \nabla_{\theta} \log \mu(a' | x', s'; \theta) \big|_{\theta=\theta_k} \frac{\tilde{A}^{\theta_k, v^*(\theta_k)}(x', s', a')}{1 - \gamma} \\ & - \nabla_{\theta} \log \mu(a_k | x_k, s_k; \theta) \big|_{\theta=\theta_k} \frac{\delta_k(v^*(\theta_k))}{1 - \gamma}. \end{aligned}$$

and

$$\begin{aligned} \delta\theta_{\epsilon} = & \sum_{x', a', s'} \pi_{\gamma}^{\theta_k}(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta_k)) \cdot \\ & \frac{\nabla_{\theta} \log \mu(a' | x', s'; \theta) \big|_{\theta=\theta_k} (A^{\theta_k}(x', s', a') - \tilde{A}^{\theta_k, v^*(\theta_k)}(x', s', a'))}{1 - \gamma} \end{aligned}$$

First, one can show that  $\delta\theta_{k+1}$  is square integrable, specifically,

$$\begin{aligned} & \mathbb{E}[\|\delta\theta_{k+1}\|^2 \mid \mathcal{F}_{\theta, k}] \\ & \leq \frac{2}{1 - \gamma} \|\nabla_{\theta} \log \mu(u | x, s; \theta) \big|_{\theta=\theta_k} \mathbf{1}_{\{\mu(u | x, s; \theta_k) > 0\}}\|_{\infty}^2 \left( \|\tilde{A}^{\theta_k, v^*(\theta_k)}(x, s, a)\|_{\infty}^2 + |\delta_k(v^*(\theta_k))|^2 \right) \\ & \leq \frac{2}{1 - \gamma} \cdot \frac{\|\nabla_{\theta} \log \mu(u | x, s; \theta) \big|_{\theta=\theta_k}\|_{\infty}^2}{\min\{\mu(u | x, s; \theta_k) \mid \mu(u | x, s; \theta_k) > 0\}^2} \left( \|\tilde{A}^{\theta_k, v^*(\theta_k)}(x, s, a)\|_{\infty}^2 + |\delta_k(v^*(\theta_k))|^2 \right) \\ & \leq 64 \frac{K^2 \|\theta_k\|^2}{1 - \gamma} \left( \max_{x, s, a} |\bar{C}_{\lambda}(x, s, a)|^2 + 2 \max_{x, s} \|\phi(x, s)\|^2 \sup_k \|v_k\|^2 \right) \\ & \leq 64 \frac{K^2 \|\theta_k\|^2}{1 - \gamma} \left( \left| \max \left\{ C_{\max}, \frac{2\lambda D_{\max}}{\gamma^T (1 - \alpha)(1 - \gamma)} \right\} \right|^2 + 2 \max_{x, s} \|\phi(x, s)\|^2 \sup_k \|v_k\|^2 \right), \end{aligned}$$

for some Lipschitz constant  $K$ , where the indicator function in the second line can be explained by the fact that  $\pi_{\gamma}^{\theta_k}(x, s, u) = 0$  whenever  $\mu(u | x, s; \theta_k) = 0$  and because the expectation is taken with respect to  $\pi_{\gamma}^{\theta_k}$ . The third inequality uses Assumption 3.2.3 and the fact that  $\mu$  takes on finitely-many values (and thus its nonzero values are bounded away from zero). Finally,  $\sup_k \|v_k\| < \infty$  follows from the Lyapunov analysis in the critic update.

Second, note that

$$\delta\theta_{\epsilon} \leq \frac{(1 + \gamma) \|\psi_{\theta_k}\|_{\infty}}{(1 - \gamma)^2} \epsilon_{\theta_k}(v^*(\theta_k)), \quad (7.52)$$

where  $\psi_\theta(x, s, a) = \nabla_\theta \log \mu(a|x, s; \theta)$  is the “compatible feature.” The last inequality is due to the fact that since  $\pi_\gamma^\theta$  is a probability measure, convexity of quadratic functions implies

$$\begin{aligned}
& \sum_{x', a', s'} \pi_\gamma^\theta(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta)) (A^\theta(x', s', a') - \tilde{A}^{\theta, v}(x', s', a')) \\
& \leq \sum_{x', a', s'} \pi_\gamma^\theta(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta)) (Q^\theta(x', s', a') - \tilde{Q}^{\theta, v}(x', s', a')) \\
& \quad + \sum_{x', s'} d_\gamma^\theta(x', s' | x_0 = x^0, s_0 = \nu^*(\theta)) (V^\theta(x', s') - \tilde{V}^{\theta, v}(x', s')) \\
& = \gamma \sum_{x', a', s'} \pi_\gamma^\theta(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta)) \sum_{x'', s''} \bar{P}(x'', s'' | x', s', a') (V^\theta(x'', s'') - \phi^\top(x'', s'')v) \\
& \quad + \sqrt{\sum_{x', s'} d_\gamma^\theta(x', s' | x_0 = x^0, s_0 = \nu^*(\theta)) (V^\theta(x', s') - \tilde{V}^{\theta, v}(x', s'))^2} \\
& \leq \gamma \sqrt{\sum_{x', a', s'} \pi_\gamma^\theta(x', s', a' | x_0 = x^0, s_0 = \nu^*(\theta)) \sum_{x'', s''} \bar{P}(x'', s'' | x', s', a') (V^\theta(x'', s'') - \phi^\top(x'', s'')v)^2} \\
& \quad + \frac{\epsilon_\theta(v)}{1 - \gamma} \\
& \leq \sqrt{\sum_{x'', s''} (d_\gamma^\theta(x'', s'' | x^0, \nu^*(\theta)) - (1 - \gamma)1\{x^0 = x'', \nu = s''\}) (V^\theta(x'', s'') - \phi^\top(x'', s'')v)^2} + \frac{\epsilon_\theta(v)}{1 - \gamma} \\
& \leq \left( \frac{1 + \gamma}{1 - \gamma} \right) \epsilon_\theta(v).
\end{aligned}$$

Then by Lemma 7.3.3, if the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$  is used to generate samples  $(x_k, s_k, a_k)$ , one obtains

$$\mathbb{E}[\delta\theta_{k+1} \mid \mathcal{F}_{\theta, k}] = 0,$$

where  $\mathcal{F}_{\theta, k} = \sigma(\theta_m, \delta\theta_m, m \leq k)$  is the filtration generated by different independent trajectories. On the other hand, we have that

$$|\delta\theta_\epsilon| \rightarrow 0 \text{ as } \epsilon_{\theta_k}(v^*(\theta_k)) \rightarrow 0.$$

Therefore, the  $\theta$ -update in (7.51) is a stochastic approximation of the continuous system  $\theta(t)$ , described by the ODE

$$\dot{\theta} = \Upsilon_\theta [-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}],$$

with an error term that is a sum of a vanishing bias and a Martingale difference. Thus, the convergence analysis of  $\theta$  follows analogously from Step 2 in the proof of Theorem 3.3.2, i.e., the sequence of  $\theta$ -updates in (3.21) converges to  $\theta^*$  almost surely, where  $\theta^*$  is the equilibrium point of the continuous system  $\theta$  satisfying

$$\Upsilon_\theta [-\nabla_\theta L(\nu, \theta, \lambda)|_{\nu=\nu^*(\theta)}] = 0. \quad (7.53)$$

■

**Step 4 (Local minimum)** The proof that  $(\theta^*, \nu^*)$  is a local minimum follows directly from the arguments in Step 3 in the proof of Theorem 3.3.2.

**Step 5 ( $\lambda$ -update and convergence to saddle point)** Note that the  $\lambda$ -update converges on the slowest time scale, thus, (3.20) may be rewritten using the converged  $v^*(\lambda)$ ,  $\theta^*(\lambda)$ , and  $\nu^*(\lambda)$  as

$$\lambda_{k+1} = \Gamma_\Lambda \left( \lambda_k + \zeta_1(k) \left( \nabla_\lambda L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda_k} + \delta\lambda_{k+1} \right) \right), \quad (7.54)$$

where

$$\delta\lambda_{k+1} = -\nabla_\lambda L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda), \lambda=\lambda_k} + \left( \nu^*(\lambda_k) + \frac{(-s_k)^+}{(1-\alpha)(1-\gamma)} \mathbf{1}\{x_k = x_{\text{Tar}}\} - \beta \right). \quad (7.55)$$

From (7.18),  $\nabla_\lambda L(\nu, \theta, \lambda)$  does not depend on  $\lambda$ . Similar to the  $\theta$ -update, one can easily show that  $\delta\lambda_{k+1}$  is square integrable, specifically,

$$\mathbb{E}[\|\delta\lambda_{k+1}\|^2 \mid \mathcal{F}_{\lambda,k}] \leq 8 \left( \beta^2 + \left( \frac{D_{\max}}{1-\gamma} \right)^2 + \left( \frac{2D_{\max}}{(1-\gamma)^2(1-\alpha)} \right)^2 \right),$$

where  $\mathcal{F}_{\lambda,k} = \sigma(\lambda_m, \delta\lambda_m, m \leq k)$  is the filtration of  $\lambda$  generated by different independent trajectories. Similar to the  $\theta$ -update, using the  $\gamma$ -occupation measure  $\pi_\gamma^\theta$ , one obtains  $\mathbb{E}[\delta\lambda_{k+1} \mid \mathcal{F}_{\lambda,k}] = 0$ . As above, the  $\lambda$ -update is a stochastic approximation for the continuous system  $\lambda(t)$  described by the ODE

$$\dot{\lambda} = \Upsilon_\lambda \left[ \nabla_\lambda L(\nu, \theta, \lambda) \Big|_{\theta=\theta^*(\lambda), \nu=\nu^*(\lambda)} \right],$$

with an error term that is a Martingale difference. Then the  $\lambda$ -convergence and the analysis of local optima follow from analogous arguments in Steps 4 and 5 in the proof of Theorem 3.3.2.

## 7.4 Technical Results in Chapter 4

In this section we present the detailed proofs to the technical results in Chapter 4.

### 7.4.1 Proof of Lemma 4.5.1

From the time consistency, monotonicity, translational invariance, and positive homogeneity of Markov dynamic polytopic risk measures, condition (4.8) implies

$$\begin{aligned}\rho_{0,k+1}(0, \dots, 0, b_1 \|x_{k+1}\|^2) &\leq \rho_{0,k+1}(0, \dots, 0, V(x_{k+1})) = \rho_{0,k}(0, \dots, 0, V(x_k) + \rho(V(x_{k+1}) - V(x_k))) \\ &\leq \rho_{0,k}(0, \dots, 0, V(x_k) - b_3 \|x_k\|^2) \leq \rho_{0,k}(0, \dots, 0, (b_2 - b_3) \|x_k\|^2).\end{aligned}$$

Also, since  $\rho_{0,k+1}$  is monotonic, one has  $b_1 \rho_{0,k+1}(0, \dots, 0, \|x_{k+1}\|^2) \geq 0$ , which implies  $b_2 \geq b_3$  and in turn  $(1 - b_3/b_2) \in [0, 1)$ . Since  $V(x_k)/b_2 \leq \|x_k\|^2$ , by using the previous inequalities one can write:

$$\rho_{0,k+1}(0, \dots, 0, V(x_{k+1})) \leq \rho_{0,k}(0, \dots, 0, V(x_k) - b_3 \|x_k\|^2) \leq \left(1 - \frac{b_3}{b_2}\right) \rho_{0,k}(0, \dots, 0, V(x_k)).$$

Repeating this bounding process, one obtains:

$$\begin{aligned}\rho_{0,k+1}(0, \dots, 0, V(x_{k+1})) &\leq \left(1 - \frac{b_3}{b_2}\right)^k \rho_{0,1}(V(x_1)) \\ &= \left(1 - \frac{b_3}{b_2}\right)^k \rho(V(x_1)) \leq \left(1 - \frac{b_3}{b_2}\right)^k (V(x_0) - b_3 \|x_0\|^2) \leq b_2 \left(1 - \frac{b_3}{b_2}\right)^{k+1} \|x_0\|^2.\end{aligned}$$

Again, by monotonicity, the above result implies

$$\rho_{0,k+1}(0, \dots, 0, x_{k+1}^\top x_{k+1}) \leq \frac{b_2}{b_1} \left(1 - \frac{b_3}{b_2}\right)^{k+1} x_0^\top x_0.$$

By setting  $c = b_2/b_1$  and  $\lambda = (1 - b_3/b_2) \in [0, 1)$ , the claim is proven.

### 7.4.2 Proof of Theorem 4.6.1

The strategy of this proof is to show that  $J_k^*$  is a valid Lyapunov function in the sense of Lemma 4.5.1. Specifically, we want to show that  $J_k^*$  satisfies the two inequalities in equation (4.8); the claim then follows by simply noting that, in our time-invariant setup,  $J_k^*$  does not depend on  $k$ .

We start by focusing on the bottom inequality in equation (4.8). Consider a time  $k$  and an initial condition  $x_{k|k} \in \mathbb{R}^{N_x}$  for problem  $\mathcal{MPC}$ . The sequence of optimal control policies is given by  $\{\pi_{k+h|k}^*\}_{h=0}^{N-1}$ . Let us



define a sequence of control policies from time  $k + 1$  to  $N$  according to

$$\pi_{k+h|k+1}(x_{k+h|k}) := \begin{cases} \pi_{k+h|k}^*(x_{k+h|k}) & \text{if } h \in [1, N-1], \\ F x_{k+N|k} & \text{if } h = N. \end{cases} \quad (7.56)$$

This sequence of control policies is essentially the concatenation of the sequence  $\{\pi_{k+h|k}^*\}_{h=1}^{N-1}$  with a linear feedback control law for stage  $N$  (the reason why we refer to this policy with the subscript “ $k + h|k + 1$ ” is that we will use this policy as a feasible policy for problem  $\mathcal{MPC}$  starting at stage  $k + 1$ ).

Consider the  $\mathcal{MPC}$  problem at stage  $k + 1$  with initial condition given by  $x_{k+1|k+1} = A(w_k)x_{k|k} + B(w_k)\pi_{k|k}^*(x_{k|k})$ , and denote with  $\bar{J}_{k+1}(x_{k+1|k+1})$  the cost of the objective function for the  $\mathcal{MPC}$  problem assuming that the sequence of control policies is given by  $\{\pi_{k+h|k+1}\}_{h=1}^N$ . Note that  $x_{k+1|k+1}$  (and therefore  $\bar{J}_{k+1}(x_{k+1|k+1})$ ) is a random variable with  $L$  realizations, given  $x_{k|k}$ . Define

$$\begin{aligned} Z_{k+N} &:= -x_{k+N|k}^\top P x_{k+N|k} + x_{k+N|k}^\top Q x_{k+N|k} + (F x_{k+N|k})^\top R F x_{k+N|k} \\ Z_{k+N+1} &:= ((A(w_{k+N|k}) + B(w_{k+N|k})F)x_{k+N|k})^\top P ((A(w_{k+N|k}) + B(w_{k+N|k})F)x_{k+N|k}). \end{aligned}$$

By exploiting the dual representation of Markov polytopic risk metrics, one can write

$$\begin{aligned} Z_{k+N} + \rho_{k+N}(Z_{k+N+1}) &= x_{k+N|k}^\top (-P + Q + F^\top R F) x_{k+N|k} + \\ &\quad \max_{q \in \mathcal{U}^{\text{poly}}(p)} \sum_{j=1}^L q(j) x_{k+N|k}^\top (A_j + B_j F)^\top P (A_j + B_j F) x_{k+N|k}. \end{aligned}$$

Combining the equation above with equation (4.11), one readily obtains the inequality

$$Z_{k+N} + \rho_{k+N}(Z_{k+N+1}) \leq 0. \quad (7.57)$$

One can then write the following chain of inequalities:

$$\begin{aligned} J_k^*(x_{k|k}) &= x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho_k \left( \rho_{k+1,N} \left( c(x_{k+1|k}, \pi_{k+1}^*(x_{k+1|k})), \dots, x_{k+N|k}^\top Q x_{k+N|k} + \right. \right. \\ &\quad \left. \left. (F x_{k+N|k})^\top R F x_{k+N|k} + \rho_{k+N}(Z_{k+N+1}) - Z_{k+N} - \rho_{k+N}(Z_{k+N+1}) \right) \right) \\ &\geq x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho_k \left( \rho_{k+1,N} \left( c(x_{k+1|k}, \pi_{k+1}^*(x_{k+1|k})), \dots, \right. \right. \\ &\quad \left. \left. x_{k+N|k}^\top Q x_{k+N|k} + (F x_{k+N|k})^\top R F x_{k+N|k} + \rho_{k+N}(Z_{k+N+1}) \right) \right) \\ &= x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho_k \left( \bar{J}_{k+1}(x_{k+1|k+1}) \right) \\ &\geq x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho_k \left( J_{k+1}^*(x_{k+1|k+1}) \right), \end{aligned} \quad (7.58)$$

where the first equality follows from the definitions of  $Z_{k+N}$  and of dynamic, time-consistent risk measures, the second inequality follows from equation (7.57) and the monotonicity property of Markov polytopic risk metrics (see also [119, Page 242]), the third equality follows from the fact that the sequence of control policies  $\{\pi_{k+h|k+1}\}_{h=1}^N$  is a feasible sequence for the  $\mathcal{MPC}$  problem starting at stage  $k+1$  with initial condition  $x_{k+1|k+1} = A(w_k)x_{k|k} + B(w_k)\pi_{k|k}^*(x_{k|k})$ , and the fourth inequality follows from the definition of  $J_{k+1}^*$  and the monotonicity of Markov polytopic risk metrics.

We now turn our focus to the top inequality in equation (4.8). One can easily bound  $J_k^*(x_{k|k})$  from below according to:

$$J_k^*(x_{k|k}) \geq x_{k|k}^\top Q x_{k|k} \geq \lambda_{\min}(Q) \|x_{k|k}\|^2, \quad (7.59)$$

where  $\lambda_{\min}(Q) > 0$  by assumption. To bound  $J_k^*(x_{k|k})$  from above, define:

$$M_A := \max_{r \in \{0, \dots, N-1\}} \max_{j_0, \dots, j_r \in \{1, \dots, L\}} \|A_{j_r} \dots A_{j_1} A_{j_0}\|_2.$$

Since the problem is unconstrained (and, hence, zero is a feasible control input) and by exploiting the monotonicity property, one can write:

$$\begin{aligned} J_k^*(x_{k|k}) &\leq c(x_{k|k}, 0) + \rho_k \left( C(x_{k+1|k}, 0) + \rho_{k+1} \left( C(x_{k+2|k}, 0) + \dots + \rho_{k+N-1} \left( x_{k+N|k}^\top P x_{k+N|k} \right) \dots \right) \right) \\ &\leq \|Q\|_2 \|x_{k|k}\|_2^2 + \rho_k \left( \|Q\|_2 \|x_{k+1|k}\|_2^2 + \rho_{k+1} \left( \|Q\|_2 \|x_{k+2|k}\|_2^2 + \dots + \rho_{k+N-1} (\|P\|_2 \|x_{k+N|k}\|_2^2) \dots \right) \right). \end{aligned}$$

Therefore, by using the translational invariance and monotonicity property of Markov polytopic risk measures, one obtains the upper bound:

$$J_k^*(x_{k|k}) \leq (N \|Q\|_2 + \|P\|_2) M_A \|x_{k|k}\|_2^2. \quad (7.60)$$

Combining the results in equations (7.58), (7.59), (7.60), and given the time-invariance of our problem setup, one concludes that  $J_k^*(x_{k|k})$  is a “risk-sensitive” Lyapunov function for the closed-loop system (4.1), in the sense of Lemma 4.5.1. This concludes the proof.

### 7.4.3 Proof of Lemma 4.6.2

We first prove statements 1) and 2) and thereby establish  $a(x) = Fx$  as a feasible control law within the set  $\mathcal{E}_{\max}(W)$ . Notice that the condition  $\|T_a Fx\|_2 \leq a_{\max}$  holds if and only if:

$$\|T_a F W^{\frac{1}{2}} (W^{-\frac{1}{2}} x)\|_2 \leq a_{\max}.$$

From (4.12), and by applying the Schur complement, we know that  $\|W^{-\frac{1}{2}}x\|_2 \leq 1$  for any  $x \in \mathcal{E}_{\max}(W)$ . Thus, by the Cauchy Schwarz inequality, a sufficient condition for the control constraint is given by

$$\|T_a F W^{\frac{1}{2}}\|_2 \leq a_{\max},$$

which can be written as

$$(F W^{\frac{1}{2}})^\top T_a^\top T_a (F W^{\frac{1}{2}}) \preceq a_{\max}^2 I \iff F^\top T_a^\top T_a F \preceq a_{\max}^2 W^{-1}.$$

Re-arranging the inequality above yields the expression given in (7.70). The state constraint can be proved in an identical fashion by leveraging conditions (4.12) and (7.71). It is omitted for brevity.

We now prove the third statement. By definition of a robust control invariant set, we are required to show that for any  $x \in \mathcal{E}_{\max}(W)$ , that is, for all  $x$  that satisfy the inequality:  $x^\top W^{-1}x \leq 1$ , application of the control law  $a(x) = Fx$  yields the following inequality:

$$(A_j x + B_j Fx)^\top W^{-1} (A_j x + B_j Fx) \leq 1, \quad \forall j \in \{1, \dots, L\}.$$

Using the S-procedure [161], it is equivalent to show that there exists  $\lambda \geq 0$  such that the following condition holds:

$$\begin{bmatrix} \lambda W^{-1} - (A_j + B_j F)^\top W^{-1} (A_j + B_j F) & 0 \\ * & 1 - \lambda \end{bmatrix} \succeq 0, \quad \forall j \in \{1, \dots, L\}.$$

By setting  $\lambda = 1$ , one obtains the largest feasibility set for  $W$  and  $F$ . The expression in (7.72) corresponds to the (1,1) block in the matrix above.

#### 7.4.4 Proof of Theorem 4.6.4

Given  $x_{k|k} \in \mathcal{X}_N$ , problem  $\mathcal{MPC}$  may be solved to yield a closed-loop optimal control policy:

$$\{\pi_{k|k}^*(x_{k|k}), \dots, \pi_{k+N-1|k}^*(x_{k+N-1|k})\},$$

such that  $x_{k+N|k} \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$ . Consider problem  $\mathcal{MPC}$  at stage  $k+1$  with initial condition  $x_{k+1|k+1}$ . From Lemma 4.6.2, we know that the set  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$  is robust control invariant under the feasible feedback control law  $a(x) = Fx$ . Thus,

$$\{\pi_{k+1|k}^*(x_{k+1|k}), \dots, \pi_{k+N-1|k}^*(x_{k+N-1|k}), Fx_{k+N|k}\}, \quad (7.61)$$

is a feasible control policy at stage  $k+1$ . Note that this is simply a concatenation of the optimal tail policy from the previous iteration  $\{\pi_{k+h|k}^*(x_{k+h|k})\}_{h=1}^{N-1}$ , with the state feedback law  $Fx_{k+N|k}$  for the final step.

Since a feasible control policy exists at stage  $k+1$ ,  $x_{k+1|k+1} = A_j x_{k|k} + B_j \pi_{k|k}^*(x_{k|k}) \in \mathcal{X}_N$  for any

$j \in \{1, \dots, L\}$ , completing the proof.

### 7.4.5 Proof of Theorem 4.6.6

The first part of the proof is identical to the reasoning presented in the proof for Theorem 4.6.1. In particular, we leverage the policy given in (7.61) as a feasible policy for problem  $\mathcal{MPC}$  at stage  $k + 1$  and inequality (7.73) to show:

$$J_k^*(x_{k|k}) \geq C(x_{k|k}, \pi_{k|k}^*(x_{k|k})) + \rho_k(J_{k+1}^*(x_{k+1|k+1})), \quad (7.62)$$

for all  $x_{k|k} \in \mathcal{X}_N$ . Additionally, we retain the same lower bound for  $J_k^*(x_{k|k})$  as given in (7.59). The upper bound for  $J_k^*(x_{k|k})$  is derived in two steps. First, define

$$M_A := \max_{r \in \{0, \dots, N-1\}} \max_{j_0, \dots, j_r \in \{1, \dots, L\}} \alpha_{j_r} \dots \alpha_{j_1} \alpha_{j_0},$$

$$\text{where } \alpha_j := \|A_j + B_j F\|_2, \theta_f := \|Q + F^\top R F\|_2.$$

Suppose  $x_{k|k} \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$ . From Lemma 4.6.2, we know that the control policy  $\pi_{k+h|k}(x_{k+h|k}) = \{F x_{k+h|k}\}_{h=0}^{N-1}$  is feasible and consequently,  $\mathbb{X} \cap \mathcal{E}_{\max}(W) \subseteq \mathcal{X}_N$ . Thus,

$$\begin{aligned} J_k^*(x_{k|k}) &\leq C(x_{k|k}, F x_{k|k}) + \rho_k \left( C(x_{k+1|k}, F x_{k+1|k}) + \dots + \rho_{k+N-1} \left( x_{k+N|k}^\top P x_{k+N|k} \right) \dots \right) \\ &\leq \theta_f \|x_{k|k}\|_2^2 + \rho_k \left( \theta_f \|x_{k+1|k}\|_2^2 + \dots + \rho_{k+N-1} (\|P\|_2 \|x_{k+N|k}\|_2^2) \dots \right), \end{aligned}$$

for all  $x_{k|k} \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$ . Exploiting the translational invariance and monotonicity property of Markov polytopic risk metrics, one obtains the upper bound:

$$J_k^*(x_{k|k}) \leq \underbrace{(N \theta_f + \|P\|_2) M_A}_{:= \beta > 0} \|x_{k|k}\|_2^2, \quad \forall x_{k|k} \in \mathbb{X} \cap \mathcal{E}_{\max}(W). \quad (7.63)$$

In order to derive an upper bound for  $J_k^*(x_{k|k})$  with the above structure for all  $x_{k|k} \in \mathcal{X}_N$ , we draw inspiration from a similar proof in [110]. Notice that there exists some constant  $\Gamma > 0$  such that  $J_k^*(x_{k|k}) \leq \Gamma$  for all  $x_{k|k} \in \mathcal{X}_N$ . That  $\Gamma$  is finite follows from the fact that  $\{\|x_{k+h|k}\|_2\}_{h=0}^N$  and  $\{\|\pi_{k+h|k}(x_{k+h|k})\|_2\}_{h=0}^{N-1}$  are finitely bounded for all  $x_{k|k} \in \mathcal{X}_N$ . Now since  $\mathcal{E}_{\max}(W)$  is compact and non-empty, there exists a  $d > 0$  such that  $\mathcal{E}_d := \{x \in \mathbb{R}^{N_x} \mid \|x\|_2 \leq d\} \subset \mathcal{E}_{\max}(W)$ . Let  $\hat{\beta} = \max\{\beta \|x\|_2^2 \mid \|x\|_2 \leq d\}$ . Consider now, the function:  $(\Gamma/\hat{\beta})\beta \|x\|_2^2$ . Then since  $\beta \|x\|_2^2 > \hat{\beta}$  for all  $x \in \mathcal{X}_N \setminus \mathcal{E}_d$  and  $\Gamma \geq \hat{\beta}$ , it follows that

$$J_k^*(x_{k|k}) \leq \left( \frac{\Gamma \beta}{\hat{\beta}} \right) \|x_{k|k}\|_2^2, \quad \forall x_{k|k} \in \mathcal{X}_N, \quad (7.64)$$

as desired. Combining the results in equations (7.62), (7.59), (7.64), and given the time-invariance of our problem setup, one concludes that  $J_k^*(x_{k|k})$  is a “risk-sensitive” Lyapunov function for the closed-loop system (4.1), in the sense of Lemma 4.5.1. This concludes the proof.

### 7.4.6 Proof of Theorem 4.7.1

Consider a symmetric matrix  $X$  such that the LMI in equation (4.19) is satisfied. Also, let  $\pi$  be a stationary feedback control policy that is feasible for problem  $\mathcal{OPT}_{RS}$ . At stage  $k$ , consider a state  $x_k$  (reachable under policy  $\pi$ ) and the corresponding control action  $a_k = \pi(x_k)$  (since  $\pi$  is a feasible policy, the pair  $(x_k, a_k)$  clearly satisfies the state-control constraints). By pre- and post-multiplying the LMI (4.19) with  $[a_k^\top, x_k^\top]$  and its transpose, one obtains the inequality

$$a_k^\top R a_k + a_k^\top \bar{B}^\top (\Sigma_l \otimes X) \bar{B} a_k + 2a_k^\top \bar{B}^\top (\Sigma_l \otimes X) \bar{A} x_k + x_k^\top (\bar{A}^\top (\Sigma_l \otimes X) \bar{A} - (X - Q)) x_k \geq 0,$$

which implies

$$x_k^\top X x_k - \sum_{j=1}^L q_l(j) (A_j x_k + B_j a_k)^\top X (A_j x_k + B_j a_k) \leq a_k^\top R a_k + x_k^\top Q x_k,$$

for all  $l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}$ . Since  $\mathcal{U}^{\text{poly}}(p)$  is a convex polytope of probability vectors  $q$  (with vertex set  $\mathcal{U}^{\text{poly}, V}(p)$ ), then for any  $q \in \mathcal{U}^{\text{poly}}(p)$  one has the inequality

$$x_k^\top X x_k - \sum_{j=1}^L q(j) (A_j x_k + B_j a_k)^\top X (A_j x_k + B_j a_k) \leq a_k^\top R a_k + x_k^\top Q x_k.$$

Exploiting the dual representation of Markov polytopical risk measures, one has

$$\rho_k(x_{k+1}^\top X x_{k+1}) = \max_{q \in \mathcal{U}^{\text{poly}}(p)} E_q[x_{k+1}^\top X x_{k+1}],$$

which leads to the inequality

$$x_k^\top X x_k - \rho_k(x_{k+1}^\top X x_{k+1}) \leq a_k^\top R a_k + x_k^\top Q x_k.$$

As the above inequality holds for all  $k \in \mathbb{N}$ , one can write, for all  $k \in \mathbb{N}$ ,

$$\sum_{h=0}^k (x_h^\top X x_h - \rho_h(x_{h+1}^\top X x_{h+1})) \leq \sum_{h=0}^k (u_h^\top R u_h + x_h^\top Q x_h).$$

Since each single-period risk measure is monotone, their composition  $\rho_0 \circ \dots \circ \rho_{k-1}$  is monotone as well.

Hence by applying  $\rho_0 \circ \dots \circ \rho_{k-1}$  to both sides one obtains

$$\rho_0 \circ \dots \circ \rho_{k-1} \left( \sum_{h=0}^k (x_h^\top X x_h - \rho_h(x_{h+1}^\top X x_{h+1})) \right) \leq \rho_0 \circ \dots \circ \rho_{k-1} \left( \sum_{h=0}^k (u_h^\top R u_h + x_h^\top Q x_h) \right).$$

By repeatedly applying the translational invariance property (see Definition 1.3.7), the right-hand side can be written as

$$\begin{aligned} & \rho_0 \circ \dots \circ \rho_{k-1} \left( \sum_{h=0}^k u_h^\top R u_h + x_h^\top Q x_h \right) \\ &= a_0^\top R a_0 + x_0^\top Q x_0 + \rho_0(u_1^\top R u_1 + x_1^\top Q x_1 + \dots + \rho_{k-1}(a_k^\top R a_k + x_k^\top Q x_k) \dots) \\ &= \rho_{0,k} (a_0^\top R a_0 + x_0^\top Q x_0, \dots, a_k^\top R a_k + x_k^\top Q x_k) = J_{0,k}(x_0, \pi), \end{aligned}$$

where the last equality follows from the definition of dynamic, time-consistent risk measures (Theorem 1.3.8). As for the left-hand side, note that the translation invariance and positive homogeneity property imply that a coherent one-step conditional risk measure is subadditive, i.e.,  $\rho_h(Z + W) \leq \rho_h(Z) + \rho_h(W)$ , where  $Z, W \in \mathcal{Z}_{h+1}$ . In turn, subadditivity implies that  $\rho_h(Z - W) \geq \rho_h(Z) - \rho_h(W)$ . Hence, by repeatedly applying the translation invariance and monotonicity property and the inequality  $\rho_h(Z - W) \geq \rho_h(Z) - \rho_h(W)$ , one obtains

$$\begin{aligned} & \rho_0 \circ \dots \circ \rho_{k-1} \left( \sum_{h=0}^k (x_h^\top X x_h - \rho_h(x_{h+1}^\top X x_{h+1})) \right) = x_0^\top X x_0 - \rho_0(x_1^\top X x_1) + \\ & \quad \rho_0(x_1^\top X x_1 - \rho_1(x_2^\top X x_2)) + \rho_1(x_2^\top X x_2 - \rho_2(x_3^\top X x_3)) + \dots + \rho_{k-1}(x_k^\top X x_k - \rho_k(x_{k+1}^\top X x_{k+1})) \geq \\ & \quad x_0^\top X x_0 - \rho_0 \circ \dots \circ \rho_{k-1} \circ \rho_k(x_{k+1}^\top X x_{k+1}) = x_0^\top X x_0 - \rho_{0,k+1}(0, \dots, x_{k+1}^\top X x_{k+1}). \end{aligned}$$

Since  $\pi$  is a feasible policy,  $\lim_{k \rightarrow \infty} \rho_{0,k}(0, \dots, x_{k+1}^\top X x_{k+1}) = 0$  almost surely. Hence, one readily obtains (using the monotonicity and positive homogeneity property)

$$\lim_{k \rightarrow \infty} \rho_{0,k+1}(0, \dots, 0, x_{k+1}^\top X x_{k+1}) \leq \lambda_{\max}(X) \lim_{k \rightarrow \infty} \rho_{0,k+1}(0, \dots, x_{k+1}^\top X x_{k+1}) = 0,$$

almost surely. Collecting all results so far, one has the inequality

$$x_0^\top X x_0 \leq \lim_{k \rightarrow \infty} J_{0,k}(x_0, \pi),$$

for all symmetric matrices satisfying the LMI (4.19) and all feasible policies  $\pi$ . By maximizing the left-hand side and minimizing the right-hand side one obtains the claim.

#### 7.4.7 Proof of Theorem 4.7.2

For all  $k \in \mathbb{N}$ , inequality (7.62) in the proof of Theorem 4.6.6 provides the relation

$$J_k^*(x_{k|k}) \geq C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \rho_k(J_{k+1}^*(x_{k+1|k+1})).$$

Since  $x_0 \in \mathcal{X}_N$  and problem  $\mathcal{MPC}$  is recursively feasible, we obtain the following sequence of state-control pairs:  $\{(x_{k|k}, \pi^{MPC}(x_{k|k}))\}_{k=0}^\infty$ . Applying inequality (7.62) recursively and using the monotonicity property of coherent one-step risk measures, we deduce the following:

$$\begin{aligned}
& J_0^*(x_{0|0}) \\
& \geq C(x_{0|0}, \pi^{MPC}(x_{0|0})) + \rho_0(J_1^*(x_{1|1})) \\
& \geq C(x_{0|0}, \pi^{MPC}(x_{0|0})) + \rho_0(C(x_{1|1}, \pi^{MPC}(x_{1|1})) + \rho_1(J_2^*(x_{2|2}))) \\
& \geq \dots \geq C(x_{0|0}, \pi^{MPC}(x_{0|0})) + \rho_0(C(x_{1|1}, \pi^{MPC}(x_{1|1})) + \dots + \rho_{k-1}(C(x_{k|k}, \pi^{MPC}(x_{k|k})) + J_{k+1}^*(x_{k+1|k+1}))) \dots \\
& \geq C(x_{0|0}, \pi^{MPC}(x_{0|0})) + \rho_0(C(x_{1|1}, \pi^{MPC}(x_{1|1})) + \dots + \rho_{k-1}(C(x_{k|k}, \pi^{MPC}(x_{k|k}))) \dots) \\
& = \rho_{0,k}(C(x_{0|0}, \pi^{MPC}(x_{0|0})), \dots, C(x_{k|k}, \pi^{MPC}(x_{k|k}))) \quad \forall k \in \mathbb{N},
\end{aligned}$$

where the second to last inequality follows from the fact that  $J_{k+1}^*(x_{k+1|k+1}) \geq 0$ , and the equality follows from the definition of dynamic, time-consistent risk metrics. Noting that  $x_{k|k} = x_k$  for all  $k \in \mathbb{N}$  and by taking the limit  $k \rightarrow \infty$  on both sides, one obtains the claim.

#### 7.4.8 Proof of Corollary 4.8.4

From Theorem 4.8.1, we know that the set of LMIs in (4.20) is equivalent to the expression in (7.73) when  $F = YG^{-1}$ . Then since  $x_0 \in \mathbb{X} \cap \mathcal{E}_{\max}(W)$ , a robust control invariant set under the local feedback control law  $a(x) = YG^{-1}x$ , exploiting the dual representation of Markov polytopic risk measures yields the inequality

$$\rho_k(x_{k+1}^\top P x_{k+1}) - x_k^\top P x_k \leq -x_k^\top L x_k \quad \forall k \in \mathbb{N}, \quad (7.65)$$

where  $L = Q + (YG^{-1})^\top R (YG^{-1}) = L^\top \succ 0$ . Define the Lyapunov function  $V(x) = x^\top P x$ . Set  $b_1 = \lambda_{\min}(P) > 0$ ,  $b_2 = \lambda_{\max}(P) > 0$  and  $b_3 = \lambda_{\min}(L) > 0$ . Then by Lemma 4.5.1, this stochastic system is ULRSSES with domain  $\mathbb{X} \cap \mathcal{E}_{\max}(W)$ .

#### 7.4.9 Proof of Theorem 4.8.1 and Corollary 4.8.2

We first present the Projection Lemma:

**Lemma 7.4.1** (Projection Lemma). *For matrices  $\Omega(X)$ ,  $U(x)$ ,  $V(X)$  of appropriate dimensions, where  $X$  is a matrix variable, the following statements are equivalent:*

1. *There exists a matrix  $W$  such that*

$$\Omega(X) + U(x)WV(X) + V(X)^\top W^\top U(x)^\top \prec 0.$$

2. The following inequalities hold:

$$U(x)^\perp \Omega(X)(U(x)^\perp)^\top \prec 0, \quad (V(X)^\top)^\perp \Omega(X)((V(X)^\top)^\perp)^\top \prec 0,$$

where  $A^\perp$  is the orthogonal complement of  $A$ .

*Proof.* See Chapter 2 in [135]. □

We now give the proof for Theorem 4.8.1 by leveraging the Projection lemma:

*Proof.* (Proof of Theorem 4.8.1) Using simple algebraic factorizations,  $\forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}$ , inequality (7.73) can be rewritten as

$$\begin{bmatrix} I \\ \Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ F \\ Q^{\frac{1}{2}} \end{bmatrix}^\top \begin{bmatrix} P & 0 & 0 & 0 \\ 0 & -I_{L \times L} \otimes P & 0 & 0 \\ 0 & 0 & -R & 0 \\ 0 & 0 & 0 & -I \end{bmatrix} \begin{bmatrix} I \\ \Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ F \\ Q^{\frac{1}{2}} \end{bmatrix} \succ 0.$$

By Schur complement, the above expression is equivalent to

$$\begin{bmatrix} I & 0 & 0 & \Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ 0 & I & 0 & F \\ 0 & 0 & I & Q^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} I_{L \times L} \otimes \bar{Q} & 0 & 0 & 0 \\ 0 & R^{-1} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & -\bar{Q} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ (\bar{A} + \bar{B}F)^\top \Sigma_l^{\frac{1}{2}} & F^\top & Q^{\frac{1}{2}} \end{bmatrix} \succ 0,$$

where  $\bar{Q} = P^{-1}$ . Now since  $\bar{Q} = \bar{Q}^\top \succ 0$  and  $R = R^\top \succ 0$ , we also have the following identity:

$$\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \end{bmatrix} \begin{bmatrix} I_{L \times L} \otimes \bar{Q} & 0 & 0 & 0 \\ 0 & R^{-1} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & -\bar{Q} \end{bmatrix} \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{bmatrix} \succ 0.$$

Next, notice that

$$\begin{bmatrix} -\Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ -F \\ -Q^{\frac{1}{2}} \\ I \end{bmatrix}^\perp = \begin{bmatrix} I & 0 & 0 & \Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ 0 & I & 0 & F \\ 0 & 0 & I & Q^{\frac{1}{2}} \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix}^\perp = \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \end{bmatrix}.$$



Now, set:

$$\Omega = - \begin{bmatrix} I_{L \times L} \otimes \bar{Q} & 0 & 0 & 0 \\ 0 & R^{-1} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & -\bar{Q} \end{bmatrix}, U = \begin{bmatrix} -\Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ -F \\ -Q^{\frac{1}{2}} \\ I \end{bmatrix}, V^T = \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix}.$$

Then by Lemma 7.4.1, it is equivalent to find a matrix  $G$  that satisfies the following inequality  $\forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V})\}$

$$\begin{bmatrix} I_{L \times L} \otimes \bar{Q} & 0 & 0 & 0 \\ 0 & R^{-1} & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & -\bar{Q} \end{bmatrix} + \begin{bmatrix} -\Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ -F \\ -Q^{\frac{1}{2}} \\ I \end{bmatrix} G \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix}^T + \begin{bmatrix} 0 \\ 0 \\ 0 \\ I \end{bmatrix} G^T \begin{bmatrix} -\Sigma_l^{\frac{1}{2}}(\bar{A} + \bar{B}F) \\ -F \\ -Q^{\frac{1}{2}} \\ I \end{bmatrix}^T \succ 0. \quad (7.66)$$

Setting  $F = YG^{-1}$  and pre-and post-multiplying the above inequality by  $\text{diag}(I, R^{\frac{1}{2}}, I, I)$  yields the LMI given in (4.20). Furthermore, from the inequality  $-\bar{Q} + G + G^T \succ 0$  where  $\bar{Q} \succ 0$ , we know that  $G + G^T \succ 0$ . Thus, by the Lyapunov stability theorem, the linear time-invariant system  $\dot{x} = -Gx$  with Lyapunov function  $x^T x$  is asymptotically stable (i.e. all eigenvalues of  $G$  have positive real part). Therefore,  $G$  is an invertible matrix and  $F = YG^{-1}$  is well defined.  $\square$

*Proof.* (Proof of Corollary 4.8.2) We will prove that the third inequality in (4.21) implies inequality (7.72). Details of the proofs on the implications of the first two inequalities in (4.21) follow from identical arguments and will be omitted for the sake of brevity. Using simple algebraic factorizations, inequality (7.72) may be rewritten (in strict form) as:

$$\begin{bmatrix} I \\ A_j + B_j F \end{bmatrix}^T \begin{bmatrix} W^{-1} & 0 \\ 0 & -W^{-1} \end{bmatrix} \begin{bmatrix} I \\ A_j + B_j F \end{bmatrix} \succ 0, \forall j \in \{1, \dots, L\}.$$

By Schur complement, the above expression is equivalent to

$$\begin{bmatrix} I & A_j + B_j F \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & -W \end{bmatrix} \begin{bmatrix} I \\ (A_j + B_j F)^T \end{bmatrix} \succ 0, \forall j \in \{1, \dots, L\}.$$

Furthermore since  $W \succ 0$ , we also have the identity

$$\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} W & 0 \\ 0 & -W \end{bmatrix} \begin{bmatrix} I \\ 0 \end{bmatrix} \succ 0.$$

Now, notice that:

$$\begin{bmatrix} -(A_j + B_j F) \\ I \end{bmatrix}^\perp = \begin{bmatrix} I & A_j + B_j F \end{bmatrix}, \quad \begin{bmatrix} 0 \\ I \end{bmatrix}^\perp = \begin{bmatrix} I & 0 \end{bmatrix}.$$

Then by Lemma 7.4.1, it is equivalent to find a matrix  $G$  such that the following inequality holds  $\forall j \in \{1, \dots, L\}$ :

$$\begin{bmatrix} W & 0 \\ 0 & -W \end{bmatrix} + \begin{bmatrix} -(A_j + B_j F) \\ I \end{bmatrix} G \begin{bmatrix} 0 \\ I \end{bmatrix}^\top + \begin{bmatrix} 0 \\ I \end{bmatrix} G^\top \begin{bmatrix} -(A_j + B_j F) \\ I \end{bmatrix}^\top \succ 0. \quad (7.67)$$

Note that Lemma 7.4.1 provides an equivalence (necessary and sufficient) condition between (7.67) and (7.72) if  $G$  is allowed to be any arbitrary LMI variable. However, in order to restrict  $G$  to be the same variable as in Theorem 4.8.1, the equivalence relation reduces to sufficiency only. Setting  $F = YG^{-1}$  in the above expression gives the claim.  $\square$

#### 7.4.10 Convex Programming Formulation of Problem $\mathcal{MPC}$

Next, we provide a technical Lemma that transforms the multi-period risk sensitive objective function of Problem  $\mathcal{MPC}$  using its epigraph form. The major reasons behind this transformation is to obtain a tractable formulation via convex programming.

**Lemma 7.4.2.** *The solution of Problem  $\mathcal{MPC}$  equals to the solution of following optimization problem:*

$$\begin{aligned} \min \quad & \gamma_1 \\ \text{subject to } & \gamma_1, W = W^\top \succ 0, G, Y, \bar{Q} = \bar{Q}^\top \succ 0, \gamma_2(j_0, \dots, j_{N-1}), \bar{a}_0, \bar{a}_h(j_0, \dots, j_{h-1}), \\ & \bar{x}_h(j_0, \dots, j_{h-1}), h \in \{1, \dots, N\}, j_0, \dots, j_{N-1} \in \{1, \dots, L\} \end{aligned} \quad (7.68)$$

subjected to the following constraints:

- the LMIs in expression (4.24) and (4.25);
- the system dynamics in equation (4.22);
- the control constraints in expression (4.26);
- the state constraints in expression (4.27);
- the objective epigraph constraint:

$$\rho_{k,k+N}(c(x_{k|k}, \bar{a}_0), \dots, c(\bar{x}_{N-1}(j_0, \dots, j_{N-2}), \bar{a}_{N-1}(j_0, \dots, j_{N-2})), \gamma_2(j_0, \dots, j_{N-1})) \leq \gamma_1.$$

*Proof.* First, let  $\gamma_1^*$  be the optimal value for the above problem in expression (7.68), corresponding to the minimizers:  $\gamma_2^*(j_0, \dots, j_{N-1})$ ,  $\bar{a}_0^*$ ,  $W^*$ ,  $G^*$ ,  $Y^*$ ,  $\bar{Q}^*$ ,  $\bar{x}_h^*(j_0, \dots, j_{h-1})$ ,  $\bar{a}_h^*(j_0, \dots, j_{h-1})$ , for  $j_0, \dots, j_{N-1} \in$

$\{1, \dots, L\}$  and  $h \in \{1, \dots, N\}$ . Now let

$$\{W^{\text{fs}}, G^{\text{fs}}, Y^{\text{fs}}, \bar{a}_0^{\text{fs}}, \{\bar{x}_h^{\text{fs}}(j_0, \dots, j_h)\}_{h=1}^N, \{\bar{a}_h^{\text{fs}}(j_0, \dots, j_h)\}_{h=1}^N, P^{\text{fs}}\}$$

be a set of arbitrary feasible solutions for Problem  $\mathcal{MPC}$  such that

$$J(x_{k|k}, \bar{a}_0^{\text{fs}}, \dots, \bar{a}_{N-1}^{\text{fs}}(j_0, \dots, j_{N-1}), P^{\text{fs}}) < \gamma_1^*.$$

Then there exists  $\gamma_1^{fs}$  such that

$$J(x_{k|k}, \bar{a}_0^{\text{fs}}, \dots, \bar{a}_{N-1}^{\text{fs}}(j_0, \dots, j_{N-1}), P^{\text{fs}}) \leq \gamma_1^{fs} < \gamma_1^*.$$

By setting  $(\bar{Q}_i^{\text{fs}})^{-1} = P^{\text{fs}}$  and

$$\gamma_2^{\text{fs}}(j_0, \dots, j_{N-1}) = \bar{x}_N^{\text{fs}}(j_0, \dots, j_{N-1})^\top P^{\text{fs}} \bar{x}_N^{\text{fs}}(j_0, \dots, j_{N-1}),$$

it implies  $\{W^{\text{fs}}, G^{\text{fs}}, Y^{\text{fs}}, \bar{a}_0^{\text{fs}}, \{\bar{x}_h^{\text{fs}}(j_0, \dots, j_h)\}_{h=1}^N, \{\bar{a}_h^{\text{fs}}(j_0, \dots, j_h)\}_{h=1}^N, \bar{Q}^{\text{fs}}, \gamma_1^{fs}, \gamma_2^{\text{fs}}(j_0, \dots, j_{N-1})\}$  is a feasible solution of problem (7.68). But this contradicts with the fact that  $\gamma^{fs} < \gamma_1^*$ . Thus, for any feasible solutions in Problem  $\mathcal{MPC}$ ,

$$J(x_{k|k}, \bar{a}_0^{\text{fs}}, \dots, \bar{a}_{N-1}^{\text{fs}}(j_0, \dots, j_{N-1}), P^{\text{fs}}) \geq \gamma_1^*.$$

Now we want to prove the equality when an optimal solution is substituted to the left side. Let

$$\{W^{\text{opt}}, G^{\text{opt}}, Y^{\text{opt}}, \bar{a}_0^{\text{opt}}, \{\bar{x}_h^{\text{opt}}(j_0, \dots, j_h)\}_{h=1}^N, \{\bar{a}_h^{\text{opt}}(j_0, \dots, j_h)\}_{h=1}^N, \bar{Q}^{\text{opt}}\}$$

be a set of optimal solutions in Problem  $\mathcal{MPC}$ . This implies for  $(\bar{Q}^{\text{opt}})^{-1} = P^{\text{opt}}$ ,

$$J_k^*(x_{k|k}) = J(x_{k|k}, \bar{a}_0^{\text{opt}}(j_0), \dots, \bar{a}_{N-1}^{\text{opt}}(j_0, \dots, j_{N-1}), P^{\text{opt}}) \leq J(x_{k|k}, \bar{a}_0^*(j_0), \dots, \bar{a}_{N-1}^*(j_0, \dots, j_{N-1}), P^*)$$

where  $(\bar{Q}^*)^{-1} = P^*$ . But by the nature of the optimization problem in (7.68), the objective epigraph constraint implies

$$J_k^*(x_{k|k}) \leq J(x_{k|k}, \bar{a}_0^*(j_0), \dots, \bar{a}_{N-1}^*(j_0, \dots, j_{N-1}), P^*) \leq \gamma_1^*.$$

Thus, by combining all arguments, we conclude that  $\gamma_1^* = J_k^*(x_{k|k})$ , which completes the proof of this Lemma.  $\square$

### 7.4.11 A Generalized Stability Condition

In this section, we want to extend the stability analysis of Problem  $\mathcal{MPC}$  to history dependent policies. At time  $k$ , define *truncated history* as  $\mathcal{H}_{k+h} = \{x_{k|k}, w_k, \dots, w_{k+h-1|k}\}$  for  $h \in \{0, \dots, N-1\}$  and the *truncated history dependent policy*  $\pi = \{\pi_{k|k}, \dots, \pi_{k+N-1|k}\}$  as

$$\pi_{k+h|k}(\mathcal{H}_{k+h}) = \begin{cases} \pi_{k|k}(x_{k|k}) & \text{if } h = 0 \\ \pi_{k+h|k}(x_{k|k}, w_k, \dots, w_{k+h-1|k}) & \text{otherwise} \end{cases}$$

In order to introduce extra freedom the constraints, we will modify model predictive control problem as follows. Define scenario dependent terminal cost function, control gain and the MPC cost function as follows:

$$\begin{aligned} P(w = w^{[i]}) &= P_i, \quad F(w = w^{[i]}) = F_i, \quad i \in \{1, \dots, L\}, \\ J(x_{k|k}, \pi_{k|k}, \dots, \pi_{k+N-1|k}, \{P_i\}_{i=1}^L) &:= \\ \rho_{k,k+N} &\left( C(x_{k|k}, \pi_{k|k}(\mathcal{H}_{k|k}), \dots, C(x_{k+N-1|k}, \pi_{k+N-1|k}(\mathcal{H}_{k+N-1|k}), x_{k+N}^\top P(w_{k+N-1})x_{k+N|k}) \right). \end{aligned}$$

Now, we modify optimization problem  $\mathcal{PE}$  as follows:

**Optimization problem  $\mathcal{PE}$**  — Given an initial state  $x_0 \in \mathbb{R}^{N_x}$  such that  $\|T_x x_0\|_2 \leq x_{\max}$ , solve

$$\begin{aligned} \max_{W_i = W_i^\top \succ 0, F_i, P_i = P_i^\top \succ 0, \forall i} \quad & \sum_{i=1}^L \log \det(W_i) \\ \text{such that} \quad & W_i \succeq x_0 x_0^\top, \end{aligned} \tag{7.69}$$

$$F_i^\top \frac{T_a^\top T_a}{a_{\max}^2} F_i - W_i^{-1} \preceq 0, \tag{7.70}$$

$$(A_i + B_i F_i)^\top \frac{T_x^\top T_x}{x_{\max}^2} (A_i + B_i F_i) - W_i^{-1} \preceq 0, \tag{7.71}$$

$$(A_i + B_i F_i)^\top W_i^{-1} (A_i + B_i F_i) - W_i^{-1} \preceq 0, \tag{7.72}$$

$$\sum_{j=1}^L q_l(j) (A_j + B_j F_j)^\top P_j (A_j + B_j F_j) - P_i + (F_i^\top R F_i + Q) \prec 0$$

$$\forall l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}, \forall i \in \{1, \dots, L\}. \tag{7.73}$$

We are now in position to modify Theorem 4.6.6 to prove stochastic stability for history dependent policies and terminal cost matrix.

**Theorem 7.4.3.** (*Stochastic Stability for Model Predictive Control Law*) Consider the model predictive control law in equation (4.18) and the corresponding closed-loop dynamics for system (4.1) with initial condition  $x_0 \in \mathbb{R}^{N_x}$ . By implementing the MPC control law, the closed loop system (4.1) is UGRSES.

*Proof.* The strategy of this proof is to show that  $J_k^*$  is a valid Lyapunov function in the sense of Lemma 4.5.1. Specifically, we want to show that  $J_k^*$  satisfies the two inequalities in equation (4.8); the claim then follows by simply noting that, in our time-invariant setup,  $J_k^*$  does not depend on  $k$ .

We start by focusing on the bottom inequality in equation (4.8). Consider a time  $k$  and an initial condition  $x_{k|k} \in \mathbb{R}^{N_x}$  for problem  $\mathcal{MPC}$ . The sequence of optimal randomized control policies is given by  $\{\pi_{k+h|k}^*\}_{h=0}^{N-1}$ . Define the following control policy sequence from time  $k+1$  to  $N$ :

$$\tilde{\pi}_{k+h|k}(x_{k|k}, w_k, \dots, w_{k+h-1|k}) := \begin{cases} \pi_{k+h|k}^*(x_{k|k}, w_k, \dots, w_{k+h-1|k}) & \text{if } h \in [1, N-1], \\ \sum_{i=1}^L F_i x_{k+N|k} \mathbf{1}\{w_{k+N-1|k} = w^{[i]}\} & \text{if } h = N. \end{cases}$$

This control policy is essentially the concatenation of the sequence  $\{\pi_{k+h|k}^*\}_{h=1}^{N-1}$  with a linear feedback control law for stage  $N$ . Since

$$x_{k+N|k} = A(w_{k+N-1|k})x_{k+N-1|k} + B(w_{k+N-1|k})\pi_{k+N-1|k}^*(x_{k|k}, w_k, \dots, w_{k+N-2|k})$$

we can easily justify

$$\sum_{i=1}^L F_i x_{k+N|k} \mathbf{1}\{w_{k+N-1|k} = w^{[i]}\}$$

is a function of  $(x_{k|k}, w_k, \dots, w_{k+N-1|k})$  by induction.

Consider the  $\mathcal{MPC}$  problem at stage  $k+1$  with initial condition given by

$$x_{k+1|k+1} = x_{k+1|k} = A(w_k)x_{k|k} + B(w_k)\pi_{k|k}^*(x_{k|k}).$$

At stage  $k+1$ , the disturbance  $w_k$  is realized and  $x_{k+1|k+1}$  is updated based on the information of  $x_{k|k}$  and  $w_k$ . Accordingly, we can define a sequence of control policies from time  $k+1$  to  $N$  using the following surjective mapping:

$$\begin{aligned} \pi_{k+1|k+1}(x_{k+1|k+1}) &:= \tilde{\pi}_{k+1|k}(x_{k|k}, w_k), \\ \pi_{k+h|k+1}(x_{k+1|k+1}, w_{k+1|k}, \dots, w_{k+h-1|k}) &:= \tilde{\pi}_{k+h|k}(x_{k|k}, w_k, \dots, w_{k+h-1|k}), \quad h \in \{2, \dots, N\}. \end{aligned}$$

Denote with  $\bar{J}_{k+1}(x_{k+1|k+1})$  the cost of the objective function assuming that the sequence of control policies is given by  $\{\pi_{k+h|k+1}\}_{h=1}^N$ . Note that  $x_{k+1|k+1}$  (and therefore  $\bar{J}_{k+1}(x_{k+1|k+1})$ ) is a random variable with  $L$  possible realizations. For any  $i, j \in \{1, \dots, L\}$ , define:

$$\begin{aligned} Z_{k+N}(w_{k+N-1|k} = w^{[i]}) &:= -x_{k+N|k}^\top P_i x_{k+N|k} + x_{k+N|k}^\top Q x_{k+N|k} + (F_i x_{k+N|k})^\top R F_i x_{k+N|k}, \\ Z_{k+N+1}(w_{k+N|k} = w^{[j]}) &:= \left( \left( A(w_{k+N|k} = w^{[j]}) + B(w_{k+N|k} = w^{[j]}) F_j \right) x_{k+N|k} \right)^\top \cdot P_j \\ &\quad \left( \left( A(w_{k+N|k} = w^{[j]}) + B(w_{k+N|k} = w^{[j]}) F_j \right) x_{k+N|k} \right). \end{aligned}$$

By the dual representation of Markov polytopic risk, the condition in  $\mathcal{S}(\{Y_i\}_{i=1}^L, \{G_i\}_{i=1}^L)$  implies

$$Z_{k+N}(w_{k+N-1|k} = w^{[i]}) + \rho(Z_{k+N+1}(w_{k+N|k})) \leq 0, \quad \text{surely, } \forall i \in \{1, \dots, L\}. \quad (7.74)$$

Similar to the arguments in expression (7.62), one can then write the following chain of inequalities:

$$\begin{aligned} & J_k^*(x_{k|k}) \\ &= x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) \\ & \quad + \rho \left( \rho_{k+1,N} \left( C(x_{k+1|k}, \pi_{k+1}^*(x_{k|k}, w_k)), \dots, x_{k+N|k}^\top P(w_{k+N-1|k}) x_{k+N|k} \right) \right) \\ & \geq x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho \left( \rho_{k+1,N} \left( C(x_{k+1|k}, \pi_{k+1}^*(x_{k|k}, w_k)), \dots, x_{k+N|k}^\top Q x_{k+N|k} + \right. \right. \\ & \quad \left. \left. (F(w_{k+N-1|k}) x_{k+N|k})^\top R F(w_{k+N-1|k}) x_{k+N|k} + \rho(Z_{k+N+1}(w_{k+N|k})) \right) \right) \\ &= x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho(\bar{J}_{k+1}(x_{k+1|k+1})) \\ & \geq x_{k|k}^\top Q x_{k|k} + (\pi_{k|k}^*(x_{k|k}))^\top R \pi_{k|k}^*(x_{k|k}) + \rho(J_{k+1}^*(x_{k+1|k+1})). \end{aligned} \quad (7.75)$$

Notice that the analysis of the top inequality in equation (4.8) follows analogously to the arguments in Theorem 4.6.6. Combining the above results and given the time-invariance of our problem setup, one concludes that  $J_k^*(x_{k|k})$  is a risk-sensitive Lyapunov function for the closed-loop system (4.1), in the sense of Lemma 4.5.1. This concludes the proof.  $\square$

Furthermore, corresponding to the inequality in  $\mathcal{S}(\{Y_i\}_{i=1}^L, \{G_i\}_{i=1}^L)$  and based on Projection Lemma, one can derive the following semi-definite feasibility condition for  $Y_i, G_i, \bar{Q}_i = \bar{Q}_i^\top \succ 0, i \in \{1, \dots, L\}$ :

$$\bar{Q} = \begin{bmatrix} \bar{Q}_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \bar{Q}_L \end{bmatrix}, \quad \begin{bmatrix} \bar{Q} & 0 & 0 & -\Sigma_l^{\frac{1}{2}}(\bar{A}G_i + \bar{B}Y_i) \\ * & R^{-1} & 0 & -Y_i \\ * & * & I & -Q_i^{\frac{1}{2}}G_i \\ * & * & * & -\bar{Q}_i + G_i + G_i^\top \end{bmatrix} \succ 0, \quad i \in \{1, \dots, L\} \quad (7.76)$$

where  $l \in \{1, \dots, \text{cardinality}(\mathcal{U}^{\text{poly}, V}(p))\}$ ,  $\bar{Q}_i = P_i^{-1}$  and  $F_i = Y_i G_i^{-1}$ . Thus, based on the new problem formulation, one can also slightly modify Lemma 7.4.2 to get a modified MPC solution algorithm for history dependent policies with scenario dependent terminal cost functions.

### 7.4.12 Alternative Formulation of Problem $\mathcal{PE}$ and $\mathcal{MPC}$

In this section we present alternative formulations of problems  $\mathcal{PE}$  and  $\mathcal{MPC}$  inspired by the approach in [15]. The methodology here is to design (offline) an equivalent control invariant set  $\mathcal{E}_{\max}$  and a *robust* Lyapunov function such that ULRSES and constraint fulfillment are guaranteed using a local state feedback control law  $a(x) = Fx$ . Let  $P = P^\top \succ 0$  and  $L = L^\top \succ 0$ . Define  $V(x_k) = x_k^\top P x_k$ . If

$$V(x_{k+1}) - V(x_k) \leq -x_k^\top L x_k, \quad \text{surely, } \forall k \in \mathbb{N}, \quad (7.77)$$

then  $V(x_k)$  is a *robust* Lyapunov function for system (4.1). In the online problem, the inequality above is relaxed to its stochastic counterpart as shown in (4.8). We first formalize the offline optimization problem:

**Optimization problem  $\mathcal{PE}$**  — Given an initial state  $x_0 \in \mathbb{X}$ , and a matrix  $L = L^\top \succ 0$ , solve

$$\begin{aligned} & \max_{W=W^\top \succ 0, Y, \gamma > 0} \quad \log \det(W) \\ & \text{such that} \quad x_0^\top W^{-1} x_0 \leq 1, \\ & \quad Y^\top \frac{T_a^\top T_a}{a_{\max}^2} Y \preceq W, \\ & \quad (A_j W + B_j Y)^\top \frac{T_x^\top T_x}{x_{\max}^2} (A_j W + B_j Y) \preceq W, \quad \forall j \in \{1, \dots, L\}, \\ & \quad \begin{bmatrix} W & (L^{1/2} W)^\top & (A_j W + B_j Y)^\top \\ * & \gamma I_{N_x} & 0 \\ * & 0 & W \end{bmatrix} \succeq 0, \quad \forall j \in \{1, \dots, L\}. \end{aligned} \quad (7.78)$$

Suppose problem  $\mathcal{PE}$  above is feasible. Set  $P = \gamma W^{-1}$ . The control invariant set is then defined to be the intersection  $\mathbb{X} \cap \mathcal{E}_{\max}$ , where

$$\mathcal{E}_{\max}(W) := \{x \in \mathbb{R}^{N_x} \mid x^\top W^{-1} x \leq 1\} = \{x \in \mathbb{R}^{N_x} \mid x^\top P x \leq \gamma\}.$$

Note that  $\mathbb{X} \cap \mathcal{E}_{\max}$  is a robust control invariant set under the feasible local feedback control law  $a(x) = YW^{-1}x$ . The constraint in (7.78) is an equivalent reformulation of the robust Lyapunov condition given in (7.77) where  $x_{k+1} = (A(w) + B(w)YG^{-1})x$ . That is, the closed-loop dynamics are ULRSES with domain  $\mathbb{X} \cap \mathcal{E}_{\max}$  under the feedback control law  $a(x) = YG^{-1}x$ . In an attempt to improve the stability properties of the system beyond what is achievable via this feedback control law, the online MPC problem is formalized as follows:

**Optimization problem MPC** — Given an initial state  $x_{k|k} \in \mathbb{X} \cap \mathcal{E}_{\max}$  and a prediction horizon  $N \geq 1$ , solve

$$\begin{aligned}
 & \min_{\pi_{k|k}, \dots, \pi_{k+N-1|k}} J(x_{k|k}, \pi_{k|k}, \dots, \pi_{k+N-1|k}, P) \\
 & \text{such that } x_{k+h+1|k} = A(w_{k+h})x_{k+h|k} + B(w_{k+h})\pi_{k+h|k}(x_{k+h|k}), \\
 & \|T_a \pi_{k+h|k}(x_{k+h|k})\|_2 \leq a_{\max}, \|T_x x_{k+h+1|k}\|_2 \leq x_{\max}, h \in \{0, \dots, N-1\}, \\
 & x_{k+1|k} \in \mathcal{E}_{\max}(W) \text{ surely,} \tag{7.79} \\
 & \rho((Ax_{k|k} + B\pi_{k|k}(x_{k|k}))^\top P(Ax_{k|k} + B\pi_{k|k}(x_{k|k}))) - x_k^\top P x_k \leq -x_k^\top L x_{k|k}. \tag{7.80}
 \end{aligned}$$

where the final constraint may be enforced by evaluating the expectation form of the one-step conditional risk measure at each vertex of the polytope:  $\mathcal{U}^{\text{poly}, V}(p)$ .

Provided problem MPC is recursively feasible, ULRSES with domain  $\mathbb{X} \cap \mathcal{E}_{\max}$  is enforced automatically via the constraint in (7.80) which leverages the risk-sensitive Lyapunov function  $x_k^\top P x_k$ , where  $P$  is the solution to the offline problem. Persistent feasibility however, is guaranteed by the constraint in (7.79).

#### 7.4.13 Suboptimality Performance of $\pi^{\text{MPC}}$

For  $k \geq 0$  consider the  $N$ -step optimal cost function with terminal cost  $x_{k+N|k}^\top P x_{k+N|k}$ :

$$\begin{aligned}
 J(x_{k|k}, N, P) &:= \\
 & \min_{\pi_{k|k}, \dots, \pi_{k+N-1|k}} \underbrace{\rho \circ \dots \circ \rho}_N (C(x_{k|k}, \pi_{k|k}(x_{k|k})), \dots, C(x_{k+N-1|k}, \pi_{k+N-1|k}(x_{k+N-1|k})), x_{k+N|k}^\top P x_{k+N|k}).
 \end{aligned}$$

The theory of solving this optimal control problem by dynamic programming can be found in [119]. Before getting to the main result, we have the following technical lemma.

**Lemma 7.4.4.** *Let . For any initial state  $x_{k|k} \in \mathbb{R}^{N_x}$ , we have that  $J(x_{k|k}, N, P) \leq (\gamma+1)C(x_{k|k}, \pi_k^*(x_{k|k}))$ , where the constant  $\gamma$  is given by:*

$$\gamma = \frac{c}{\underline{\sigma}(Q + F^\top R F)} \left( \bar{\sigma}(Q + F^\top R F) \frac{1}{1-\lambda} + \bar{\sigma}(\lambda^N P) \right) - 1. \tag{7.81}$$

with  $F = YG^{-1}$  is the state-feedback control gain that satisfies (7.70), (7.71), (7.72), (4.20).

*Proof.* First, consider the following inequalities for each  $j \geq 0$ :

$$\begin{aligned}
 & \underbrace{\rho \circ \dots \circ \rho}_j (0, \dots, C(x_{k+j|k}, Fx_{k+j|k})) \leq \underbrace{\rho \circ \dots \circ \rho}_j (0, \dots, x_{k+j|k}^\top (Q + F^\top R F) x_{k+j|k}) \\
 & \leq \bar{\sigma}(Q + F^\top R F) \underbrace{\rho \circ \dots \circ \rho}_j (0, \dots, x_{k+j|k}^\top x_{k+j|k}) \leq c\lambda^j \bar{\sigma}(Q + F^\top R F) x_k^\top x_{k|k}
 \end{aligned} \tag{7.82}$$



for some  $c > 0$  and  $\lambda \in (0, 1)$ . The first inequality is due to monotonicity of time consistent Markov risk measures and the second inequality is based on the UGRSES condition of  $\{x_{k|k}\}$  induced by the closed loop control sequence  $\pi_{k|k}(x_{k|k}) = Fx_{k|k}$ . The proof of UGRSES when  $\pi_{k|k}(x_{k|k}) = Fx_{k|k}$  follows from the stochastic stability analysis of Algorithm  $\mathcal{MPC}^0$  using stochastic Lyapunov function  $x^\top Px$ ,  $P = P^\top \succ 0$ .

Then, for any  $i \geq 0$ , we have the following inequalities:

$$\begin{aligned}
J(x_{k|k}, N, P) &:= \underbrace{\rho \circ \dots \circ \rho}_N (C(x_{k|k}, \pi_{k|k}^*(x_{k|k})), \dots, C(x_{N-1+k|k}, \pi_{k+N-1|k}^*(x_{N-1+k|k})), P) \\
&\leq \underbrace{\rho \circ \dots \circ \rho}_N (C(x_{k|k}, Fx_{k|k}), \dots, C(x_{N-1+k|k}, Fx_{N-1+k|k}), P) \\
&\leq \lim_{M \rightarrow \infty} \underbrace{\rho \circ \dots \circ \rho}_M (C(x_{k|k}, Fx_{k|k}), \dots, C(x_{M+k|k}, Fx_{M+k|k})) + \underbrace{\rho \circ \dots \circ \rho}_N (x_{k+N-1|k}^\top P x_{k+N-1|k}) \\
&\leq \lim_{M \rightarrow \infty} \sum_{j=0}^M \underbrace{\rho \circ \dots \circ \rho}_j (0, \dots, C(x_{k+j|k}, Fx_{k+j|k})) + \underbrace{\rho \circ \dots \circ \rho}_N (x_{k+N-1|k}^\top P x_{k+N-1|k}) \\
&\leq \lim_{M \rightarrow \infty} \sum_{j=0}^M c\lambda^j \bar{\sigma}(Q + F^\top RF) x_{k|k}^\top x_{k|k} + c\lambda^N \bar{\sigma}(P) x_{k|k}^\top x_{k|k} \\
&= \left( \bar{\sigma}(Q + F^\top RF) \frac{c}{1-\lambda} + c\bar{\sigma}(\lambda^N P) \right) x_{k|k}^\top x_{k|k} \\
&\leq \frac{c}{\underline{\sigma}(Q + F^\top RF)} \left( \bar{\sigma}(Q + F^\top RF) \frac{1}{1-\lambda} + \bar{\sigma}(\lambda^N P) \right) C(x_{k|k}, \pi_{k|k}^*(x_{k|k})).
\end{aligned} \tag{7.83}$$

The first inequality is due to the fact that  $Fx_{k+j|k}$  is a feasible solution to the  $N$  step MPC problem. The second inequality is due to monotonicity and sub-additivity of Markov risk measures, and the third inequality is due to sub-additivity (convexity) of Markov risk measures. The fourth inequality follows from the expression in (7.82), and the last inequality is due to the fact that

$$C(x_{k|k}, Fx_{k|k}) \geq \underline{\sigma}(Q + F^\top RF) x_{k|k}^\top x_{k|k}.$$

Substituting the definition of  $\gamma$  to the above expression completes the proof of this Lemma.  $\square$

Recall that the MPC control law  $\pi^{MPC}$  is the optimal policy corresponds to the current initial state  $x_{k|k}$  in the finite horizon problem. That is,  $\pi^{MPC}(x_{k|k}) = \pi_{k|k}^*(x_{k|k})$  at stage  $k$  where  $\{\pi_k^*, \dots, \pi_{k+N-1}^*\}$  is the sequence of optimal policies. From Lemma 7.4.4 and the Bellman optimality of value function  $J(x_{k|k}, N, 0)$ , it can be easily seen that

$$\rho(J(x_{k+1|k}, N-1, P)) = J(x_{k|k}, N, P) - C(x_{k|k}, \pi^{MPC}(x_{k|k})) \leq \gamma C(x_{k|k}, \pi^{MPC}(x_{k|k})). \tag{7.84}$$

Recall the realization constant  $\gamma$  in Lemma 7.4.4. Define the following constant:

$$\eta_N = \frac{(\gamma + 1)^{N-2}}{(\gamma + 1)^{N-2} + \gamma^{N-1}} \in (0, 1). \quad (7.85)$$

Again by Bellman optimality of  $J(x_{k|k}, N, 0)$  and the state update  $x_{k+1|k+1} = x_{k+1|k}$ , above expression immediately implies that

$$\begin{aligned} J(x_{k|k}, N, P) &= \rho(J(x_{k+1|k}, N-1, P)) + C(x_{k|k}, \pi^{MPC}(x_{k|k})) \\ &\geq \left(1 - \gamma \frac{1 - \eta_N}{\gamma + \eta_N}\right) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \left(1 + \frac{1 - \eta_N}{\gamma + \eta_N}\right) \rho(J(x_{k+1|k}, N-1, P)) \\ &= \left(1 - \gamma \frac{1 - \eta_N}{\gamma + \eta_N}\right) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \left(1 + \frac{1 - \eta_N}{\gamma + \eta_N}\right) \rho(J(x_{k+1|k+1}, N-1, P)) \end{aligned}$$

Recall the result in Lemma 7.4.4 and the definition of  $\eta_N$ . Starting at state  $x_{k+1|k+1} = x_{k+1|k}$  the following expression holds

$$J(x_{k+1|k+1}, 0, P) \geq \frac{1}{1 + \gamma} J(x_{k+1|k+1}, 1, P).$$

Thus for  $N = 2$ , from the definition of  $\eta_N$  in (7.85),

$$J(x_{k+1|k+1}, N-1, P) \geq \eta_N J(x_{k+1|k+1}, N-2, P).$$

Followed by inductive arguments, the following expression holds for any  $x_{k|k} \in \mathbb{R}^{N_x}$  and  $N > 2$ :

$$\begin{aligned} J(x_{k|k}, N, P) &\geq \left(1 - \gamma \frac{1 - \eta_N}{\gamma + \eta_N}\right) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \eta_N \left(1 + \frac{1 - \eta_N}{\gamma + \eta_N}\right) \rho(J(x_{k+1|k+1}, N-2, P)) \\ &= \left(1 - \gamma \frac{1 - \eta_N}{\gamma + \eta_N}\right) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \eta_N \left(1 + \frac{1 - \eta_N}{\gamma + \eta_N}\right) \rho(J(x_{k+1|k}, N-2, P)) \\ &= \eta_N \frac{\gamma + 1}{\gamma + \eta_N} J(x_{k|k}, N-1, P) = \eta_{N+1} J(x_{k|k}, N-1, P) \end{aligned} \quad (7.86)$$

The last equality is due to the fact that

$$\eta_{N+1} = \frac{(\gamma + 1)^{N-1}}{(\gamma + 1)^{N-1} + \gamma^N} = \frac{(\gamma + 1)^{N-2}}{(\gamma + 1)^{N-2} + \gamma^{N-1}} \left( \frac{\gamma + 1}{\gamma + \frac{(\gamma + 1)^{N-2}}{(\gamma + 1)^{N-2} + \gamma^{N-1}}} \right) = \eta_N \left( 1 + \frac{1 - \eta_N}{\gamma + \eta_N} \right).$$

Notice that inequality (7.86) is still be valid if  $x_{k|k}$  is replaced by  $x_{k+1|k+1}$  for any fixed  $x_{k+1|k+1} \in \mathbb{R}^{N_x}$ .

Then the above expression implies

$$\begin{aligned} J(x_{k+1|k+1}, N, P) - J(x_{k+1|k+1}, N-1, P) &\leq \left( \frac{1}{\eta_{N+1}} - 1 \right) J(x_{k+1|k+1}, N, P) \\ &= \left( \frac{1}{\eta_{N+1}} - 1 \right) J(x_{k+1|k}, N, P) \leq \left( \frac{1}{\eta_{N+1}} - 1 \right) \gamma C(x_{k|k}, \pi^{MPC}(x_{k|k})) \end{aligned}$$

where the second inequality is due to expression (7.84). By re-arranging this inequality, one obtains

$$\begin{aligned} J(x_{k+1|k+1}, N, P) &\leq \left( \frac{1}{\eta_{N+1}} - 1 \right) \gamma C(x_{k|k}, \pi^{MPC}(x_{k|k})) + J(x_{k+1|k+1}, N-1, P), \\ &= (1 - \beta_N) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + J(x_{k+1|k+1}, N-1, P) \\ &= (1 - \beta_N) C(x_{k|k}, \pi^{MPC}(x_{k|k})) + J(x_{k+1|k}, N-1, P) \end{aligned}$$

where the constant  $\beta_N$  is defined as

$$\beta_N = 1 - \left( \frac{(\gamma + 1)^{N-2} + \gamma^{N-1}}{(\gamma + 1)^{N-2}} - 1 \right) \gamma = 1 - \frac{\gamma^N}{(\gamma + 1)^{N-2}} = \frac{(\gamma + 1)^{N-2} - \gamma^N}{(\gamma + 1)^{N-2}} \in (0, 1). \quad (7.87)$$

Notice that by applying the risk measure  $\rho$  on both sides, the above expression becomes

$$\beta_N C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \rho(J(x_{k+1|k+1}, N, P)) \leq J(x_{k|k}, N, P).$$

Therefore, one obtains the following expression by recursively expanding  $J(x_{k+1|k+1}, N, 0)$  with the above arguments:

$$\begin{aligned} &\beta_N C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \rho(\beta_N C(x_{k+1|k+1}, \pi^{MPC}(x_{k+1|k+1})) + \rho(J(x_{k+2|k+2}, N, P))) \\ &\leq \beta_N C(x_{k|k}, \pi^{MPC}(x_{k|k})) + \rho(J(x_{k+1|k+1}, N, P)) \leq J(x_{k|k}, N, P). \end{aligned}$$

Furthermore, for any  $M \in \mathbb{Z}$ , by repeating the above analysis from  $k$  to  $k + M - 1$  and noticing that  $J(x_{k+M|k+M}, N, P) \geq 0$ , monotonicity and positive homogeneity of risk measure  $\rho$  imply

$$\begin{aligned} &\beta_N \underbrace{\rho \circ \dots \circ \rho}_M (C(x_{k|k}, \pi^{MPC}(x_{k|k})), \dots, C(x_{k+M-1|k+M-1}, \pi^{MPC}(x_{k+M-1|k+M-1}))) \\ &\leq \underbrace{\rho \circ \dots \circ \rho}_M \left( \beta_N C(x_{k|k}, \pi^{MPC}(x_{k|k})), \dots, \beta_N C(x_{k+M-1|k+M-1}, \pi^{MPC}(x_{k+M-1|k+M-1})) \right) \\ &\quad + J(x_{k+M|k+M}, N, P) \leq J(x_{k|k}, N, P). \end{aligned}$$

Finally, when  $M$  goes to infinity, the above expression implies

$$\begin{aligned}
& \beta_N J_{0,\infty}^*(x_{k|k}) \\
& \leq \beta_N \lim_{M \rightarrow \infty} \underbrace{\rho \circ \dots \circ \rho}_M (C(x_{k|k}, \pi^{MPC}(x_{k|k})), \dots, C(x_{k+M-1|k+M-1}, \pi^{MPC}(x_{k+M-1|k+M-1}))) \\
& \leq J(x_{k|k}, N, P) \leq J_{0,\infty}^*(x_{k|k}).
\end{aligned}$$

The first inequality is based on the fact that MPC control policies are feasible, thus the induced cost function is larger than the optimal value function  $J_{0,\infty}^*(x_{k|k})$ . The second inequality is by the recursive analysis given above. The third inequality is from the fact that with nonnegative stage-wise cost  $C(x, a)$ , monotonicity of risk measure  $\rho$  implies

$$\begin{aligned}
& J(x_{k|k}, N, P) \\
& = \min_{\pi_{k|k}, \dots, \pi_{k+N-1|k}} \underbrace{\rho \circ \dots \circ \rho}_N (C(x_{k|k}, \pi_{k|k}(x_{k|k})), \dots, C(x_{k+N-1|k}, \pi_{k+N-1|k}(x_{k+N-1|k})), P) \\
& \leq \min_{\pi_{k|k}, \dots, \pi_{k+N|k}} \underbrace{\rho \circ \dots \circ \rho}_{N+1} (C(x_{k|k}, \pi_{k|k}(x_{k|k})), \dots, C(x_{k+N|k}, \pi_{k+N|k}(x_{k+N|k})), P) \\
& \leq \dots \leq \lim_{N \rightarrow \infty} \min_{\pi_{k|k}, \dots, \pi_{k+N|k}} \underbrace{\rho \circ \dots \circ \rho}_{N+1} (C(x_{k|k}, a_k), \dots, C(x_{k+N|k}, \pi_{k+N|k}(x_{k+N|k})), P) \\
& = J_{0,\infty}^*(x_{k|k}).
\end{aligned}$$

The equality follows from the analysis in (7.83) with  $\lambda^N \rightarrow 0$  as  $N \rightarrow \infty$ . The result of the sub-optimality performance bound is summarized in the following theorem.

**Theorem 7.4.5.** *Let  $x_{k|k} \in \mathbb{R}^{N_x}$  be the initial state at stage  $k$  and  $N$  be the MPC lookahead horizon. The infinite horizon cost function induced by the MPC control policy  $\pi^{MPC}$  has the following sub-optimal performance bound:*

$$\begin{aligned}
& J_{0,\infty}^*(x_{k|k}) \\
& \leq \lim_{M \rightarrow \infty} \underbrace{\rho \circ \dots \circ \rho}_M (C(x_{k|k}, \pi^{MPC}(x_{k|k})), \dots, C(x_{k+M-1|k+M-1}, \pi^{MPC}(x_{k+M-1|k+M-1}))) \\
& \leq \frac{J_{0,\infty}^*(x_{k|k})}{\beta_N}.
\end{aligned}$$

where  $J_{0,\infty}^*(x_{k|k})$  is the optimal solution of problem  $\mathcal{OPT}_{RS}$  with  $x_{k|k}$  being the initial state and the performance coefficient  $\beta_N$  is given in (7.87).

The above theorem shows that the MPC solution is  $1/\beta_N$ -optimal for  $\beta_N \in (0, 1)$ . When  $N$  tends to infinity, the definition of  $\beta_N$  implies that  $\beta_N$  goes to 1, which means that the MPC solution is optimal.

## 7.5 Technical Results in Chapter 5

### 7.5.1 Proof of Theorem 5.3.2

The proof style is similar to that of Theorem 3.1 in [40]. The proof consists of two steps. First, we show that  $V^*(x, d) \geq \mathbf{T}[V^*](x, d)$  for all pairs  $(x, d) \in \mathcal{X} \times \mathbb{R}$ . Second, we show  $V^*(x, d) \leq \mathbf{T}[V^*](x, d)$  for all pairs  $(x, d) \in \mathcal{X} \times \mathbb{R}$ . These two results will prove the claim that  $V^*$  is a fixed point solution to the Bellman's equation.

*Step (1).* If  $d \notin \Phi(x)$ , then, by definition,  $V^*(x, d) = \infty$ . Also,  $d \notin \Phi(x)$  implies that  $F(x, d)$  is empty. Hence,  $\mathbf{T}[V^*](x, d) = \infty$ . Therefore, if  $d \notin \Phi(x_k)$ ,

$$V^*(x, d) = \infty = \mathbf{T}[V^*](x, d), \quad (7.88)$$

i.e.,  $V^*(x, d) \geq \mathbf{T}[V^*](x, d)$ .

Now assume  $x_0 = x$  and  $d_0 = d$  such that  $d \in \Phi(x)$ . Let  $\pi^* \in \Pi_H$  be an optimal policy that yields the optimal cost  $V^*(x, d)$ . Construct the “truncated” policy  $\bar{\pi} = \{\bar{\mu}_1, \bar{\mu}_2, \dots\}$  according to:

$$\bar{\mu}_j(h_{1,j}) := \mu_j^*(x_0, \mu_0^*(x_0), h_{1,j}), \quad \text{for } j \geq 1.$$

In other words,  $\bar{\pi}$  is a tail policy prescribed by  $\pi^*$ . By applying the law of total expectation, we can write:

$$\begin{aligned} V^*(x, d) &= \lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=0}^{N-1} \gamma^t C(x_t, \mu_t^*(h_{0,t})) \right] = C(x, \mu_0^*(x)) + \lim_{N \rightarrow \infty} \gamma \mathbb{E} \left[ \sum_{t=1}^{N-1} \gamma^t C(x_t, \mu_j^*(h_{0,t})) \right] \\ &= C(x, \mu_0^*(x)) + \gamma \mathbb{E} \left[ \lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=1}^{N-1} \gamma^t C(x_t, \mu_j^*(h_{0,t})) \mid h_{0,1} \right] \right]. \end{aligned}$$

Note that  $\lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=1}^{N-1} \gamma^t C(x_t, \pi_j^*(h_{0,t})) \mid h_{0,1} \right] = \mathcal{C}^{\bar{\pi}}(x_1)$ . Clearly, the truncated policy  $\bar{\pi}$  is a feasible policy for the tail subproblem

$$\begin{aligned} &\min_{\pi \in \Pi_H} \mathcal{C}^{\pi}(x_1) \\ &\text{subject to } \mathcal{D}^{\pi}(x_1) \leq \mathcal{D}^{\bar{\pi}}(x_1). \end{aligned}$$

Collecting the above results, we can write

$$V(x, d) = C(x_0, \pi_0^*(x_0)) + \gamma \mathbb{E} [\mathcal{C}^{\bar{\pi}}(x_1)] \geq C(x_0, \pi_0^*(x_0)) + \gamma \mathbb{E} [V_1(x_1, \mathcal{D}^{\bar{\pi}}(x_1))] \geq \mathbf{T}[V^*](x, d),$$

where the last inequality follows from the fact that  $\mathcal{D}^{\bar{\pi}}(\cdot)$  can be viewed as a valid threshold function in the minimization in equation (5.1).

*Step (2).* If  $d \notin \Phi(x)$ , equation (7.88) holds and, therefore,  $V(x, d) \leq \mathbf{T}[V^*](x, d)$ .

Assume  $d \in \Phi(x)$  (which implies that  $F(x, d)$  is non-empty). For a given pair  $(x, d)$ , where  $d \in \Phi(x)$ , let  $a^*$  and  $d'^*$  be minimizers in equation (5.1) (here we are exploiting the assumption that the minimization

problem in equation (5.1) admits a minimizer). By definition,  $d'^*,(x') \in \Phi(x')$  for all  $x' \in \mathcal{X}$ . Also, let  $\pi^* \in \Pi_H$  be an optimal policy for the tail subproblem:

$$\begin{aligned} & \min_{\pi \in \Pi_H} \mathcal{C}^\pi(x') \\ & \text{subject to } \mathcal{D}^\pi(x') \leq d'^*,(x'). \end{aligned}$$

Construct the “extended” policy  $\bar{\pi} \in \Pi_H$  as follows:

$$\bar{\pi}_0(x) = a^*, \text{ and } \bar{\pi}_j(h_{0,j}) = \pi_j^*(h_{1,j}) \text{ for } j \geq 1.$$

Since  $\pi^*$  is an optimal, and a fortiori feasible, policy for the tail subproblem (from stage 1, starting at state  $x_1 = x'$  and constraint threshold  $d_1 = d'^*,(x')$ ) with threshold function  $d'^*,$  the policy  $\bar{\pi} \in \Pi_H$  is a feasible policy for the tail subproblem (from stage 0, starting at state  $x_0 = x$  and constraint threshold  $d_0 = d$ ):

$$\begin{aligned} & \min_{\pi \in \Pi_H} \mathcal{C}^\pi(x) \\ & \text{subject to } \mathcal{D}^\pi(x) \leq d. \end{aligned}$$

Hence, we can write

$$V^*(x, d) \leq \mathcal{C}^{\bar{\pi}}(x) = C(x, \bar{\mu}_0(x)) + \lim_{N \rightarrow \infty} \mathbb{E} \left[ \mathbb{E} \left[ \sum_{t=1}^{N-1} \gamma^t C(x_t, \bar{\mu}_t(h_{0,t})) \mid h_{0,1} \right] \right].$$

Note that

$$\lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=1}^{N-1} \gamma^{t-1} C(x_t, \bar{\mu}_t(h_{0,t})) \mid h_{0,1} \right] = \mathcal{C}^{\pi^*}(x').$$

Hence, from the definition of  $\pi^*$ , one easily obtains:

$$V^*(x, d) \leq C(x, \bar{\pi}_0(x)) + \gamma \mathbb{E} [\mathcal{C}^{\pi^*}(x')] = C(x, a^*) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a^*) V^*(x', d'^*,(x')) = \mathbf{T}[V^*](x, d).$$

Collecting the above results, we have shown that  $V^*$  is a fixed point solution to Bellman’s equation  $V(x, d) = \mathbf{T}[V](x, d), \forall x, d$ . To show that  $V^*$  is the unique solution to the fixed point equation, according to Lemma 5.3.1,  $\mathbf{T}$  is a contraction mapping. Therefore Proposition 2.2 of [17] immediately implies that the fixed point equation  $\mathbf{T}[V](x, d) = V(x, d), \forall x, d$ , has a unique solution, which is  $V^*$ . This completes the proof of this theorem.

### 7.5.2 Proof of Theorem 5.3.6

Similar to the definition of the optimal Bellman operator  $\mathbf{T}$ , for any augmented stationary Markovian policy  $u : \mathcal{X} \times \mathbb{R} \rightarrow \mathcal{A}$  and any risk-to-go function  $d'(x, d)(\cdot)$  such that

$$D(x, u(x, d)) + \gamma \rho(d'(x, d)(x')) \leq d, \quad (7.89)$$

we define the policy induced Bellman operator  $\mathbf{T}_u$  as

$$\mathbf{T}_u[V](x, d) = C(x, u(x, d)) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, u(x, d)) V(x', d'(x, d)(x')).$$

Analogous to Theorem 5.3.2, we can easily show that the fixed point solution to  $\mathbf{T}_u[V](x, d) = V(x, d)$  is uniquely equal to the value function

$$V_u(x, d) = \lim_{N \rightarrow \infty} \mathbb{E} \left[ \sum_{t=0}^{N-1} \gamma^t C(x_t, a_t) \mid x_0 = x, \pi_H \right],$$

where the history dependent policy  $\pi_H = \{\mu_0, \mu_1, \dots\}$  is given by  $\mu_k(h_k) = u(x_k, d_k)$  for any  $k \geq 0$  with initial state  $x_0 = x$ , constraint threshold  $d_0 = d$ , and the state transitions are given by expression (5.3), but with augmented stationary Markovian policy  $u^*$  replaced by  $u$  and the risk-to-go  $d'^*(x, d)(\cdot)$  replaced by  $d'(x, d)(\cdot)$ . On the other hand, by recursively applying (7.89) at state  $(x_k, d_k)$ , for  $k \in \{0, 1, \dots\}$ , we immediately show that policy  $\pi_H$  is feasible to problem  $\mathcal{OPT}_{RC}$ , i.e.,

$$\lim_{N \rightarrow \infty} \rho_{0, N-1} \left( D(x_0, a_0), \dots, \gamma^{N-1} D(x_{N-1}, a_{N-1}) \right) \mid x_0 = x, \pi_H \leq d.$$

To complete the proof of this theorem, we need to show that the augmented stationary Markovian policy  $u^*$  is optimal if and only if

$$\mathbf{T}[V^*](x, d) = \mathbf{T}_{u^*}[V^*](x, d), \quad \forall x \in \mathcal{X}, d \in \mathbb{R}, \quad (7.90)$$

where  $V^*(x, d)$  is the unique fixed point solution of  $\mathbf{T}[V](x, d) = V(x, d)$ . Here an augmented stationary Markovian policy  $u^*$  is optimal if and only if the induced history dependent policy  $\pi_H^*$  in (5.2) is optimal to problem  $\mathcal{OPT}_{RC}$ .

First suppose  $u^*$  is an optimal augmented stationary Markovian policy. Then using the definition of  $u^*$  and the result from Theorem 5.3.2, we immediately show that  $V^*(x, d) = V_{u^*}(x, d)$ , where by definition  $V_{u^*}$  is the fixed point solution to  $V(x, d) = \mathbf{T}_{u^*}[V](x, d)$  for any  $x, d$ . By the fixed point equation  $\mathbf{T}[V^*](x, d) = V^*(x, d)$  and  $\mathbf{T}_{u^*}[V_{u^*}](x, d) = V_{u^*}(x, d)$ , this further implies (7.90) holds.

Second suppose  $u^*$  satisfies the equality in (7.90). Then by the fixed point equality  $\mathbf{T}[V^*](x, d) = V^*(x, d)$ , we immediately obtain the equation  $V^*(x, d) = \mathbf{T}_{u^*}[V^*](x, d)$  for any  $x \in \mathcal{X}$  and  $d \in \mathbb{R}$ . since the fixed point solution to  $\mathbf{T}_{u^*}[V](x, d) = V(x, d)$  is unique, we further show that  $\mathbf{T}[V^*](x, d) =$

$V^*(x, d) = V_{u^*}(x, d)$ . Furthermore by Theorem 5.3.2 we have that

$$\begin{aligned} V_{u^*}(x, d) &= \min_{\pi \in \Pi_H} \mathcal{C}^\pi(x) \\ \text{subject to } &\mathcal{D}^\pi(x) \leq d. \end{aligned}$$

By using the policy construction formula in (5.2) to obtain the history dependent policy  $\pi_H^*$  and following the above arguments (where the augmented Markovian stationary policy  $u$  is replaced by  $u^*$ , and the risk-to-go function  $d'$  is replaced by  $d'^*$ ), this further implies

$$\begin{aligned} \mathcal{C}^{\pi_H^*}(x) &= V_{u^*}(x, d) = \min_{\pi \in \Pi_H} \mathcal{C}^\pi(x) \\ \text{subject to } &\mathcal{D}^\pi(x) \leq d, \end{aligned}$$

and  $\mathcal{D}^{\pi_H^*}(x) \leq d$ , i.e.,  $u^*$  is an optimal augmented stationary Markovian policy.

### 7.5.3 Proof of Lemma 5.4.3

Before proving the main result, we first show the Lipschitz-ness of set-valued mapping  $\mathcal{U}(x, a, P)$  in the following technical result.

**Proposition 7.5.1.** *For any  $\xi \in \mathcal{U}(x, a, P)$ , there exists a  $M_\xi > 0$  such that for some  $\tilde{\xi} \in \mathcal{U}(x, a, \tilde{P})$ , and  $q(x') = \xi(x')P(x'|x, a)$ ,  $\tilde{q}(x') = \tilde{\xi}(x')P(x'|x, a)$ ,*

$$\sum_{x' \in \mathcal{X}} |q(x') - \tilde{q}(x')| \leq M_\xi \sum_{x' \in \mathcal{X}} |P(x'|x, a) - \tilde{P}(x'|x, a)|.$$

*Proof.* We know that  $\mathcal{U}(x, a, P)$  is a closed, bounded, convex set of probability mass functions. Since any conditional probability mass function  $Q$  is in the interior of  $\text{dom}(\mathcal{U})$  and the graph of  $\mathcal{U}(x, a, P)$  is closed, by Theorem 2.7 in [91],  $\mathcal{U}(x, a, P)$  is a Lipschitz set-valued mapping with respect to the Hausdorff distance. Thus, for any  $\xi \in \mathcal{U}(x, a, P)$ , the following expression holds for some  $M_\xi > 0$ :

$$\inf_{\hat{q} = \hat{\xi}P: \hat{\xi} \in \mathcal{U}(x, a, \tilde{P})} \sum_{x' \in \mathcal{X}} |q(x') - \hat{q}(x')| \leq M_\xi \sum_{x' \in \mathcal{X}} |P(x'|x, a) - \tilde{P}(x'|x, a)|.$$

Next, we want to show that the infimum of the left side is attained. Since the objective function is convex, and  $\mathcal{U}(x, a, \tilde{P})$  is a convex compact set, there exists  $\tilde{\xi} \in \mathcal{U}(x, a, \tilde{P})$  such that infimum is attained.  $\square$

Now we turn to the main proof of Lemma 5.4.3. First, we want to show that

$$\alpha(a, \underline{d}') := D(x, a) + \gamma \rho(d'(x'))$$



is a Lipschitz function. Define

$$\{q^*(x')\}_{x' \in \mathcal{X}} \in \arg \max_{q \in \xi P: \xi \in \mathcal{U}(x, a, P(x'|x, a))} \left\{ D(x, a) + \gamma \sum_{x' \in \mathcal{X}} q(x') d'(x') \right\}.$$

Then, there exists a  $\tilde{\xi} \in \mathcal{U}(x, a, P(x'|x, \tilde{a}))$ ,  $\tilde{q} = \tilde{\xi} P$ , such that the following expressions hold:

$$\begin{aligned} & \alpha(a, \underline{d}') - \alpha(\tilde{a}, \underline{\tilde{d}}') \\ &= D(x, a) + \gamma \rho(d'(x')) - D(x, \tilde{a}) - \gamma \rho(\tilde{d}'(x')) \\ &\leq D(x, a) - D(x, \tilde{a}) + \gamma \sum_{x' \in \mathcal{X}} (q^*(x') - \tilde{q}(x')) d'(x') + \gamma \sum_{x' \in \mathcal{X}} \tilde{q}(x') (d'(x') - \tilde{d}'(x')) \\ &\leq |D(x, a) - D(x, \tilde{a})| + \gamma \sum_{x' \in \mathcal{X}} |d'(x') - \tilde{d}'(x')| + \gamma \max_{x \in \mathcal{X}} |d'(x)| \sum_{x' \in \mathcal{X}} |q^*(x') - \tilde{q}(x')|. \end{aligned} \quad (7.91)$$

The first equality follows from definitions of coherent risk measures. The first inequality is due to the representation theorem (Theorem 1.3.3) and the definition of  $q^* = \xi^* P$ ,  $\xi^* \in \mathcal{U}(x, a, P(x'|x, a))$ . The second inequality is due to the fact that  $\tilde{q}$  is a probability mass functions with  $\tilde{\xi} \in \mathcal{U}(x, a, P(x'|x, \tilde{a}))$ . Then, by Proposition 7.5.1, there exists  $M_\xi > 0$  such that

$$\sum_{x' \in \mathcal{X}} |q^*(x') - \tilde{q}(x')| \leq M_\xi \sum_{x' \in \mathcal{X}} |P(x'|x, a) - P(x'|x, \tilde{a})|.$$

Furthermore, by Assumptions (7.3.2) to (5.4.2) and the definition of  $\Phi(x')$ , expression (7.91) implies

$$\alpha(a, \underline{d}') - \alpha(\tilde{a}, \underline{\tilde{d}}') \leq M_A \left( |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - d'(x')| \right)$$

where

$$M_A = \max \{ M_D + \gamma M_P \bar{d} M_\xi, 1 \}.$$

By a symmetric argument, we can also show that

$$\alpha(\tilde{a}, \underline{\tilde{d}}') - \alpha(u, \underline{d}') \leq M_A \left( |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - d'(x')| \right).$$

Thus, by combining both arguments, we have shown that  $\alpha(a, \underline{d}')$  is a Lipschitz function. Next, for any  $(a, d') \in F(x, d)$ , where

$$F(x, d) = \left\{ (a, d') \mid u \in \mathcal{A}(x), d'(x') \in \Phi(x'), \forall x' \in \mathcal{X}, \alpha(a, \underline{d}') \leq d \right\},$$

consider the following optimization problem:

$$\mathcal{P}_{x,a,\underline{d}'}(d) = \inf_{(\tilde{a},\tilde{\underline{d}}') \in F(x,d)} |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - d'(x')|.$$

Since  $(a, \underline{d}')$  is in  $F(x, d)$ , it is a feasible solution to the above problem which yields  $\mathcal{P}_{x,a,\underline{d}'}(d) = 0$ . By our assumptions, both  $\mathcal{A}(x)$  and  $\Phi(x')$  are compact sets of real numbers. Note that both  $|\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - d'(x')|$  and  $\alpha(\tilde{a}, \tilde{\underline{d}}')$  are Lipschitz functions in  $(\tilde{a}, \tilde{\underline{d}}')$ . Also, consider the sub-gradient of  $f(\tilde{a}, \tilde{\underline{d}}', d) := \alpha(\tilde{a}, \tilde{\underline{d}}') - d^1$ :

$$\partial f(\tilde{a}, \tilde{\underline{d}}', d) = \bigcap_{(\hat{a}, \hat{\underline{d}}', \hat{d}) \in \text{dom}(f)} \left\{ \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix} \in \mathbb{R}^{|\mathcal{U}|} \times \mathbb{R}^{|\mathcal{X}|} \times \mathbb{R} : f(\hat{a}, \hat{\underline{d}}', \hat{d}) \geq f(\tilde{a}, \tilde{\underline{d}}', d) + \begin{bmatrix} g_1 \\ g_2 \\ g_3 \end{bmatrix}^T \left( \begin{bmatrix} \tilde{a} \\ \tilde{\underline{d}}' \\ d \end{bmatrix} - \begin{bmatrix} \hat{a} \\ \hat{\underline{d}}' \\ \hat{d} \end{bmatrix} \right) \right\}.$$

For any  $(g_1, g_2, g_3) \in \partial f(u, \underline{d}', d)$ , this implies

$$\alpha(\tilde{a}, \tilde{\underline{d}}') - \alpha(\hat{a}, \hat{\underline{d}}') \geq (g_3 + 1)(d - \hat{d}) + g_1(\tilde{a} - \hat{a}) + g_2^T(\tilde{\underline{d}}' - \hat{\underline{d}}'),$$

$\forall (\hat{a}, \hat{\underline{d}}', \hat{d}) \in \text{dom} f$ . Suppose  $g_3 > -1$ , then there exists  $\epsilon > 0$  such that  $g_3 + 1 = \epsilon$ . Also, by the Lipschitz-ness of  $\alpha(\tilde{a}, \tilde{\underline{d}}')$  and Cauchy Schwarz inequality, we get

$$(M_A + |g_1|)|\tilde{a} - \hat{a}| + (1 + \|g_2\|) \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - \hat{d}'(x')| \geq \epsilon(d - \hat{d}), \quad \forall (\hat{a}, \hat{\underline{d}}', \hat{d}) \in \text{dom}(f)$$

Since  $\tilde{a}, \hat{a}$  are finite and  $\tilde{\underline{d}}', \hat{\underline{d}}'$  are bounded, the above inequality fails if  $\hat{d} \rightarrow -\infty$ . This yields a contradiction. Similarly, by considering  $\hat{d} \rightarrow \infty$ , we can also arrive at a contradiction for the case of  $g_3 < -1$ . Therefore, the set of the third element of  $\partial f(\tilde{a}, \tilde{\underline{d}}', r)$  is a singleton and it equals to  $\{-1\}$ .

Since  $\alpha(\tilde{a}, \tilde{\underline{d}}') - d$  is differentiable on  $r$ , the third element of  $\partial f(\tilde{a}, \tilde{\underline{d}}', r)$  is a singleton and it equals to  $\{-1\}$ . Next, consider the sub-gradient of  $h(\tilde{a}, \tilde{\underline{d}}', r) = |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - d'(x')|$ . By identical arguments, we can show that the set of the third element of  $\partial h(\tilde{a}, \tilde{\underline{d}}', r)$  is a singleton and it equals to  $\{0\}$ . Therefore, Theorem 4.2 in [78] implies  $\mathcal{P}_{x,a,\underline{d}'}(d)$  is strictly differentiable (Lipschitz continuous) in  $r$ <sup>2</sup>. Then, for any  $(a, \underline{d}') \in F(x, d)$ , there exists  $M_R > 0$  such that

$$\inf_{(\tilde{a}, \tilde{\underline{d}}') \in F(x, \tilde{d})} |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - \tilde{d}'(x')| \leq M_R |\tilde{d} - d|.$$

Finally we want to show that the infimum on the left side of the above expression is attained. First,  $|\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |\tilde{d}'(x') - \tilde{d}'(x')|$  is coercive and continuous in  $(\tilde{a}, \tilde{\underline{d}}')$ . By Example 14.29 in [111], this

<sup>1</sup> A sub-gradient of a function  $f : \mathcal{X} \rightarrow \mathbb{R}$  at a point  $x_0 \in \mathcal{X}$  is a real vector  $g$  such that for all  $x \in \mathcal{X}$ ,  $f(x) - f(x_0) \geq g^T(x - x_0)$ ,  $\forall x \in \mathcal{X}$ .

<sup>2</sup> Theorem 4.2 in [78] implies both  $\partial \mathcal{P}_{x,a,\underline{d}'}(d), \partial^\infty \mathcal{P}_{x,a,\underline{d}'}(d) \subseteq \{0\}$  for  $d \in \Phi_k(d)$ . This result further implies  $\mathcal{P}_{x,a,\underline{d}'}(d)$  is strictly differentiable. For details, please refer to this paper.

function is a Caratheodory integrand and is also a normal integrand. Furthermore, since  $F(x, \tilde{d})$  is a closed set (since  $\mathcal{A}(x)$  is a finite set,  $\Phi(x')$  is a compact set and the constraint inequality is non-strict) and  $\alpha(\tilde{a}, \tilde{d}') - \tilde{d}$  is a normal integrand (see the proof of Theorem 5.3.6), by Theorem 14.36 and Example 14.32 in [111], one can show that the following indicator function:

$$\mathbb{I}_x(\tilde{a}, \tilde{d}', \tilde{d}) := \begin{cases} 0 & \text{if } (\tilde{a}, \tilde{d}') \in F(x, \tilde{d}) \\ \infty & \text{otherwise} \end{cases}$$

is a normal integrand. Furthermore, by Proposition 14.44 in [111], the function

$$g_x(\tilde{a}, \tilde{d}', \tilde{d}) := |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |d'(x') - \tilde{d}'(x')| + \mathbb{I}_x(\tilde{a}, \tilde{d}', \tilde{d})$$

is a normal integrand. Also,  $\inf_{\tilde{a}} g_x(\tilde{a}, \tilde{d}', \tilde{d}) = \inf_{(\tilde{a}, \tilde{d}') \in F(x, \tilde{d})} |\tilde{a} - a| + \sum_{x' \in \mathcal{X}} |d'(x') - \tilde{d}'(x')|$ . By Theorem 14.37 in [111], there exists  $(\hat{a}, \hat{d}') \in F(x, \tilde{d})$  such that  $(\hat{a}, \hat{d}') \operatorname{argmin} g_x(\tilde{a}, \tilde{d}', \tilde{d})$ . Furthermore, the right side of the above equality is finite since  $F(x, \tilde{d})$  is a non-empty set. The definition of  $\mathbb{I}_x(\hat{a}, \hat{d}', \tilde{d})$  implies that  $(\hat{a}, \hat{d}') \in F(x, \tilde{d})$ . Therefore this implies expression (5.6) holds for any  $(a, \underline{d}') \in F(x, d)$ .

### 7.5.4 Proof of Theorem 5.4.4

The proof of the main result of this paper relies on three technical lemmas. The first lemma provides a sensitivity result for the value function  $V^*(x, d)$ .

**Lemma 7.5.2.** *Suppose that  $F(x, d)$  and  $F(x, \tilde{d})$  are non-empty sets for  $x \in \mathcal{X}$  and  $d, \tilde{d} \in \Phi(x)$ . Then, the following expression holds:*

$$0 \leq V^*(x, \tilde{d}) - V^*(x, d) \leq \underbrace{\left( \frac{M_C}{1 - \gamma} + \frac{M_P C_{\max}}{(1 - \gamma)^2} \right)}_{M_V} M_R(d - \tilde{d}), \quad (7.92)$$

where  $M_R$  is the constant defined in inequality (5.6).

*Proof.* First, when  $\tilde{d} \leq d$ , by the definition of the value function in problem  $\mathcal{OPT}_{\text{RC}}$ , we know that  $V^*(x, \tilde{d}) \geq V^*(x, d)$ . The proof is completed if we can show that for  $\tilde{d} \leq d$ ,

$$V^*(x, \tilde{d}) - V^*(x, d) \leq M_V M_R(d - \tilde{d}).$$

Now let

$$V_0(x, d) = \frac{C_{\max}}{1 - \gamma}$$

be the initial value function estimate and the sequence of estimate is updated by value iteration, i.e.,

$$V_{k+1}(x, d) = \mathbf{T}[V_k](x, d).$$

From this update sequence one can immediately show that  $\|V_k\|_\infty \leq C_{\max}/(1 - \gamma)$  for every  $k$ .

Furthermore for any given  $x \in \mathcal{X}$ ,  $d \in \Phi(x)$ , let  $(a^*, d^{*,'})$  be the minimizer of  $\mathbf{T}[V_k](x, d)$ . For notional convenience here we omit the dependency of  $k$  in the set of minimizers  $(a^*, d^{*,'})$ . Then, there exists  $(\hat{a}, \hat{d}') \in F(x, \tilde{d})$ , such that inequality (5.6) and the following expressions hold:

$$\begin{aligned}
& V_{k+1}(x, \tilde{d}) - V_{k+1}(x, d) \\
& \leq C(x, \hat{a}) - C(x, a^*) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, \hat{a}) V_k(x', \hat{d}'(x')) - \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a^*) V_k(x', d^{*,'}(x')) \\
& = C(x, \hat{a}) - C(x, a^*) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, \hat{a}) \left( V_k(x', \hat{d}'(x')) - V_k(x', d^{*,'}(x')) \right) \\
& \quad + \gamma \sum_{x' \in \mathcal{X}} (P(x'|x, \hat{a}) - P(x'|x, a^*)) V_k(x', d^{*,'}(x')) \\
& \leq \gamma \|V_k\|_\infty \sum_{x' \in \mathcal{X}} |P(x'|x, \hat{a}) - P(x'|x, a^*)| + \gamma \sum_{x' \in \mathcal{X}} \left\{ \left| V_k(x', d^{*,'}(x')) - V_k(x', \hat{d}'(x')) \right| \right\} \\
& \quad + |C(x, \hat{a}) - C(x, a^*)|.
\end{aligned}$$

The second inequality follows from  $\sum_{x' \in \mathcal{X}} P(x'|x, \hat{a}) = 1$  and the definition of  $\|V_k\|_\infty \leq C_{\max}/(1 - \gamma)$ . Now consider the following sequence of constants

$$M_{V,k} = \frac{1 - \gamma^k}{1 - \gamma} (M_C + M_P C_{\max}/(1 - \gamma)).$$

Obviously, at  $k = 0$ , using the results in Assumption 7.3.2 and Lemma 5.4.3, the above expression implies

$$V_1(x, \tilde{d}) - V_1(x, d) \leq (M_C + M_P C_{\max}/(1 - \gamma)) M_R |\tilde{d} - d|.$$

Now at  $k = j$  by induction we assume

$$V_j(x, \tilde{d}) - V_j(x, d) \leq \underbrace{\frac{1 - \gamma^j}{1 - \gamma} \left( M_C + M_P \frac{C_{\max}}{1 - \gamma} \right)}_{M_{V,j}} M_R |\tilde{d} - d|.$$

Equipped with the induction's assumption, at  $k = j + 1$  the above expression further implies

$$\begin{aligned}
V_{j+1}(x, \tilde{d}) - V_{j+1}(x, d) & \leq (M_C + M_P \|V_j\|_\infty) |\hat{a} - a^*| + \gamma M_{V,j} \sum_{x' \in \mathcal{X}} |\hat{d}'(x') - d^{*,'}(x')| \\
& \leq (M_C + M_P C_{\max}/(1 - \gamma) + \gamma M_{V,j}) M_R |\tilde{d} - d| \\
& \leq M_{V,j+1} M_R |\tilde{d} - d|.
\end{aligned} \tag{7.93}$$

Thus by induction we have that  $V_k(x, \tilde{d}) - V_k(x, d) \leq M_{V,k} M_R |\tilde{d} - d|$ . Note that, due to the contraction property of  $\mathbf{T}$ , the sequence of value function estimates  $\{V_k\}$  converges to  $V^*$ . Finally combining this

property with the fact that  $\lim_{k \rightarrow \infty} M_{V,k} = M_C/(1 - \gamma) + M_P C_{\max}/(1 - \gamma)^2$ , one immediately shows that expression (7.92) holds.  $\square$

The second lemma shows that the “difference” between the dynamic programming operators  $\mathcal{T}_{\mathcal{D}}[V^*](x, d)$  and  $\mathbf{T}[V^*](x, d)$  is bounded.

**Lemma 7.5.3.** *For any  $x \in \mathcal{X}$  and  $d \in \Phi(x)$ , the following inequality holds:*

$$0 \leq \mathcal{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \leq \gamma M_V M_R \Delta$$

where  $M_V$  is given in Lemma 7.5.2,  $M_R$  is the constant defined in inequality (5.6), and  $\Delta$  is the step size for the discretization of the threshold state  $d$ .

*Proof.* First, by the definition of  $F_{\mathcal{D}}(x, d)$ , we know that  $F_{\mathcal{D}}(x, d) \subseteq F(x, d)$ . Since, the objective functions and all other constraints in  $\mathcal{T}_{\mathcal{D}}[V^*](x, d)$  and  $\mathbf{T}[V^*](x, d)$  are identical, we can easily conclude that  $\mathcal{T}_{\mathcal{D}}[V^*](x, d) \geq \mathbf{T}[V^*](x, d)$  for all  $x \in \mathcal{X}$ ,  $d \in \Phi(x)$ . The proof is completed if we can show

$$\mathcal{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \leq \gamma M_V M_R \Delta.$$

By Theorem 5.3.6 we know that the infimum of  $\mathbf{T}[V^*](x, d)$  is attained. Let  $(a^*, d^{*,'}) \in F(x, d)$  be the minimizer of  $\mathbf{T}[V^*](x, d)$ . Also, for every fixed  $x' \in \mathcal{X}$ , let  $\theta(x') \in \{0, \dots, \Theta\}$  such that  $d^{*,'}(x') \in \Phi^{(\theta(x'))}(x')$ . Now, construct

$$\tilde{d}'(x') := d^{(\theta(x'))} \in \Phi^{(\theta(x'))}(x').$$

By definition of  $\bar{\Phi}(x')$ , we know that  $\tilde{d}'(x') \in \bar{\Phi}(x')$ ,  $\forall x' \in \mathcal{X}$ . Since  $d^{(\theta(x'))}$  is the lower bound of  $\Phi^{(\theta(x'))}(x')$ , we have  $d^{(\theta(x'))} \leq d^{*,'}(x')$ . Furthermore, since the size of  $\Phi^{(\theta(x'))}(x')$  is  $\Delta$ , we know that  $|d^{(\theta(x'))} - d^{*,'}(x')| \leq \Delta$  for any  $x' \in \mathcal{X}$ . By monotonicity of coherent risk measures,

$$D(x, a^*) + \gamma \rho(\tilde{d}'(x_{k+1})) \leq D(x, a^*) + \gamma \rho(d^{*,'}(x_{k+1})) \leq d.$$

Therefore, we conclude that  $(a^*, \tilde{d}') \in F_{\mathcal{D}}(x, d)$  is a feasible solution to the problem in  $\mathcal{T}_{\mathcal{D}}[V^*](x, d)$ . From this fact, we get the following inequalities:

$$\begin{aligned} \mathcal{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d) &\leq \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a^*) \left( V^*(x', \tilde{d}'(x')) - V^*(x', d^{*,'}(x')) \right) \\ &\leq \gamma \sup_{x' \in \mathcal{X}} \left\{ \left| V^*(x', \tilde{d}'(x')) - V^*(x', d^{*,'}(x')) \right| \right\} \\ &\leq \gamma M_V M_R \sup_{x' \in \mathcal{X}} |\tilde{d}'(x') - d^{*,'}(x')| \leq \gamma M_V M_R \Delta. \end{aligned}$$

The first inequality is due to substitutions of the feasible solution of  $\mathcal{T}_{\mathcal{D}}[V^*](x, d)$  and the optimal solution of  $\mathbf{T}[V^*](x, d)$ . The second inequality is trivial. The third inequality is a result of Lemma 7.5.2 and the fourth

inequality is due to the definition of  $\tilde{d}'(x')$ , for all  $x' \in \mathcal{X}$ . This completes the proof.  $\square$

The third lemma characterizes the error bound between the dynamic programming operators  $\mathbf{T}[V^*](x, d)$  and  $\mathbf{T}_{\mathcal{D}}[V^*](x, d)$ .

**Lemma 7.5.4.** *Suppose Assumptions (7.3.2) to (5.4.2) hold. Then,*

$$\|\mathbf{T}_{\mathcal{D}}[V^*] - \mathbf{T}[V^*]\|_{\infty} \leq (1 + \gamma)M_V M_R \Delta, \quad (7.94)$$

where  $\mathbf{T}_{\mathcal{D}}[V^*](x, d)$  is defined in equation (5.4),  $\Delta$  is the discretization step size,  $M_V$  is given in Lemma 7.5.2 and  $M_R$  is the constant defined in inequality (5.6).

*Proof.* For any given  $x \in \mathcal{X}$  and  $d \in \Phi(x)$ , let  $\theta \in \{0, \dots, \Theta\}$  such that  $d \in \Phi^{(\theta)}(x)$ . Then, by definition of  $\mathbf{T}_{\mathcal{D}}[V^*](x, d)$  and Theorem 5.3.2, the following expression holds:

$$|\mathbf{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d)| \leq |V^*(x, d^{(\theta)}) - V^*(x, d)| + |\mathcal{T}_{\mathcal{D}}[V^*](x, d^{(\theta)}) - \mathbf{T}[V^*](x, d^{(\theta)})|.$$

By Lemma 7.5.2 and 7.5.3, the above equation implies that

$$|\mathbf{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d)| \leq M_V M_R \Delta + \gamma M_V M_R |d - d^{(\theta)}| \leq (1 + \gamma)M_V M_R \Delta.$$

The last inequality follows from the fact that  $d \in \Phi^{(\theta)}(x)$  implies  $|d^{(\theta)} - d| \leq \Delta$ , where  $d^{(\theta)}$  is the lower bound for the discretized region of risk threshold  $\Phi^{(\theta)}(x)$ . By taking the supremum with respect to  $x \in \mathcal{X}$  and  $d \in \Phi(x)$  on both sides of the above inequality, the proof is completed.  $\square$

Now we turn to the main proof of Theorem 5.4.4. First from Lemma 7.5.4, we have that  $\|\mathbf{T}_{\mathcal{D}}[V^*] - \mathbf{T}[V^*]\|_{\infty} \leq (1 + \gamma)M_V M_R \Delta$ . This implies the following chain of inequalities:

$$\begin{aligned} \|V_{\mathcal{D}}^* - V^*\|_{\infty} &= \|\mathbf{T}_{\mathcal{D}}[V_{\mathcal{D}}^*] - \mathbf{T}[V^*]\|_{\infty} \leq \|\mathbf{T}_{\mathcal{D}}[V_{\mathcal{D}}^*] - \mathbf{T}_{\mathcal{D}}[V^*]\|_{\infty} + \|\mathbf{T}_{\mathcal{D}}[V^*] - \mathbf{T}[V^*]\|_{\infty} \\ &\leq \gamma \|V_{\mathcal{D}}^* - V^*\|_{\infty} + (1 + \gamma)M_V M_R \Delta. \end{aligned}$$

The first equality is due to Theorem 5.3.2 and the fact that  $V_{\mathcal{D}}^*(x, d) = \mathbf{T}_{\mathcal{D}}[V_{\mathcal{D}}^*](x, d)$ . The second inequality follows from triangular inequality and the third inequality follows from the contraction property in Lemma 5.3.1 and the arguments in Lemma 7.5.4.

Then one concludes the proof of this theorem by having the following inequality:

$$\|V_{\mathcal{D}}^* - V^*\|_{\infty} \leq \frac{1 + \gamma}{1 - \gamma} M_V M_R \Delta,$$

where  $M_V$  is defined in Lemma 7.5.2, with its expression is given by:

$$M_V = \left( \frac{M_C}{1 - \gamma} + \frac{M_P C_{\max}}{(1 - \gamma)^2} \right),$$

and  $M_R$  is the constant defined in inequality (5.6).

### 7.5.5 Proof of Theorem 5.4.7

The error bound analysis of the interpolated Bellman iteration is similar to the analysis of the discretized Bellman iteration described in Theorem 5.4.4. Analogous to Lemma 7.5.3, we have the following technical lemma.

**Lemma 7.5.5.** *For any  $x \in \mathcal{X}$  and  $d \in \Phi(x)$ , the following inequality holds:*

$$-\gamma M_V M_R \Delta \leq \mathbf{T}_{\mathcal{I}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \leq \gamma M_V M_R \Delta$$

where  $M_V$  is given in Lemma 7.5.2,  $M_R$  is the constant defined in inequality (5.6), and  $\Delta$  is the step size for the discretization of the threshold state  $d$ .

*Proof.* The proof of this lemma is split into two parts. First we want to show that

$$\mathbf{T}_{\mathcal{I}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \leq \gamma M_V M_R \Delta. \quad (7.95)$$

Before getting into the analysis directly, a crucial intermediate step is to derive the following inequality for any function  $V : B(\mathcal{X} \times \mathbb{R}) \rightarrow B(\mathcal{X} \times \mathbb{R})$ :

$$\mathbf{T}_{\mathcal{I}}[V](x, d) \leq \mathcal{T}_{\mathcal{D}}[V](x, d), \quad \forall x \in \mathcal{X}, \quad d \in \mathbb{R}. \quad (7.96)$$

Obviously an equivalent re-formulation of optimization problem  $\mathcal{T}_{\mathcal{D}}[V](x, d)$  is given by:

$$\begin{aligned} & \min_{a, d'} C(x, a) + \gamma \sum_{x' \in \mathcal{X}} P(x'|x, a) \mathcal{I}_{x'}[V](d) \\ & \text{subject to } a \in \mathcal{A}(x), d'(x') \in \bar{\Phi}(x') \text{ for all } x' \in \mathcal{X}, \text{ and } D(x, a) + \gamma \rho(d'(x')) \leq d. \end{aligned}$$

Using this relationship, at any state  $x \in \mathcal{X}$  and constraint threshold  $d \in \mathbb{R}$ , the optimizers of problem  $\mathcal{T}_{\mathcal{D}}[V](x, d)$  are indeed feasible solutions to problem  $\mathbf{T}_{\mathcal{I}}[V](x, d)$ , with integer-valued constraints. This further implies that inequality (7.96) holds. From this fact, we get the following inequalities:

$$\begin{aligned} \mathbf{T}_{\mathcal{I}}[V^*](x, d) - \mathbf{T}[V^*](x, d) & \leq \mathcal{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \\ & \leq \mathbf{T}_{\mathcal{D}}[V^*](x, d) - \mathbf{T}[V^*](x, d) \\ & \leq \gamma \sup_{x' \in \mathcal{X}} \left\{ \left| V^*(x', \tilde{d}'(x')) - V^*(x', d^{*'}(x')) \right| \right\} \leq \gamma M_V M_R \Delta. \end{aligned}$$

The second inequality follows from the non-increasing property of value function  $V^*$  in  $d$  and the definition of  $\mathbf{T}_{\mathcal{D}}$ . The third inequality and fourth inequality follow from the same lines of arguments in Lemma 7.5.3.

On the other hand, by using analogous arguments, we can show that

$$\begin{aligned} \mathbf{T}[V^*](x, d) - \mathbf{T}_{\mathcal{I}}[V^*](x, d) &\geq \mathbf{T}[V^*](x, d) - \mathbf{T}_{\mathcal{D}}[V^*](x, d) \\ &\geq -\gamma \sup_{x' \in \mathcal{X}} \left\{ \left| V^*(x', \tilde{d}'(x')) - V^*(x', d^{*,'}(x')) \right| \right\} \geq -\gamma M_V M_R \Delta. \end{aligned} \quad (7.97)$$

Therefore the proof of this lemma is completed by combining both inequality (7.95) and (7.97).  $\square$

Equipped with this result, the rest of the error bound proof follows identical arguments from the proof of Theorem 5.4.4 (and Lemma 7.5.4), with value function estimate  $V_{\mathcal{D}}^*$  replaced by  $V_{\mathcal{I}}^*$  and Bellman operator  $\mathbf{T}_{\mathcal{D}}$  replaced by  $\mathbf{T}_{\mathcal{I}}$ . Details of these steps will be omitted for the sake of brevity.

On the other hand, we now show the claim  $V_{\mathcal{I}}^*(x, d) \leq V_{\mathcal{D}}^*(x, d)$ . Recall from inequality (7.96) and the definition of discretized Bellman operator  $\mathbf{T}_{\mathcal{D}}$  that  $\mathbf{T}_{\mathcal{I}}[V](x, d) \leq \mathcal{T}_{\mathcal{D}}[V](x, d) \leq \mathbf{T}_{\mathcal{D}}[V](x, d)$  for any function  $V : B(\mathcal{X} \times \mathbb{R}) \rightarrow B(\mathcal{X} \times \mathbb{R})$ . By putting  $V = V_{\mathcal{D}}^*$  and applying  $\mathbf{T}_{\mathcal{I}}$  on both sides, we have that

$$\mathbf{T}_{\mathcal{I}}^2[V_{\mathcal{D}}^*](x, d) \leq \mathbf{T}_{\mathcal{I}}[V_{\mathcal{D}}^*](x, d) \leq V_{\mathcal{D}}^*(x, d)$$

Repeating this procedure, and noticing that  $\lim_{N \rightarrow \infty} \mathbf{T}_{\mathcal{I}}^N[V_{\mathcal{D}}^*](x, d) = V_{\mathcal{I}}^*(x, d)$ , the inequality  $V_{\mathcal{I}}^*(x, d) \leq V_{\mathcal{D}}^*(x, d)$  is concluded.

Combining all the above arguments, the proof of this theorem is completed.



# Bibliography

- [1] B. Acciaio, H. Föllmer, and I. Penner. *Dynamic convex risk measures*, chapter 1, pages 1–34. Springer-Verlag, 2011.
- [2] B. Açıkmese and L. Blackmore. Lossless Convexification of a Class of Optimal Control Problems with Non-convex Control Constraints. *Automatica*, 47(2):341–347, February 2011.
- [3] G. Alexander. From Markowitz to modern risk management. *The European Journal of Finance*, 15(5-6):451–461, 2009.
- [4] E. Altman. *Constrained Markov decision processes*. CRC Press, 1999.
- [5] E. Altman, K. Avrachenkov, and R. Núñez Queija. Perturbation analysis for denumerable Markov chains with application to queueing models. *Advances in Applied Probability*, pages 839–853, 2004.
- [6] P. Apkarian and P. Gahinet. A convex characterization of gain-scheduled  $H_\infty$  controllers. *IEEE Transactions on Automatic Control*, 40(5):853–864, 1995.
- [7] P. Artzner, F. Delbaen, J. Eber, and D. Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1998.
- [8] O. Bardou, N. Frikha, et al. Computing VaR and CVaR using stochastic approximation and adaptive unconstrained importance sampling. *Monte Carlo Methods and Applications*, 15(3):173–210, 2009.
- [9] N. Bäuerle and A. Mundt. Dynamic mean-risk optimization in a binomial model. *Mathematics of Operations Research*, 70(2):219–239, 2009.
- [10] N. Bäuerle and J. Ott. Markov decision processes with average-value-at-risk criteria. *Mathematics of Operations Research*, 74(3):361–379, 2011.
- [11] J. Baxter and P. Bartlett. Infinite-Horizon Policy-Gradient Estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.
- [12] A. Ben Tal, L. El Ghaoui, and A. Nemirovski. *Robust optimization*. Princeton Series of Applied Mathematics, 2008.

- [13] A. Ben Tal and M. Teboulle. An old-new concept of convex risk measure: The optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476, 2007.
- [14] M. Benaïm, J. Hofbauer, and S. Sorin. Stochastic approximations and differential inclusions, Part II: Applications. *Mathematics of Operations Research*, 31(4):673–695, 2006.
- [15] D. Bernardini and A. Bemporad. Stabilizing model predictive control of stochastic constrained linear systems. *IEEE Transactions on Automatic Control*, 57(6):1468–1480, 2012.
- [16] D. Bertsekas. *Nonlinear programming*. Athena Scientific, 1999.
- [17] D. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2005.
- [18] D. Bertsekas. Min common/max crossing duality: A geometric view of conjugacy in convex optimization. *Lab. for Information and Decision Systems, MIT, Tech. Rep. Report LIDS-P-2796*, 2009.
- [19] D. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.
- [20] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, 1997.
- [21] D. Bertsimas and D. Brown. Constructing uncertainty sets for robust linear optimization. *Operations Research*, 57(6):1483–1495, 2009.
- [22] L. F. Bertuccelli, N. Pellegrino, and M. L. Cummings. Choice Modeling of Relook Tasks for UAV Search Missions. In *American Control Conference*, pages 2410–2415, Baltimore, MD, June 2010.
- [23] M. Best and R. Grauer. On the Sensitivity of Mean-variance-efficient Portfolios to Changes in Asset Means: Some Analytical and Computational Results. *Review of Financial Studies*, 4(2):315–342, 1991.
- [24] S. Bhatnagar. An actor-critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters*, 59(12):760–766, 2010.
- [25] S. Bhatnagar and K. Lakshmanan. An Online Actor-Critic Algorithm with Function Approximation for Constrained Markov Decision Processes. *Journal of Optimization Theory and Applications*, pages 1–21, 2012.
- [26] S. Bhatnagar, H. Prasad, and L. Prashanth. *Stochastic recursive algorithms for optimization*, volume 434. Springer, 2013.
- [27] S. Bhatnagar, R. Sutton, M. Ghavamzadeh, and M. Lee. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

- [28] M. Bichi, G. Ripaccioli, S. Di Cairano, D. Bernardini, A. Bemporad, and I. Kolmanovsky. Stochastic model predictive control with driver behavior learning for improved powertrain control. In *Proc. IEEE Conf. on Decision and Control*, pages 6077–6082. IEEE, 2010.
- [29] L. Blackmore, M. Ono, A. Bektassov, and B. C. Williams. A probabilistic particle-control approximation of chance-constrained stochastic predictive control. *IEEE Transactions on Robotics*, 26(3):502–517, 2010.
- [30] K. Boda and J. Filar. Time Consistent Dynamic Risk Measures. *Mathematics of Operations Research*, 63(1):169–186, 2006.
- [31] K. Boda, J. Filar, Y. Lin, and L. Spanjers. Stochastic target hitting time and the problem of early retirement. *IEEE Transactions on Automatic Control*, 49(3):409–419, 2004.
- [32] V. Borkar. A Sensitivity Formula for the Risk-sensitive Cost and the Actor-Critic Algorithm. *Systems & Control Letters*, 44:339–346, 2001.
- [33] V. Borkar.  $Q$ -learning for Risk-sensitive Control. *Mathematics of Operations Research*, 27:294–311, 2002.
- [34] V. Borkar. An actor-critic algorithm for constrained Markov decision processes. *Systems & Control Letters*, 54(3):207–213, 2005.
- [35] V. Borkar. *Stochastic approximation: A dynamical systems viewpoint*. Cambridge University Press, 2008.
- [36] V. Borkar and R. Jain. Risk-Constrained Markov Decision Processes. *IEEE Transactions on Automatic Control*, 59(9):2574 – 2579, 2014.
- [37] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- [38] M. Cannon, B. Kouvaritakis, and X. Wu. Model predictive control for systems with stochastic multiplicative uncertainty and probabilistic constraints. *Automatica*, 45(1):167–172, 2009.
- [39] M. Cannon, B. Kouvaritakis, and X. Wu. Probabilistic constrained MPC for multiplicative and additive stochastic uncertainty. *IEEE Transactions on Automatic Control*, 54(7):1626–1632, 2009.
- [40] R. Chen. Constrained stochastic control with probabilistic criteria and search optimization. In *Proc. IEEE Conf. on Decision and Control*, 2004.
- [41] R. Chen and E. Feinberg. Non-randomized policies for constrained Markov decision process. *Mathematical Methods in Operations Research*, 66:165–179, 2007.
- [42] P. Cheridito and M. Kupper. Composition of time consistent dynamic monetary risk measures in discrete time. *International Journal of Theoretical and Applied Finance*, 14(1):137–162, 2011.

- [43] P. Cheridito and M. Stadjie. Time inconsistency of VaR and time-consistent alternatives. *Finance Research Letters*, 6(1):40–46, 2009.
- [44] M. Chilali and P. Gahinet.  $H_\infty$  design with pole placement constraints: an LMI approach. *IEEE Transactions on Automatic Control*, 41(3):358–367, 1996.
- [45] C. Chow and J. Tsitsiklis. An optimal one-way multigrid algorithm for discrete-time stochastic control. *IEEE Transactions on Automatic Control*, 36(8):898–914, 1991.
- [46] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.
- [47] Y. Chow and M. Pavone. A Uniform-Grid Discretization Algorithm for Stochastic Optimal Control with Risk Constraints. In *Proc. IEEE Conf. on Decision and Control*, pages 2465–2470, 2013.
- [48] Y. Chow and M. Pavone. Stochastic Optimal Control with Dynamic, Time-Consistent Risk Constraints. In *American Control Conference*, pages 390–395, 2013.
- [49] E. Collins. Using Markov decision processes to optimize a nonlinear functional of the final distribution, with manufacturing applications. In *Stochastic Modelling in Innovative Manufacturing*, pages 30–45. Springer, 1997.
- [50] C. De Souza. Robust stability and stabilization of uncertain discrete-time Markovian jump linear systems. *IEEE Transactions on Automatic Control*, 51(5):836–841, 2006.
- [51] B. Derfer, N. Goodyear, K. Hung, C. Matthews, G. Paoni, K. Rollins, R. Rose, M. Seaman, and J. Wiles. Online marketing platform, August 17 2007. US Patent App. 11/893,765.
- [52] K. Dowd. *Measuring market risk*. John Wiley & Sons, 2007.
- [53] A. Eichhorn and W. Römisch. Polyhedral risk measures in stochastic programming. *SIAM Journal on Optimization*, 16(1):69–95, 2005.
- [54] J. Filar, L. Kallenberg, and H. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.
- [55] J. Filar, D. Krass, and K. Ross. Percentile Performance Criteria for limiting Average Markov Decision Processes. *IEEE Transactions on Automatic Control*, 40(1):2–10, 1995.
- [56] K. Fisher and M. Statman. The Mean-variance-Optimization Puzzle: Security Portfolios and Food Portfolios. *Financial Analysts Journal*, pages 41–50, 1997.
- [57] E. Frazzoli, M. Dahleh, and E. Feron. Real-time motion planning for agile autonomous vehicles. *AIAA Journal of Guidance, Control, and Dynamics*, 25(1):116–129, 2002.

- [58] G. Gordon. Approximate solutions to Markov decision processes. *Robotics Institute*, page 228, 1999.
- [59] R. Green and B. Hollifield. When Will Mean-Variance Efficient Portfolios Be Well Diversified? *The Journal of Finance*, 47(5):1785–1809, 1992.
- [60] W. Haskell and R. Jain. A convex analytic approach to risk-aware Markov Decision Processes. *SIAM Journal of Control and Optimization*, 2014.
- [61] R. Howard and J. Matheson. Risk-Sensitive Markov Decision Processes. *Management Science*, 18(7):356–369, 1972.
- [62] D. Iancu, M. Petrik, and D. Subramanian. Tight Approximations of Dynamic Risk Measures. *arXiv preprint arXiv:1106.6102*, 2013.
- [63] G. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.
- [64] G. Iyengar and A. Ma. Fast gradient descent method for mean-CVaR optimization. *Annals of Operations Research*, 205(1):203–212, 2013.
- [65] T. Jaksch, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.
- [66] S. Kakade and J. Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pages 267–274, 2002.
- [67] M. Kearns and S. Singh. Finite-sample convergence rates for  $Q$ -learning and indirect algorithms. *Advances in Neural Information Processing Systems*, pages 996–1002, 1999.
- [68] H. K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002.
- [69] V. Konda and J. Tsitsiklis. Actor-Critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.
- [70] G. Konidaris, S. Osentoski, and P. Thomas. Value Function Approximation in Reinforcement Learning Using the Fourier Basis. In *AAAI*, 2011.
- [71] Y. Korilis and A. Lazar. On the existence of equilibria in noncooperative optimal flow control. *Journal of the Association for Computing Machinery*, 42(3):584–613, May 1995.
- [72] M. Kothare, V. Balakrishnan, and M. Morari. Robust constrained model predictive control using linear matrix inequalities. *Automatica*, 32(10):1361–1379, 1996.
- [73] H. Kushner and G. Yin. *Stochastic approximation algorithms and applications*. Springer, 1997.

- [74] Y. Kuwata, M. Pavone, and J. Balaram. A risk-constrained multi-stage decision making approach to the architectural analysis of Mars missions. In *Proc. IEEE Conf. on Decision and Control*, pages 2102–2109, 2012.
- [75] M. Kvasnica, P. Grieder, M. Baotić, and M. Morari. Multi-parametric toolbox (MPT). In *Hybrid Systems: Computation and Control*, pages 448–462. Springer, 2004.
- [76] C. Liang and H. Peng. Optimal adaptive cruise control with guaranteed string stability. *Vehicle System Dynamics*, 32(4-5):313–330, 1999.
- [77] J. Löfberg. YALMIP : A toolbox for modeling and optimization in MATLAB. In *IEEE International Symposium on Computer Aided Control Systems Design*, pages 284–289, 2004.
- [78] Y Lucet and J. Ye. Sensitivity analysis for the value function for optimization problems with variational inequalities constraints. *SIAM Journal on Optimization*, 40(3):699–723, 2002.
- [79] D. Mankowitz, A. Tamar, and S. Mannor. Situational Awareness by Risk-Conscious Skills. *arXiv preprint arXiv:1610.02847*, 2016.
- [80] S. Mannor, O. Mebel, and H. Xu. Lightning Does Not Strike Twice: Robust MDPs with Coupled Uncertainty. In *International Conference on Machine Learning*, pages 385–392, 2012.
- [81] S. Mannor, D. Simester, P. Sun, and J. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.
- [82] S. Mannor and J. N. Tsitsiklis. Mean-variance optimization in Markov decision processes. In *International Conference on Machine Learning*, 2011.
- [83] P. Marbach. *Simulated-based methods for Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 1998.
- [84] H. Markowitz. Portfolio selection. *Journal of Finance*, 7(1):77–91, 1952.
- [85] D. Mayne, J. Rawlings, C. Rao, and P. Scokaert. Constrained Model Predictive Control: Stability and Optimality. *Automatica*, 36(6):789–814, 2000.
- [86] P. Milgrom and I. Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, pages 583–601, 2002.
- [87] T. Moldovan and P. Abbeel. Risk aversion in Markov decision processes via near optimal Chernoff bounds. In *Advances in Neural Information Processing Systems*, pages 3140–3148, 2012.
- [88] T. Morimura, M. Sugiyama, M. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *International Conference on Machine Learning*, pages 799–806, 2010.

- [89] K. Murty. A problem in enumerating extreme points, and an efficient algorithm. *Optimization Letters*, 3(2):211–237, 2007.
- [90] A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, pages 663–670, 2000.
- [91] N. Nguyen and G. Lafferriere. Lipschitz properties of non-smooth functions and set-valued mappings via generalized differentiation and applications. *arXiv preprint arXiv:1302.1794*, 2013.
- [92] A. Nilm and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.
- [93] W. Ogryczak and A. Ruszczyński. From stochastic dominance to mean-risk models: Semi-deviations as risk measures. *European Journal of Operational Research*, 116(1):33–50, 1999.
- [94] M. Ono, M. Pavone, Y. Kuwata, and J. Balaram. Chance-Constrained Dynamic Programming with Application to Risk-Aware Robotic Space Exploration. *Autonomous Robots*, 39(4):555–571, 2015.
- [95] T. Osogami. Robustness and risk-sensitivity in Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 233–241, 2012.
- [96] J. Ott. *A Markov decision model for a surveillance application and risk-sensitive Markov decision processes*. PhD thesis, Karlsruhe Institute of Technology, 2010.
- [97] P. P. Nain and K. Ross. Optimal priority assignment with hard constraint. *IEEE Transactions on Automatic Control*, 31(10):883 – 888, October 1986.
- [98] B. Park and W. Kwon. Robust one-step receding horizon control of discrete-time Markovian jump uncertain systems. *Automatica*, 38(7):1229–1235, 2002.
- [99] J. Peters, S. Vijayakumar, and S. Schaal. Natural actor-critic. In *Proceedings of the Sixteenth European Conference on Machine Learning*, pages 280–291, 2005.
- [100] M. Petrik, M. Ghavamzadeh, and Y. Chow. Safe Policy Improvement by Minimizing Robust Baseline Regret. *Advances in Neural Information Processing Systems*, 2016.
- [101] M. Petrik and D. Subramanian. An approximate solution method for large risk-averse Markov decision processes. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2012.
- [102] G. Pflug and A. Pichler. Time consistent decisions and temporal decomposition of coherent risk functionals. *Optimization online*, 2015.
- [103] M. Phillips. *Interpolation and approximation by polynomials*, volume 14. Springer Science & Business Media, 2003.

- [104] A. Piunovskiy. Dynamic programming in constrained Markov decision process. *Control and Cybernetics*, 35(3):646–660, 2006.
- [105] L. Prashanth. Policy gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory*, pages 155–169. Springer, 2014.
- [106] L. Prashanth and M. Ghavamzadeh. Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, pages 252–260, 2013.
- [107] J. Primbs A and C. Sung. Stochastic receding horizon control of constrained linear systems with state and control multiplicative noise. *IEEE Transactions on Automatic Control*, 54(2):221–230, 2009.
- [108] J. Qin and T. Badgwell. A survey of industrial model predictive control technology. *Control Engineering Practice*, 11(7):733–764, 2003.
- [109] A. Rajeswaran, S. Ghotra, S. Levine, and B. Ravindran. EPOpt: Learning Robust Neural Network Policies using Model Ensembles. *arXiv preprint arXiv:1610.01283*, 2016.
- [110] J. Rawlings and D. Mayne. *Model predictive control: Theory and design*. Nob Hill Publishing, 2013.
- [111] R. Rockafellar. *Convex analysis*, volume 28. Princeton University Press, 1997.
- [112] R. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(21-41), 2000.
- [113] R. Rockafellar and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002.
- [114] R. Rockafellar, S. Uryasev, and M. Zabarankin. Generalized deviations in risk analysis. *Finance and Stochastic*, 10(1):51–74, 2006.
- [115] B. Roorda, J. Schumacher, and J. Engwerda. Coherent acceptability measures in multi-period models. *Mathematical Finance*, 15(4):589–612, 2005.
- [116] S. Ross et al. *Stochastic processes*, volume 2. John Wiley & Sons New York, 1996.
- [117] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, page 6, 2011.
- [118] B. Rudloff, A. Street, and D. Valladao. Time consistency and risk averse dynamic decision models: Interpretation and practical consequences. *Internal Research Reports*, 17, 2011.
- [119] A. Ruszczyński. Risk averse dynamic programming for Markov decision process. *Journal of Mathematical Programming*, 125(2):235–261, 2010.



- [120] A. Ruszczyński and A. Shapiro. Optimization of risk measures. Risk and Insurance 0407002, Econ-WPA, 2004.
- [121] A. Ruszczyński and A. Shapiro. Conditional risk mappings. *Mathematics of Operations Research*, 21(3):544–561, 2006.
- [122] A. Ruszczyński and A. Shapiro. Optimization of convex risk functions. *Mathematics of Operations Research*, 31(3):433–452, 2006.
- [123] L. Ryashko and H. Schurz. Mean square stability analysis of some linear stochastic systems. *Dynamic Systems and Applications*, 6:165–190, 1997.
- [124] C. Scherer. The general nonstrict algebraic Riccati inequality. *Linear Algebra and its Applications*, 219:1–33, 1995.
- [125] C. Scherer. Mixed  $H_2/H_\infty$  control for time-varying and linear parametrically-varying systems. *International Journal on Robust and Nonlinear Control*, 6(910):929–952, 1996.
- [126] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.
- [127] G. Serraino and S. Uryasev. Conditional Value-at-Risk (CVaR). In *Encyclopedia of Operations Research and Management Science*, pages 258–266. Springer, 2013.
- [128] G. Shani, D. Heckerman, and R. Brafman. An MDP-based recommender system. *Journal of Machine Learning Research*, 6(Sep):1265–1295, 2005.
- [129] A. Shapiro. On a time consistency concept in risk averse multi-stage stochastic programming. *Operations Research Letters*, 37(3):143–147, 2009.
- [130] A. Shapiro. Dynamic programming approach to adjustable robust optimization. *Operations Research Letters*, 39(2):83–87, 2010.
- [131] A. Shapiro. Minimax and risk averse multistage stochastic programming. *European Journal of Operational Research*, 219(3):719–726, 2012.
- [132] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: Modeling and theory*. SIAM, 2009.
- [133] A. Shapiro, W. Tekaya, J. da Costa, and M. Soares. Risk neutral and risk averse stochastic dual dynamic programming method. *European Journal of Operational Research*, 224(2):375–391, 2013.
- [134] T. Shardlow and A. Stuart. A perturbation theory for ergodic Markov chains and application to numerical approximations. *SIAM journal on numerical analysis*, 37(4):1120–1137, 2000.

- [135] R. Skelton, T. Iwasaki, and K. Grigoriadis. *A unified algebraic approach to linear control design*. CRC Press, 1998.
- [136] M. Sniedovich. A variance-constrained reservoir control problem. *Water Resources Research*, 16:271–274, 1980.
- [137] M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.
- [138] J. Spall. Multivariate Stochastic Approximation using a Simultaneous Perturbation Gradient Approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.
- [139] A. Strehl, L. Li, E. Wiewiora, J. Langford, and M. Littman. PAC model-free reinforcement learning. In *International Conference on Machine Learning*, pages 881–888. ACM, 2006.
- [140] R. Stubbs and S. Mehrotra. A branch-and-cut method for 0 – 1 mixed convex programming. *Mathematical programming*, 86(3):515–532, 1999.
- [141] L. Sun and L. Hong. Asymptotic representations for importance-sampling estimators of value-at-risk and conditional value-at-risk. *Operations Research Letters*, 38(4):246–251, 2010.
- [142] R. Sutton and A. Barto. *Introduction to reinforcement learning*. MIT Press, 1998.
- [143] R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.
- [144] Y. Le Tallec. *Robust, risk-sensitive, and data-driven control of Markov decision processes*. PhD thesis, Massachusetts Institute of Technology, 2007.
- [145] A. Tamar, D. Di Castro, and S. Mannor. Policy gradients with variance related risk criteria. In *International Conference on Machine Learning*, pages 387–396, 2012.
- [146] A. Tamar, Y. Glassner, and S. Mannor. Optimizing the CVaR via Sampling. In *AAAI*, 2015.
- [147] A. Tamar, S. Mannor, and H. Xu. Scaling up robust MDPs using function approximation. In *International Conference on Machine Learning*, pages 181–189, 2014.
- [148] G. Theocharous and A. Hallak. Lifetime value marketing using reinforcement learning. *RLDM*, page 19, 2013.
- [149] P. Thomas, G. Theocharous, and M. Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning*, 2015.

- [150] O. Toker and H. Ozbay. On the NP-hardness of solving bilinear matrix inequalities and simultaneous stabilization with static output feedback. In *American Control Conference*, volume 4, pages 2525–2526. IEEE, 1995.
- [151] J. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.
- [152] V. Vilkov. Some properties of the Lagrange function in mathematical programming problems. *Cybernetics and Systems Analysis*, 22(1):75–81, 1986.
- [153] P. Vitale. *Risk-averse traders with inside information*. Department of Applied Economics, University of Cambridge, 1995.
- [154] P. Vitale. Linear risk-averse optimal control problems: Applications in economics and finance. *Available at SSRN 2334131*, 2012.
- [155] M. P. Vitus and C. J. Tomlin. On feedback design and risk allocation in chance constrained control. In *Proc. IEEE Conf. on Decision and Control*, pages 734–739, 2011.
- [156] Y. Wang and S. Boyd. Fast model predictive control using online optimization. *IEEE Transactions on Control Systems Technology*, 18(2):267–278, 2010.
- [157] D. White. Mean, variance, and probabilistic criteria in finite Markov decision processes: A review. *Journal of Optimization Theory and Applications*, 56(1):1–29, 1988.
- [158] R. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [159] C. Wu and Y. Lin. Minimizing Risk models in Markov decision process with policies depending on target values. *Journal of Mathematical Analysis and Applications*, 23(1):47–67, 1999.
- [160] H. Xu and S. Mannor. The robustness-performance tradeoff in Markov decision processes. In *Advances in Neural Information Processing Systems*, pages 1537–1544, 2006.
- [161] V. Yakubovich.  $S$ -procedure in nonlinear control theory. *Vestnik Leningrad University*, 1:62–77, 1971.