Risk-sensitive Markov Decision Processes

vorgelegt von
Diplom Informatiker
Yun Shen
geb. in Jiangsu, China

von der Fakultät IV, Elektrotechnik und Informatik
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

doctor rerum naturalium
-Dr. rer. nat.-

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Manfred Opper
Gutachter:	Prof. Dr. Klaus Obermayer
	Prof. Dr. Wilhelm Stannat
	Prof. Dr. Łukasz Stettner
	Prof. Dr. Vivek Borkar

Tag der wissenschaftlichen Aussprache: 1. Juni 2015

Berlin 2015

Dedicated to my parents.

Acknowledgments

First, I would like to express my sincere gratitude to Prof. Klaus Obermayer, the supervisor of this thesis, for giving me the opportunity to work in his Neural Information Processing Group at the Technische Universität Berlin. His guidance helped me to learn how to approach a problem scientifically. He also gave me the freedom and trust to pursue my scientific ideas. The scientific environment set up by him and meetings with collaborators greatly contributed to the success of this research project. I would also like to express my special gratitude to Prof. Wilhelm Stannat, who taught me a great deal on various mathematical subjects. Without his patient guidance, the mathematical foundation for this thesis would never have been accomplished.

I'm indebted to Steffen Grünewälder, who motivated me and initialized the research on the topic of this thesis, and to Wendelin Böhmer, with whom I had a lot of scientific discussions that were very inspiring and contributed plenty of ideas to my work.

Many thanks to Michael J. Tobia and Tobias Sommer-Blöchl for sharing experimental data and collaborating on human decision-making studies. Special thanks to Ruihong Huang for introducing me the knowledge of algorithmic trading, sharing high frequency real market data and collaborating on the data analysis. Thanks also goes to Chang Yan for analyzing the market data.

I would like to thank all past and present group members at the Neural Information Processing Group: Arno Onken, Marcel Stimberg, Deepak Srinivasan, Klaus Wimmer, Nicolas Neubauer, Aki Naito, Muamar Ahmad, Johannes Jain, Dipanjan Roy, Johannes Mohr, Sambu Seo, Michael Scholz, Maziar Hashemi-Nezhad, Audrey Houillon, Timm Lochmann, Konstantin Mergenthaler, Sinem B. Beylergil, Rong Guo, Josef Ladenbauer, Moritz Augustin, Robert Pröpper, Robert Meyer, Ivo Trowitzsch, Florian Aspart, Ningfei Li, etc. They have been a valuable source of scientific ideas, knowledge, and perspectives. I would also like to thank Roswitha Paul-Walz from the International Office of Technische Universität Berlin, who helped me, a foreigner, adjust quickly to the new living environment in Berlin.

Finally, I would like to thank my parents who supported me through the years.

※ ※ ※

During the work on this thesis, I received financial support from Technische Universität Berlin (Stipendium des Präsidenten der TU Berlin) and from German Federal Ministry of Education and Research (Bersteinfokus Lernen TP1, 01GQ0911).

Abstract

This thesis investigates risk-sensitive sequential decision-making problems in an uncertain environment.

We first introduce the axiomatic concept of valuation functions that generalize known concepts of risk measures in mathematical finance to cover most of the existing risk related models in various fields, in particular, behavioral economics and cognitive neuroscience.

By applying this concept to Markov processes, we construct valuation maps and develop thereby a unified framework for incorporating risk into Markov decision processes on general spaces. Within the framework, we study mainly two types of infinite-horizon risk-sensitive criteria, discounted and average valuations, and solve the associated optimization problems by value iteration. For the discounted case, we propose a new discount scheme, which is different from the conventional form but consistent with existing literature, while for the average criterion, we state Lyapunov-type stability conditions that generalize known conditions for Markov chains to ensure the existence of solutions to the optimality equation and a geometric convergence rate for the value iteration.

Applying a set of valuation functions, called utility-based shortfall, we derive a family of model-free risk-sensitive reinforcement learning algorithms for solving the optimization problems corresponding to risk-sensitive valuations. In addition, we find that when appropriate utility functions are chosen, agents' behaviors express key features of human behavior as predicted by prospect theory, for example, different risk preferences for gains and losses, as well as the shape of subjective probability curves.

As a proof of principle for the applicability of the new algorithms, we apply them to two tasks, 1) to quantify human behavior in a sequential investment task and 2) to perform risk control in simulated algorithmic trading of stocks. In the first task, the risk-sensitive variant provides a significantly better fit to the behavioral data and it leads to an interpretation of the subject's responses which is indeed consistent with prospect theory. The analysis of simultaneously measured fMRI signals show a significant correlation of the risk-sensitive temporal difference error with BOLD signal change in the ventral striatum. In the second task, our algorithm outperforms the risk-neutral reinforcement learning algorithm by keeping the trading cost at a substantially low level at the spot when the 2010 Flash Crash happened, and significantly reducing the risk over the whole test period.

Zusammenfassung

Diese Dissertation untersucht risikosensitive sequenzielle Entscheidungsprobleme in stochastischen Umgebungen.

Wir führen zunächst axiomatisch das Konzept von *Valuation Function* ein, welches Risikomaße aus der Finanzmathematik verallgemeinert. Dieses umfassende Modell deckt ebenfalls risikobezogene Modelle aus einer Vielfalt von anderen Disziplinen ab, insbesondere der Verhaltensökonomie und der kognitiven Neurowissenschaft.

Durch eine Erweiterung mit Markov-Prozessen konstruieren wir sogenannte *Valuation Maps*, welche einen einheitlichen Rahmen für die Berücksichtigung von Risiken in Markov-Entscheidungsprozessen auf allgemeinen Räume erlauben. Hierbei untersuchen wir hauptsächlich zwei Arten von unbegrenzten risikosensitiven Bewertungen: Eine zeitlich diskontiertes und eine zeitlich gemitteltes Kriterium. Die damit verbundenen Optimierungsprobleme werden durch Bewertungsiteration gelöst. Für den diskontierten Fall schlagen wir einen neuen Ansatz vor, welcher von etablierten Paradigmen abweicht, aber dadurch im Einklang mit allgemein akzeptierter Literatur aus der Psychologie und Verhaltensökonomie ist. Um eine geometrische Konvergenzrate der Bewertungsiteration für das zeitlich gemittelte Kriterium zu gewährleisten, geben wir Lyapunov-Typ-Stabilitätsbedingungen an, welche etablierte Bedingungen für Markov-Ketten verallgemeinern.

Unter Annahme einer bestimmten Klasse von Bewertungsfunktionen, des sogenannten *Utility based Shortfall*, leiten wir eine Familie von modellfreien risikosensitiven *Reinforcement Learning* Algorithmen ab, welche unsere Methode auf praktische Probleme anwendbar macht. Mit geeigneten Risikofunktionen können diese Algorithmen wichtige Eigenschaften des menschlichen Verhaltens aus der *Prospect Theory* replizieren, z.B. unterschiedliche Risikopräferenzen für Gewinne und Verluste, sowie die Form der subjektiven Wahrscheinlichkeitskurven.

Zur Demonstration des Prinzips und der neuen Algorithmen wenden wir diese auf zwei Aufgaben an: 1) die Quantifizierung von menschliche Verhalten in einer sequentiellen Investitionsaufgabe und 2) die Simulation von algorithmischen Handel mit Aktien. In der ersten Aufgabe zeigt unsere risikosensitive Variante eine bessere Erklährung der Verhaltensdaten, und erlaubt erstmals eine Interpretation, welche konsistent mit der Prospect Theory ist. Die Analyse der gleichzeitig gemessenen fMRI Signale zeigt eine signifikante Korrelation einiger Modellvariablen mit BOLD Signaländerungen im ventralen Striatum. Auch in der zweiten Aufgabe zeigt unser Algorithmus eine starke Performance. Sowohl das Riskio über den gesamten Testzeitraum, als auch in besonderen Kriesensituationen wie dem *2010 Flash Crash*, ist deutlich niedriger als bei gewöhnlichen, risikoneutralen Reinforcement Learning Algorithmen.

Vita

1983	Born, Changzhou, Jiangsu Province, China
2002–2005	B.Eng. (Computer Science), Shanghai Jiao Tong University, China
2005–2008	Dipl.-Inf., Technische Universität Berlin
2008–2009	M.Eng. (Computer Science), Shanghai Jiao Tong University, China
2009–	PhD student, Technische Universität Berlin

Publications

W. Böhmer, S. Grünewälder, Y. Shen, M. Musial, and K. Obermayer. (2013) Construction of approximation spaces for reinforcement learning. *Journal of Machine Learning Research*, 14:2067–2118.

Y. Shen, W. Stannat, and K. Obermayer. (2013) Risk-sensitive Markov Control Processes. *SIAM Journal on Control and Optimization*, (5):3652–3672.

Y. Shen, R. Huang, C. Yan, and K. Obermayer. (2014a) Risk-averse reinforcement learning for algorithmic trading. In A. Serguieva, D. Maringer, V. Palade, and R. Almeida, editors, *Proceedings of 2014 IEEE Computational Intelligence for Financial Engineering and Economics*, pages 391–398.

Y. Shen, W. Stannat, and K. Obermayer. (2014b) Risk-sensitive Markov control processes with general convex risk maps. *Submitted to SIAM Journal on Control and Optimization. Arxiv preprint arXiv:1403.3321.*

Y. Shen, W. Stannat, and K. Obermayer. (2014c) A unified framework for risk-sensitive Markov control processes. To appear in *Proceedings of 53rd IEEE Conference on Decision and Control.*

Y. Shen, M. Tobia, T. Sommer, and K. Obermayer. (2014d) Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328.

Contents

List of Figures

1

Introduction

Plans based on average assumptions are wrong on average.
— Sam L. Savage (2009)

Motivation

Risk arises from the uncertainties associated with future events, and is inevitable since the consequences of actions are uncertain at the time when a decision is made. Hence, risk has to be taken into account by the decision maker, consciously or unconsciously.

An economically rational decision-making rule, which is *risk-neutral*, is to select the alternative with the highest expected reward. In the context of sequential or multistage decision-making problems, the objective is then to find the best policy that maximizes the expected *cumulative* rewards in an environment typically described by a *Markov decision process* (MDP, see e.g. Puterman, 1994; White, 1993 and Hernández-Lerma and Lasserre, 1996, 1999 under the name *Markov control processes*). A computational approach called *reinforcement learning* (RL, see e.g. Sutton and Barto, 1998) is widely applied to optimize this risk-neutral objective by interacting with the environment. RL is a well-developed model not only for guiding agents to learn *how* to select actions, but also for explaining *why* humans or nonhuman animals take those actions, because similar computational structures, such as dopaminergically mediated reward prediction errors, have been identified across species (Schultz et al., 1997; Schultz, 2002).

Besides risk-neutral policies, *risk-averse* policies, which accept a choice with a more certain but possibly lower expected reward, are also considered economically rational (Gollier, 2004). For example, a risk-averse investor might choose to put money into a bank account with a low but guaranteed interest rate, rather than into a stock with possibly high expected returns but also a chance of high losses. Conversely, *risk-seeking* policies, which prefer a choice with less certain but possibly high reward, are considered economically irrational. Human agents are, however, not always economically rational (Gilboa, 2009). Behavioral stud-

ies show that human can be risk-seeking in one situation while risk-averse in another situation (Kahneman and Tversky, 1979). RL algorithms, along with the underlying MDP, developed so far cannot effectively model these complicated risk-preferences.

The aim of this thesis is, therefore, twofold: first, to develop a general theoretical framework for incorporating risk into MDPs; and second, to derive RL-type algorithms for solving the optimization problem induced by the framework. The derived algorithms can be applied to two types of problems. One is methodological, i.e., for instance, the algorithms ought to tell an agent how to avoid risk, if risk-averse behavior is expected. Furthermore, the algorithms should also be able to control the degree of risk-averseness. The other is epistemological, that is, given observations of one agent's behavior (especially human agents), the algorithms can judge the agent's risk preference, and quantify its degree as well.

Related literature

Risk-sensitive decision-making problems, in the context of MDPs, have been investigated in various fields, e.g., in machine learning (Heger, 1994; Mihatsch and Neuneier, 2002), optimal control (Hernández-Hernández and Marcus, 1996), finance (Ruszczyński, 2010), operations research (Howard and Matheson, 1972 and Borkar, 2002), as well as cognitive neuroscience (Nagengast et al., 2010; Braun et al., 2011; Niv et al., 2012).

The core of MDPs consists of two sets of *objective* quantities describing the environment: immediate *rewards* obtained at states by executing actions, and *transition probabilities* for switching states when performing actions. Facing the same environment, however, different agents might have different policies, which indicates that risk is taken into account differently by different agents. Hence, to incorporate risk, which is derived from both quantities, all existing literature applies a nonlinear transformation to either the experienced reward values or to the transition probabilities, or to both. The former is the canonical approach in classical economics, as in expected utility theory (Gollier, 2004), while the latter originates from behavioral economics, as in *subjective probability* (Savage, 1972), but is also derived from a rather recent development in mathematical finance, *convex/coherent risk measures* (CRMs, Artzner et al., 1999; Föllmer and Schied, 2002). For modeling human behaviors, prospect theory (Kahneman and Tversky, 1979) suggests that we should combine both approaches, i.e., human beings have different perceptions not only for the same objective amount of rewards but also the same value of the true probability. Recently, Niv et al. (2012) combined both approaches by applying piecewise linear functions (an approximation of a nonlinear transformation) to reward prediction errors that contain the information of rewards directly and the information of transition probabilities indirectly. Importantly, the reward prediction errors that incorporated experienced risk were strongly coupled to activity in the nucleus accumbens of the ventral striatum, pro-

viding a biologically based plausibility to this combined approach. We will show (see Section 2.4.5) that this algorithm proposed by Niv et al. (2012) is a special case of our general risk-sensitive framework.

Most of the literature in economics or engineering fields focuses on economically rational risk-averse/-neutral strategies, which are not always adopted by humans. The models proposed in behavioral economics, despite allowing economic irrationality, require knowledge of the true probability, which usually is not available at the outset of a learning task. In neuroscience, on the one hand, several works (e.g., Wu et al., 2009; Preuschoff et al., 2008) follow the same line as in behavioral economics and require knowledge of the true probability. On the other hand, though different modified RL algorithms (e.g., Glimcher et al., 2008; Symmonds et al., 2011) are applied to model human behaviors in learning tasks, the algorithms often fail to generalize across different tasks.

Road map

To overcome the limitations mentioned above, we develop a novel general framework of risk-sensitive Markov decision processes by introducing nonlinear transformations to both rewards and transition probabilities. Risk-sensitive objectives, including the discounted and average criteria, are derived and optimized by value iteration or dynamic programming. This framework covers sequential decision making problems in various fields, e.g., optimal control, operations research, finance, behavioral economics and cognitive neuroscience. Based on this general framework, we derive a set of RL algorithms, which are model-free, i.e., the knowledge of the transition and reward model is not needed. Finally, as applications, we apply the RL algorithms 1) to quantify human behavior in terms of risk sensitivity and 2) to reduce risk in algorithmic trading. Note that the first application is epistemological, while the second one is methodological. They represent exactly the two types of problems, to which we hope to apply the derived algorithms.

This thesis is organized as follows.

In Chapter 2, we focus on the fundamental problem of measuring risk of choices with multiple (possibly) random outcomes. By extending the definition of CRMs, we introduce the concept of *valuation functions* in Section 2.3. We show in Section 2.4 that this concept can cover most of the existing models in various fields, e.g., optimal control, operations research, finance, behavioral economics and cognitive neuroscience. In particular, it is shown in Section 2.4.5 that the key features predicted by prospect theory (Kahneman and Tversky, 1979), e.g., different risk-preferences for gains and losses as well as the shape of subjective probability curves, can be replicated by applying a special family of valuation functions, called *utility-based shortfall* with appropriately chosen utility functions.

In Chapter 3, we apply a constructive approach that maintains the Markov property that is necessary for the existence of stationary optimal policies for two infinite-horizon objectives. To this end, we introduce the concept of *valuation*

maps, which extend valuation functions to a temporal setting. We show also in this chapter that our constructive approach coincides with the idea of applying *dynamic time-consistent risk measures* (see e.g. Cheridito et al., 2006; Ruszczyński, 2010).

Chapter 4 is devoted to the Poisson equation with nonlinear valuation maps. This sets a theoretical foundation for solving the optimization problem induced by average risk-sensitive MDPs to be studied in Chapter 5. We generalized the Lyapunov approach that was applied by Hairer and Mattingly (2011) to ensure the geometric ergodicity for Markov chains. Our assumptions and their variants for different types of valuations maps are composed of two conditions, a) the existence of a Lyapunov function to control the growth of iterations, and b) a Doeblin-like condition for local contraction. In particular, for the same problem on finite state spaces, the above two conditions can be reduced to one condition: a multistep Doeblin-like condition for global contraction.

Chapter 5 introduces a unified framework for measuring risk in the context of Markov decision processes with risk maps on general Borel spaces. Within the framework, applying weighted norm spaces to incorporate also unbounded costs, we study two types of infinite-horizon risk-sensitive criteria, discounted and average valuation, and solve the associated optimization problems by value iteration. For the discounted case, we propose a new discount scheme, which is different from the conventional form but consistent with the existing literature, while for the average criterion, we state Lyapunov-type stability conditions that generalize known conditions for Markov chains to ensure not only the existence of solutions to the optimality equation, but also the a geometric convergence rate for the value iteration.

In Chapter 6, restricting to MDPs on finite state-action spaces, we derive a risk-sensitive Q-learning algorithm, which does not require the knowledge of the underlying MDP, by applying a rich family of valuation maps, called utility-based shortfall. We prove its convergence with a stochastic approximation technique developed by Bertsekas and Tsitsiklis (1996).

In Chapter 7, as a proof of principle for the applicability of the new algorithm, we apply it to quantify human behavior in a sequential investment task. We find, that the risk-sensitive variant provides a significantly better fit to the behavioral data and that it leads to an interpretation of the subject's responses which is indeed consistent with prospect theory. The analysis of simultaneously measured fMRI signals show a significant correlation of the risk-sensitive TD error with BOLD signal change in the ventral striatum. In addition we find a significant correlation of the risk-sensitive Q-values with neural activity in the striatum, cingulate cortex and insula, which is not present if standard Q-values are used.

Finally, in Chapter 8, we apply the risk-sensitive RL algorithms to algorithmic trading. Our approach is tested in an experiment based on 1.5 years of millisecond time-scale limit order data from NASDAQ, which contain the data around the 2010 flash crash. The results show that our algorithm outperforms the risk-neutral reinforcement learning algorithm by 1) keeping the trading cost at a substantially

low level at the spot when the flash crash happened, and 2) significantly reducing the risk over the whole test period.

2

Measure of Risk: an Indirect Approach

Risk is a choice rather than a fate.
— Peter L. Bernstein (1997)

Précis In this chapter we introduce an axiomatic framework of valuation functions that generalizes the known concept of risk measures applied in mathematical finance (Artzner et al., 1999; Föllmer and Schied, 2002). In fact, a valuation function can be viewed as an indirect measure of risk in general decision-making problems and its induced risk preference is determined by its mathematical property. We will show that our framework covers several important families of examples that have been considered in literature of various related fields including mathematical finance, behavioral economics and cognitive neuroscience. Properties of each family will be studied and compared.

Publications related to this chapter Section 2.3 and 2.4.5 have been published in Shen et al., 2014d, Section 2. Most of the examples in Section 2.4 have been contained in Shen et al., 2013, Section 4.3.

A Motivating Example

In 1970s, Kahneman and Tversky (1979) conducted a series of laboratory studies on human decision making. They asked human subjects, mainly their students, to make decisions between several pairs of choices. One example of the paired choices is as follows (Kahneman and Tversky, 1979, Problem 1).

Example 2.1. 72 subjects were asked to choose between Choice A and B:

A:	2,500 with probability	.33	B:	2,400 with certainty.
	2,400 with probability	.66		
	0 with probability	.01		

The result turned out that 82% of the subjects chose B in the above problem.

In real life, the decisions we are facing are far more complicated than the above oversimplified example. Nevertheless, this example presents several important features of *choices* and also features of *risk* associated to choices.

1. Risk arises from the *uncertainties associated with future events*. For instance, in the above example, before choosing (and knowing the realization of) A, how many points we will obtain is uncertain. Conversely, if there is no uncertainty, like Choice B, there is no risk.

2. For a risky choice, like A, there are different *events* associated with different *outcomes* and different *probabilities*. Choice A has 3 events, with the outcomes of obtaining i) 2500, ii) 2400 or iii) 0 points. Their corresponding probabilities are 33%, 66% and 1% respectively. Risk arises, therefore, from both uncertain elements: outcomes and probabilities.

3. Risk is *subjective* in the sense that it is evaluated differently by different individuals. In the example, we see that facing the same pair of choices, the subjects make different decisions: the majority chose B, while the minority preferred A.

Hence, instead of asking how to measure risk, the more fundamental question to ask is how to make a decision between 2 or more choices that are possibly risky. In other words, the core of the problem of measuring risk is to determine the *preference* of (possibly) risky choices. Furthermore, the preference can be individually different. To solve this problem, first of all, we need a general quantitative model for choices.

2.1 Choices

We have seen in Example 2.1 that there are two essential elements of a choice, i) outcomes and ii) probabilities, both of which are associated to events. A risk-free choice like Choice B, where only one event exists, is a special case. For the sake of generality, we employ the language of modern probability theory (Kolmogorov, 1933) and introduce the following (mathematically rigorous) definition.

Definition **2.2** (Choice)**.** *Let Ω be a state space and $\mathcal{F}$ be an event space, which is a σ-algebra of subsets of Ω. Then a* choice *(v, μ) consists of i) an outcome function, $v : \Omega \to \mathbb{R}$, which is a real-valued $\mathcal{F}$-measurable function, and ii) a probability measure, μ on $(\Omega, \mathcal{F})$. Denote by $\mathscr{C}$ the space of all choices on $(\Omega, \mathcal{F})$.*

Remark 2.3. In the terminology of probability theory, v is also called a *random variable*.

This definition is general since it covers both discrete and continuous cases. If Ω is a space with a finite number of elements, e.g., N elements, as Choice A in Example 2.1, then a conventional σ-algebra $\mathcal{F}$ is the power set of Ω, containing

all subsets of Ω. The probability measure $\mu(\{\omega_i, i \in I\}) = \sum_{i\in I}\mu(\{\omega_i\})$, where I is any set of $\{1, 2, \ldots, N\}$. If Ω is continuous, e.g., $\Omega = \mathbb{R}$, then one example of $\mathcal{F}$ is the space containing all sets of the form (a, b), $a < b \in \mathbb{R}$, which is an open subset of $\mathbb{R}$, and also all their unions and complimentary sets. The probability measure μ is then defined on open subsets of $\mathbb{R}$ (as well as their complimentary sets).

States To apply Definition 2.2 to a specific decision problem, one needs to first determine the state space Ω, which "should be thought of as an exhaustive list of all scenarios that might unfold." (Gilboa, 2009, Chapter 10) In other words, a state "should resolves all uncertainty." (Savage, 1972) We should define states such that the outcome of each state, $v(\omega)$, is unique and deterministic. Otherwise, we can always merge those states with the same outcome to be a new single state. For instance, in Example 2.1, Choice A has three states, for there are three different outcomes, while Choice B has only one state, for only one outcome exists. Although some paradoxes, e.g., Hempel's paradox (Hempel, 1945) and Good's variation (Good, 1986), indicate that in some decision problems, it is hard to define states properly (for more details, see also Gilboa, 2009, Chapter 11), we always assume in this thesis that the state space is well defined.

Probabilities The probability measure μ in a choice C must satisfy that

(i) $\mu(\mathsf{E}) \in [0, 1]$, for each $\mathsf{E} \in \mathcal{F}$ with $\mu(\Omega) = 1$;

(ii) for all countable collections of events $\{\mathsf{E}_i\}_{i\in I}$ of pairwise disjoint sets:

$$\mu(\cup_{i\in I}\mathsf{E}_i) = \sum_{i\in I}\mu(\mathsf{E}_i).$$

Note that the key property required here is the σ-additivity implied by (ii). That is, by letting $\mathsf{B}_n := \cup_{i=1}^{n}\mathsf{E}_i$ and $\mathsf{B} := \cup_{i=1}^{\infty}\mathsf{E}_i$, $\mathsf{B}_n \nearrow \mathsf{B}$ and (ii) implies

$$\mu(\lim_{n\to\infty}\mathsf{B}_n) = \mu(\cup_{i=1}^{\infty}\mathsf{E}_i) = \sum_{i=1}^{\infty}\mu(\mathsf{E}_i) = \lim_{n\to\infty}\mu(\mathsf{B}_n). \tag{2.1}$$

It is also the reason why we need the notation of σ-algebra and the assumption of measurability for outcome functions. In fact, (2.1) gives an additional constraint of continuity, which allows to define the Lebesgue integral properly. This sufficiently general setting is the most widely used framework in probability theory. However, we will see in Section 2.4.4 that, in order to model human decision preferences, some properties in the above framework ought to be generalized, by which one can define a more generalized version of integral than Lebesgue's.

Outcome In Example 2.1, all outcomes are real-valued monetary payoffs. In this case, these outcomes are *objective*, in the sense that all decision makers obtain the same amount of (possibly random) payoff if they make the same decision.

However, the same amount of payoff might have different *subjective* values for different individuals. For instance, the same amount of money might have a higher subjective value for the poor than for the rich. This subjective value is usually called *utility* in neoclassical economics (Morgenstern and Neumann, 1944). In other cases, the outcome might not be easily quantified, e.g., happiness, satisfactory, sadness, pain, etc. It requires, therefore, a subjective valuation function that is individually different. In such cases, we may assume that there exists a real-valued $\mathcal{F}$-measurable function that "calculates" the subjective outcome for each state. In fact, for decision makers like human beings, we can reasonably assume that there exists a neurobiological mechanism that calculates the subjective value for each outcome represented as different kinds of perceptual stimuli, though the mechanism is still a "black box" that so far has not been fully understood. Therefore, in (behavioral) economical experiments, to avoid this ambiguity and uncertainty, outcomes are always represented as quantifiable monetary payoffs. In this thesis, we follow the same discipline as well. Throughout this thesis, by outcome, we mean objective payoff, whereas subjective outcome is meant by utility.

2.2 Preference and valuation

Given two choices, we define the preference as follows.

Definition **2.4** (Preference). *Let $\mathscr{C}$ be the space of all choices on $(\Omega, \mathcal{F})$. Given a real-valued function $\rho : \mathscr{C} \to \mathbb{R}$, we say C_1 is preferred to C_2 if $\rho(C_1) \geq \rho(C_2)$, $C_1, C_2 \in \mathscr{C}$.*

Note that the real set $\mathbb{R}$ is *partially ordered* (Simon and Barry, 1980, Section 1.1), i.e., for any a, b and c in $\mathbb{R}$, it satisfies

(i) (reflexivity) $a \leq a$;

(ii) (antisymmetry) if $a \leq b$ and $b \leq a$, then $a = b$;

(iii) (transitivity) if $a \leq b$ and $b \leq c$, then $a \leq c$.

In principle, the set $\mathbb{R}$ can be replaced by any other partially ordered sets. However, for the sake of computation, $\mathbb{R}$ is the most convenient set to work with.

The transitivity implies that by Definition 2.4 if C is preferred to B and B is preferred to A, then C is preferred to A. Hence, we have already assumed some degree of "rationality" in Definition 2.4. Put into the framework of von Neumann and Morgenstern axioms (Morgenstern and Neumann, 1944; see also the next subsection), it means that we assume the *axiom of weak order*.

We briefly introduce two evaluation functions that are widely used in (behavioral) economics.

2.2.1 Expected utility

The most widely used evaluation function in economic literature is the *expected utility*:

$$\rho^u(C = (v, \mu)) := \int_\Omega u(v(\omega))\mu(d\omega), C \in \mathscr{C}_u \tag{2.2}$$

Here, $\mathscr{C}_u := \{C \in \mathscr{C} | \rho^u(C) < \infty\}$ and $u : \mathbb{R} \to \mathbb{R}$ is a *utility function*, which is usually assumed to be an increasing function.

The idea of expected utility first appeared explicitly in Daniel Bernoulli's St. Petersburg paradox (Bernoulli, 1954) in 1738. However, it became prevailing in economics only after von Neumann and Morgenstern published their cornerstone book (Morgenstern and Neumann, 1944), where they proved that under 3 axioms, *weak order, continuity and independence*, the expect utility is the unique form of evaluation functions.

2.2.2 Prospect theory

Now we return to the motivating example at the beginning of this chapter by Kahneman and Tversky. Their series of laboratory experiments (Kahneman and Tversky, 1979) showed that *homo sapiens* is not so "rational" as von Neumann and Morgenstern postulated. Instead, Kahneman and Tversky proposed a modified version of expect utility, called *prospect theory*, which fits human behavior better than expected utility. For a discrete state space, their model is

$$\rho(C = (v, \mu)) = \sum_{\omega \in \Omega} u(v(\omega))w(\mu(\{\omega\})). \tag{2.3}$$

Here, u is again a utility function and $w : [0, 1] \to [0, 1]$ is a *probability weighting function*. The key observation of Kahneman and Tversky is that w is not a linear function for most of the human subjects who participated in the experiments.

The probability weighting function can be interpreted as a subjective perception of true probabilities. In other words, along with the subjective perception of outcomes, different people might also perceive the same probability differently. Both subjective perceptions contribute together to the subjective evaluation of choices and, therefore, lead to possibly different decisions even when facing with the same problem.

Let us look at the probability weighting function considered in prospect theory in more details. With a slight abuse of terminology, write $w(\mu(\{\omega\}))$ as $\nu(\{\omega\})$. Assume further that ν is a function on the σ-algebra $\mathcal{F}$ like the probability measure μ satisfying $\nu(\Omega) = 1$ and $\nu(\emptyset) = 0$. To define $\nu(\{\omega\})$ properly, one should first order states according to their outcomes as follows

$$v(\omega_1) < v(\omega_2) < \ldots < v(\omega_n).$$

Here it is assumed that outcomes of different states are different, otherwise we can always merge those states with the same outcome to be a new single state. Finally, the prospect theory employs the following definition of subjective probability for each state

$$\nu(\{\omega_i\}) := \nu(\{\omega_i, \omega_{i+1}, \dots, \omega_n\}) - \nu(\{\omega_{i+1}, \omega_{i+2}, \dots, \omega_n\}).$$

The observation that ν is not necessarily linear leads to a violation of additivity, which is the key assumption of classical probability measures. However, $\nu(\{\omega\})$ has to be nonnegative for each ω. The assumption of additivity is, therefore, replaced by a more generalized assumption:

$$\nu(A) \leq \nu(B), \text{ whenever } A \subset B \in \mathcal{F}. \tag{2.4}$$

This lies exactly in the mathematical framework of *capacity theory* by Choquet (1953) and *nonadditive measures* by Denneberg (1994), which will be briefly introduced in Section 2.4.4.

Remark 2.5. There are also other (for psychologists or economists, probably even more) important ingredients in prospect theory, e.g., framing effects, nonlinear preferences, source dependence, gain-loss asymmetry. However, since we are now focusing on the mathematical model used by prospect theory, those ingredients are not explained here and those readers who are interested in more details are referred to Kahneman and Tversky (1979) and Tversky and Kahneman (1992) or a recent book by Wakker (2010) and references therein.

2.3 Axioms for valuation functions

We now propose our normative axiomatic framework for evaluation functions based on the theory of *coherent/convex risk measures* that has been widely applied in financial mathematics since the seminal papers by Artzner et al. (1999) and Föllmer and Schied (2002) were published one and half decades ago.

We first introduce some notations. Let $\mathscr{L}$ be a linear space of real-valued $\mathcal{F}$-measurable functions containing all constant functions, which implies that, with a slight abuse of notations, $\mathbb{R} \subset \mathscr{L}$. For $v, u \in \mathscr{L}$, we say $v \leq u$ if $v(\omega) \leq u(\omega)$ for all $\omega \in \Omega$. Let $\mathscr{P}$ be the space of all probability measures on $(\Omega, \mathcal{F})$.

Definition **2.6.** *A mapping $\rho : \mathscr{L} \times \mathscr{P} \to \mathbb{R}$ is called a* valuation function*, if it satisfies for each $\mu \in \mathscr{P}$,*

(I) (monotonicity) $\rho(v, \mu) \leq \rho(u, \mu)$, whenever $v \leq u \in \mathscr{L}$;

(II) (translation invariance) $\rho(v + y, \mu) = \rho(v, \mu) + y$, for any $y \in \mathbb{R}$.

(III) (centralization) $\rho(0, \mu) = 0$.

Note that v and u are outcomes of two choices. Monotonicity reflects the intuition that given the same distribution μ, if the outcome of one choice is *always* (for all states) higher than the outcome of another choice, the *valuation* of the choice must be also higher. Under the axiom of translation invariance, the sure outcome y (equal outcome for every state) after executing decisions, is considered as a sure outcome before making decision. This also reflects the intuition that there is no risk if there is no uncertainty. Finally, the axiom of centralization sets the reference point being 0. In principle, the reference point can be variant for different agents or human subjects. This axiom is, however, not restrictive, since for any non-centralized valuation function ϱ, one can always obtain its centralized version ρ by $\rho(v,\mu) := \varrho(v,\mu) - \varrho(0,\mu)$.

2.3.1 Risk preference

We now link valuation functions to their induced risk preferences. First we define some concepts.

Definition **2.7.** *A valuation function ρ is said to be*

- convex*, if for all $\alpha \in [0,1]$, $v,u \in \mathscr{L}$ and $\mu \in \mathscr{P}$,*

$$\rho(\alpha v + (1-\alpha)u, \mu) \leq \alpha\rho(v) + (1-\alpha)\rho(u,\mu); \tag{2.5}$$

- concave*, if $\tilde{\rho}(\cdot,\mu) := -\rho(-\cdot,\mu)$ is a convex valuation function;*
- homogeneous*, if $\rho(\lambda v,\mu) = \lambda\rho(v,\mu)$ for all $\lambda \in \mathbb{R}_+$, $v \in \mathscr{L}$ and $\mu \in \mathscr{P}$;*
- coherent*, if ρ is concave and homogeneous.*

To judge the risk-preference induced by a certain type of valuation functions, we follow the rule that *diversification* should be preferred if the agent is *risk-averse*. More specifically, suppose an agent has two possible choices, one of which leads to the future reward (v,μ) while the other one leads to the future reward (u,ν). For simplicity we assume $\mu = \nu$. If the agent *diversifies*, i.e., if one spends only a fraction α of the resources on the first and the remaining amount on the second alternative, the future reward is given by $\alpha v + (1-\alpha)u$. If the applied valuation function is concave, i.e.,

$$\rho(\alpha v + (1-\alpha)u, \mu) \geq \alpha\rho(v,\mu) + (1-\alpha)\rho(u,\mu),$$

for all $\alpha \in [0,1]$ and $v,u \in \mathscr{L}$, then the diversification should increase the (subjective) valuation. Thus, we call the agent's behavior *risk-averse*. Conversely, if the applied valuation function is *convex*, the induced risk-preference should be *risk-seeking*. Concave valuation functions have the following property.

Proposition **2.8.** *Let ρ be a concave valuation function. Then*

$$-\rho(-f) \geq \rho(f), \forall f \in \mathscr{L}. \tag{2.6}$$

Proof. By the definition of concavity in Definition 2.7, we have for each $f \in \mathscr{L}$

$$\frac{1}{2}\rho(f) + \frac{1}{2}\rho(-f) \leq \rho(0) = 0$$

where the equality is due to the centralization axiom. Hence, we obtain (2.6). □

Homogeneous valuation functions satisfy:

PROPOSITION **2.9.** *Let ρ be a concave homogeneous valuation function. Then*

$$\rho(f+g) \geq \rho(f) + \rho(g), \forall f, g \in \mathscr{L}. \tag{2.7}$$

Conversely, if ρ is convex and homogeneous, then

$$\rho(f+g) \leq \rho(f) + \rho(g), \forall f, g \in \mathscr{L}. \tag{2.8}$$

Proof. We prove only the first case, where ρ is coherent. By concavity and homogeneity, we have

$$\frac{1}{2}\rho(f) + \frac{1}{2}\rho(g) \leq \rho(\frac{1}{2}f + \frac{1}{2}g) = \frac{1}{2}\rho(f+g), \forall f, g \in \mathscr{L},$$

which implies the required inequality. □

Remark 2.10. Following the literature of coherent measures (Delbaen, 2000), we call a valuation function

(a) *subadditive*, if it satisfies (2.8);

(b) *superadditive*, if it satisfies (2.7).

2.3.2 Comparison with risk measures

Comparing with risk measures defined in financial mathematics (Artzner et al., 1999; Föllmer and Schied, 2002), we state some remarks.

1. It is also easy to see that for a fixed probability measure $\mu \in \mathscr{P}$, $\varrho(\cdot) := -\rho(\cdot, \mu)$ is then a valid risk measure.

2. The risk measures applied in financial mathematics are literally defined to measure *risk*. The objective is, therefore, to minimize risk, rather than to maximize *value* in our framework, and the correspondence between risk preference and convexity/concavity in risk measure theory is reversed, i.e., convex (respectively concave) risk measures induce risk-averse (respectively -seeking) behavior.

3. The valuation function defined in Definition 2.6 needs not to be risk-averse, which is required in the definition of risk measures. This is due to the fact that humans are not always economically rational, as suggested by prospect theory. In this sense, our definition is more general than the one employed in risk measure theory.

4. It is also remarkable that our definition of coherency is different from the usual definition applied in risk measure theory, where coherency equals "convex + homogeneous".

2.3.3 The linear space $\mathscr{L}$

Bounded spaces In the seminal paper by Artzner et al. (1999), coherent risk measures are defined on finite state spaces. Therefore $\mathscr{L} = \mathbb{R}^d$. In a subsequent paper by Delbaen (2000), the result is generalized to $\mathscr{L} = L^\infty(\Omega, \mathcal{F}, \mu)$, i.e., the space of all μ-almost surely bounded $\mathcal{F}$-measurable real-valued functions, with a fixed probability measure μ. Under the same setting, Föllmer and Schied (2002) introduced the concept of convex risk measures. This setting has some nice properties.

Proposition **2.11.** *Let ρ be a valuation function. Fix a probability measure $\mu \in \mathscr{P}$ and let $\mathscr{L} = L^\infty(\Omega, \mathcal{F}, \mu)$. Assume further that the partial ordering on $\mathscr{L}$ is understood in μ-almost sure (a.s.) sense. Then*

$$\mu\text{-}essinf\, v \le \rho(v, \mu) \le \mu\text{-}esssup\, v, \quad \forall v \in \mathscr{L}.$$

Proof. We show only the first inequality. The second one can be obtained similarly. By definition, we have

$$v \ge \mu\text{-essinf}\, v =: \underline{v}, \mu\text{-a.s.}$$

Hence, by Axiom (I) and (II) in Definition 2.6, we have

$$\rho(v, \mu) \ge \rho(\underline{v}, \mu) = \rho(0, \mu) + \underline{v}.$$

Finally, by Axiom (III), we obtain the required the inequality. □

If we take furthermore a even more restrictive set

$$\mathscr{L} = L^\infty(\Omega, \mathcal{F}) := \left\{ f : \Omega \to \mathbb{R} \text{ is } \mathcal{F} \text{ measurable} \,\middle|\, \sup_{\omega \in \Omega} |f(\omega)| < \infty \right\},$$

which is apparently a subset of $L^\infty(\Omega, \mathcal{F}, \mu), \forall \mu \in \mathscr{P}$, then we obtain immediately the following corollary which is independent of the choice of μ.

Corollary **2.12.** *Let ρ be a valuation function and $\mathscr{L} = L^\infty(\Omega, \mathcal{F})$. Then*

$$\inf_{\omega \in \Omega} v(\omega) \le \rho(v, \mu) \le \sup_{\omega \in \Omega} v(\omega), \forall \mu \in \mathscr{P}.$$

The above proposition and corollary show that two "rational" constraints are automatically assumed upon the subjective evaluation, viz., it cannot be higher than the largest possible outcome, nor be lower than the smallest possible outcome.

Another important property with the setting $\mathscr{L} = L^\infty(\Omega, \mathcal{F}, \mu)$ is that under some regularity conditions, one can obtain the following dual representation for concave (in other words, risk-averse) evaluation function (see e.g., Schied et al., 2009, Theorem 1.2): there exists a corresponding *minimal penalty function* $\gamma : \mathscr{P} \to \mathbb{R}$ such that

$$\phi(v, \mu) = \min_{\nu \in \mathscr{P}: \nu \ll \mu} \left(\mathbb{E}^\nu[v] + \gamma(\nu) \right),$$

where $\mathbb{E}^\nu[v] = \int_\Omega v(\omega)\nu(d\omega)$, and " $\ll$ " denotes the absolute continuity.

This representation in fact reveals how one can construct new evaluation functions, which is not obvious at all by the original "abstract" axiomatic definition.

L^p spaces Several works, e.g., Delbaen (2000) for coherent risk measures, Svindland (2009b) and Ruszczynski and Shapiro (2006) for convex risk measures, extend the setting to the space $\mathscr{L} = L^p(\Omega, \mathcal{F}, \mu), p \in [1, \infty)$ and obtain similar dual representations.

Orlicz spaces Another extension is to use Orlicz hearts and the dual representation is obtained on the corresponding Orlicz space. For more details see Cheridito and Li (2009). This setting is useful, e.g., for entropic measures (see Subsection 2.4.2 below).

Remark 2.13. In all literature of convex/coherent risk measures mentioned above, the dependence of risk measures on the probability measure μ is usually not explicitly stated in axioms. Instead, it is implicitly implicated by the dependence of $\mathscr{L}$ on μ. In our case, however, we shall apply $\mathscr{L} = L^\infty(\Omega, \mathcal{F})$ and its generalization (see Subsection 4.2.1), which are independent of μ. Hence, we include μ in the three axioms explicitly.

2.4 Counter- and examples

We start first with counterexamples (Section 2.4.1), followed by examples (Section 2.4.2–2.4.7).

2.4.1 Counterexamples

Expected utility

ρ^u defined in Subsection 2.2.1 is in general not a valuation function, since it does not satisfy the axiom of translation invariance, except the trivial case $u(x) = x$. Due to the same reason, nor is the ρ defined in Subsection 2.2.2 for *prospect theory* a valuation function. Nevertheless, if we apply directly the utility $u \circ v$, instead of its objective outcome function v, then it becomes the standard linear expectation for the expected utility and the Choquet integral (Choquet, 1953) for prospect

theory. Both of them become, therefore, a special case of valuation functions. We will explain Choquet integral in more details in Subsection 2.4.4 below, for it is the essential progress made by prospect theory, comparing with the expected utility theory.

Mean-variance trade-off

Since Markowitz's pioneer paper (Markowitz, 1952) on portfolio selection, variance (or standard deviation) has been widely used to measure risk. Hence, the objective is to maximize the trade-off between mean and variance (or standard deviation)

$$\begin{aligned}\rho(v,\mu) &:= \mathbb{E}^{\mu}[v] - \lambda \mathrm{Var}^{\mu}(v)\\ \text{or } \rho(v,\mu) &:= \mathbb{E}^{\mu}[v] - \lambda\sqrt{\mathrm{Var}^{\mu}(v)},\end{aligned}$$

where λ control the degree of risk-sensitivity. If $\lambda > 0$, then it induces risk-averse behavior, since variance (risk) is minimized to certain degree. Conversely, $\lambda < 0$ induces risk-seeking behavior. It is easy to check that the axiom of translation invariance is satisfied, whereas the axiom of monotonicity may not be satisfied. Hence, the mean-variance trade-off is, in general, not a valid valuation function. To overcome this problem but meanwhile to keep the idea of measuring risk with 2nd order (or even higher order) statistics, the mean-semideviation trade-off is introduced by Ogryczak and Ruszczyński (1999). The details will be explained in Subsection 2.4.7.

2.4.2 Entropic measure

Although its name is taken from a rather recent paper by Föllmer and Schied (2002), as a measure of risk in the framework of Markov decision processes, the entropic measure can be traced back to Howard and Matheson (1972) under the name of exponential utility, and as a dual form of relative entropy (see (2.11) below), it can even date back to Kullback and Leibler (1951). A more detailed review of literature in MDPs and control theory will be presented in Chapter 5. The entropic measure is defined as follows.

$$\rho^{\lambda}(v,\mu) := \frac{1}{\lambda}\log\left\{\int_{\Omega} e^{\lambda v} d\mu\right\}, \lambda \neq 0 \in \mathbb{R}. \tag{2.9}$$

Here, we assume that $\int e^{\lambda v} d\mu < \infty$. It is easy to check that ρ is a valid valuation function satisfying the three axioms.

The risk sensitivity is controlled by λ. If $\lambda > 0$, then it is easy to check that ρ is convex and therefore is risk-seeking. Conversely, negative λ yields concave and risk-averse ρ.

The entropic measure has several nice properties.

- Expanding ρ w.r.t. λ leads to

$$\rho^{\lambda}(v,\mu) = \mathbb{E}^{\mu}[v] + \frac{\lambda}{2}\mathrm{Var}^{\mu}[v] + O(\lambda^2).$$

  Hence, it can be also viewed as an approximation of mean-variance trade-off when λ is close to 0. If $\lambda > 0$, the risk measured by variance is "liked" and therefore it induces risk-seeking behavior. Conversely, the risk is "disliked", i.e., risk-averse for negative λ. This judgment of risk-preference is consistent with the risk-preference judged by convexity/concavity as mentioned above.

- Assume further that $v \in L^{\infty}(\Omega,\mathcal{F},\mu)$. Then, one can show that (for a proof see e.g. Coraluppi and Marcus, 2000)

$$\begin{aligned} \lim_{\lambda\to 0}\rho^{\lambda}(v,\mu) &= \mathbb{E}^{\mu}[v] \\ \lim_{\lambda\to\infty}\rho^{\lambda}(v,\mu) &= \mu\text{-esssup } v =: \bar{v} \\ \lim_{\lambda\to-\infty}\rho^{\lambda}(v,\mu) &= \mu\text{-essinf } v =: \underline{v} \end{aligned} \tag{2.10}$$

  These limit results show that by controlling λ, we can arrive at any point in the whole range $[\underline{v},\bar{v}]$, i.e., the largest range can be covered by any valuation function on $L^{\infty}(\Omega,\mathcal{F},\mu)$ (see Proposition 2.11).

- ρ^{λ} has the following dual representation (for a proof see Föllmer and Schied (2002) on $L^{\infty}(\Omega,\mathcal{F},\mu)$ and Cheridito and Li (2009) on Orlicz hearts)

$$\rho^{\lambda}(v,\mu) = \begin{cases} \max\limits_{\nu\in\mathscr{P}:\nu\ll\mu}\left(\mathbb{E}^{\nu}[v] - \frac{1}{\lambda}H(\nu|\mu)\right), & \lambda > 0 \\ \min\limits_{\nu\in\mathscr{P}:\nu\ll\mu}\left(\mathbb{E}^{\nu}[v] - \frac{1}{\lambda}H(\nu|\mu)\right), & \lambda < 0 \end{cases},$$

$$\text{with } H(\nu|\mu) := \begin{cases} \int d\nu \log\frac{d\nu}{d\mu} & \text{if } \nu \ll \mu \\ +\infty & \text{otherwise} \end{cases}. \tag{2.11}$$

  It is also easy to check that the optimal ν for both positive and negative λ is attained at ν^* satisfying

$$d\nu^* = \frac{e^{\lambda v}d\mu}{\int e^{\lambda v}d\mu}.$$

2.4.3 Robust control

Iyengar (2005) introduced the framework of *robust dynamic programming*, by which he argues that in some applications the probability measure μ cannot be inferred exactly. Instead, he employs a set of probability measures, $\mathcal{Q}$, which

contains all possible "ambiguous" probability measures. In order to gain the "robustness", the worst case is considered, adapted in our framework, i.e.,

$$\rho(v) := \inf_{\nu \in \mathcal{Q}} \mathbb{E}^{\nu}[v]. \tag{2.12}$$

We can verify that ρ is everywhere concave and therefore risk-averse, which coincides the intuition that the worst scenario is considered. One special case of the robust dynamic programming was the *minimax control* (see e.g. Coraluppi and Marcus, 2000), which also considers the worst scenario with finite state space, $\rho(v) := \max_{\mu(y)>0} v(y)$. By (2.10), this can also be viewed as an extreme of the entropic measure by letting $\lambda \to -\infty$.

It is also notable that each concave and homogeneous valuation function has one dual presentation of the form (2.12) under some regularity conditions for the set $\mathcal{Q}$, see e.g. Delbaen (2000) for essentially bounded spaces and Svindland (2009a) for unbounded ones.

2.4.4 Choquet integral

We first introduce the concept of *nonadditive measures* (Denneberg, 1994) (also called *capacities* in Choquet (1953)) that generalize standard probability measures. A set function $\nu : \mathcal{F} \to [0,1]$ is called a nonadditive measure if

a) $\nu(\mathsf{A}) \leq \nu(\mathsf{B})$, whenever $\mathsf{A} \subset \mathsf{B} \in \mathcal{F}$

b) $\nu(\emptyset) = 0$ and $\nu(\Omega) = 1$.

Remark 2.14. Given a valuation function ρ, we can construct a nonadditive measure in the following way,

$$\nu(\mathsf{B}) := \rho(\mathbf{1}_{\mathsf{B}}, \mu), \mathsf{B} \in \mathcal{F},$$

where $1 : \Omega \to [0,1]$ denotes the indicator function, i.e.,

$$1_{\mathsf{B}}(\omega) := \begin{cases} 1, & \omega \in \mathsf{B} \\ 0, & \text{otherwise} \end{cases}. \tag{2.13}$$

Given an $\mathcal{F}$-measurable function $v : \Omega \to \mathbb{R}$ and a nonadditive measure ν, the *Choquet integral* is defined as

$$\rho(v, \nu) = \int_{\Omega}^{Ch} v(\omega)\nu(d\omega) := \int_{-\infty}^{0} [\nu(v > t) - 1]dt + \int_{0}^{\infty} \nu(v > t)dt.$$

Here, $\nu(v > t) := \nu(\{\omega | v(\omega) > t\})$. It can be shown that (Denneberg, 1994, Proposition 5.1) $\rho(v, \nu)$ satisfies the three axioms of valuation functions, and furthermore ρ is homogeneous. However, ρ is not necessarily concave nor convex. It means that the corresponding behavior is not always risk-averse nor risk-seeking, which makes it quite suitable for modeling human behavior.

This integral can be viewed as a continuous-state version of the model applied in prospect theory (see Subsection 2.2.2), by replacing the objective outcome v with its utility $u \circ v$. The level set $\nu(v > t)$ reflects exactly the rank-dependent probability applied in prospect theory. It is also remarkable that in $\rho(v, \nu)$, the objective probability measure μ is replaced by the *subjective* nonadditive measure ν, though the dependence between these two measures are not specified here. In prospect theory, it is usually assumed (Tversky and Kahneman, 1992) that ν is obtained by a nonlinear transformation of μ with some parameters that differ individually.

2.4.5 Utility-based shortfall

The entropic measure introduced in Section 2.4.2 belongs, in fact, to a large family of valuation functions called *utility-based shortfall* (see Föllmer and Schied, 2004, Section 4.6 and Schied et al., 2009, Section 2) defined as follows. Let $u : \mathbb{R} \to \mathbb{R}$ be a continuous, increasing and non-constant utility function satisfying $u(0) = 0$. Assume that there exists a constant $m \in \mathbb{R}$ such that $\int_\Omega u(v(\omega) - m)\mu(d\omega) < \infty$. Then,

$$\rho^u(v, \mu) := \sup \left\{ m \in \mathbb{R} \mid \int_\Omega u(v(\omega) - m)\mu(d\omega) \geq 0 \right\} \tag{2.14}$$

defines a utility-based shortfall. It is easy to check that ρ^u satisfies the three axioms and therefore is a valid valuation function. It is remarkable that comparing with its original definition in Föllmer and Schied (2004), we fix here the reference level to be 0 on the right-hand side of (2.14), which is due to the axiom of centralization.

Comparing with the expected utility theory, the utility function in (2.14) is applied to the relative value $v(\omega) - m$ rather than to the absolute outcome $v(\omega)$. For modeling human decisions, this reflects the intuition that human beings judge utilities usually by comparing those outcome with a reference value.

Optimality conditions

We show below in Proposition 2.15 that under some technical assumptions, the optimal m^* to the maximization problem in the definition of utility-based shortfall (2.14) satisfies, in fact, the following stochastic equation

$$\mathbb{E}^\mu \left[u(v - m^*) \right] = \int_\Omega u(v(\omega) - m^*)\mu(d\omega) = 0.$$

In other words, we have the following equation

$$\mathbb{E}^\mu \left[u\left(v - \rho^u(v, \mu) \right) \right] = 0.$$

This property enables us to implement *stochastic approximation* Kushner and Yin (2003) algorithms for approximating the shortfall $\rho^u(v, \mu)$ (see Dunkel and Weber (2010) for one-dimensional cases and Chapter 6 for multidimensional cases).

Proposition **2.15** (cf. Föllmer and Schied, 2004, Proposition 4.104)**.** *Let ρ^u be a shortfall defined in* (2.14) *with a continuous and strictly increasing utility function u which satisfies $u(0) = 0$. Then the following statements are equivalent:*

(i) $\rho^u(v, \mu) = m^*$ *and*

(ii) $\mathbb{E}^\mu[u(X - m^*)] = 0$.

Proof. (ii) $\Rightarrow$ (i). By definition, $m^* \leq \rho^u(v, \mu)$. For any $\epsilon > 0$, since u is strictly increasing, we have $u(v(\omega) - m^* - \epsilon) < u(v(\omega) - m^*), \forall \omega \in \Omega$, which implies $\mathbb{E}^\mu [u(v - m^* - \epsilon)] < \mathbb{E}^\mu u(v - m^*) = 0$. Hence, $m^* = \rho^u(v, \mu)$.

(i) $\Rightarrow$ (ii). By definition we have $\mathbb{E}^\mu [u(v - m^*)] \geq 0$. Assume that $\mathbb{E}^\mu[u(v - m^*)] > 0$. By the continuity of u, there exists an $\epsilon > 0$ such that

$$\mathbb{E}^\mu [u(v - m^* - \epsilon)] > 0,$$

which implies $\rho^u(v, \mu) \geq m^* + \epsilon > m^*$ and hence contradicts (i). Thus, (ii) holds. □

Convexity and concavity

The following proposition shows that the property of u being convex or concave determines the risk sensitivity of ρ^u.

Proposition **2.16.** *Given a concave function u, ρ is also concave (and hence risk-averse). Vice versa, ρ is convex (hence risk-seeking) for convex u.*

Proof. We prove only the convex case. The concave case can be obtained similarly. Let $m_f := \rho^u(f, \mu)$ and $m_g := \rho^u(g, \mu)$. Hence, by Proposition 2.15, we have

$$\mathbb{E}^\mu[u(f - m_f)] = 0 \text{ and } \mathbb{E}^\mu[u(g - m_g)] = 0.$$

On the one hand, due to the convexity of u, we have for each $\alpha \in [0, 1]$

$$\begin{aligned} 0 =& \alpha \mathbb{E}^\mu[u(f - m_f)] + (1 - \alpha)\mathbb{E}^\mu[u(g - m_g)] \\ \geq& \mathbb{E}^\mu[u(\alpha f + (1 - \alpha)g - \alpha m_f - (1 - \alpha)m_g)]. \end{aligned}$$

On the other hand, $m^* := \rho^u(\alpha f + (1 - \alpha)g, \mu)$ satisfies

$$\mathbb{E}^\mu[u(\alpha f + (1 - \alpha)g - m^*)] \geq 0.$$

Hence, combing above two inequalities and using the monotonicity of u yield the required inequality $m^* \leq \alpha m_f + (1 - \alpha)m_g$. □

Remark 2.17. The convex case of the proposition can be also implied by Proposition 4.61 in Föllmer and Schied (2004), which utilizes a dual representation of the utility based shortfall. Above we provide, however, a more concise proof for both convex and concave cases.

Examples

Utility-based shortfalls cover a large family of valuation functions, which have been proposed in literature of various fields.

1) $u(x) = x$ yields the standard expectation $\rho^u(v, \mu) = \mathbb{E}^\mu[v]$.

2) $u(x) = e^{\lambda x} - 1$ yields the entropic measure defined in (2.9).

3) Mihatsch and Neuneier proposed in Mihatsch and Neuneier (2002) the following setting

$$u(x) = \begin{cases} (1-\kappa)x & \text{if } x > 0 \\ (1+\kappa)x & \text{if } x \leq 0 \end{cases}, \tag{2.15}$$

where $\kappa \in (-1, 1)$ controls the degree of risk sensitivity. Its sign determines the property of the utility function u being convex vs. concave and, therefore, the risk-preference of ρ.

When quantifying human behavior, combined convex/concave utility functions, e.g.,

$$u_p(x) = \begin{cases} k_+ x^{l_+} & x \geq 0 \\ -k_-(-x)^{l_-} & x < 0 \end{cases}, \tag{2.16}$$

are of special interest, since people tend to treat gains and losses differently and, therefore, have different risk preferences on gain and loss sides. In fact, the polynomial function in (2.16) was used in the prospect theory (Kahneman and Tversky, 1979) to model human risk preferences and the results show that l_+ is usually below 1, i.e., $u_p(x)$ is concave and thus risk-averse on gains, while l_- is also below 1 and $u_p(x)$ is therefore convex and risk-seeking on losses.

Relation to prospect theory

To illustrate the risk-preferences induced by different utility functions, we consider a simple example with two events. The first event has outcome v_1 with probability p, while the other event has smaller outcome $v_2 < v_1$ with $1 - p$. Note that

$$p = \frac{\mathbb{E}[v] - v_2}{v_1 - v_2}, \text{ where } \mathbb{E}[v] = pv_1 + (1-p)v_2$$

denotes the risk-neutral mean.

Replacing $\mathbb{E}[v]$ with the valuation function $\rho(v, p)$, we can define a *subjective probability* (cf. Tversky and Kahneman, 1992) as

$$w(p) := \frac{\rho(v, p) - v_2}{v_1 - v_2}, \tag{2.17}$$

which measures agents' subjective perception of the true probability p.

In risk-neutral cases, $\rho(v,p)$ is simply the mean and $w(p) = p$. In risk-averse cases, the balance moves towards the worst scenario. Hence, the probability of the first event (with larger outcome x_1) is always underestimated. On the contrary, in risk-seeking cases, the probability of the first event is always overestimated. Behavioral studies show that human subjects usually overestimate low probabilities and underestimate high probabilities (Tversky and Kahneman, 1992). This can be quantified by applying mixed valuation functions ρ. If we apply utility-based shortfalls, it can be quantified by using mixed utility function u.

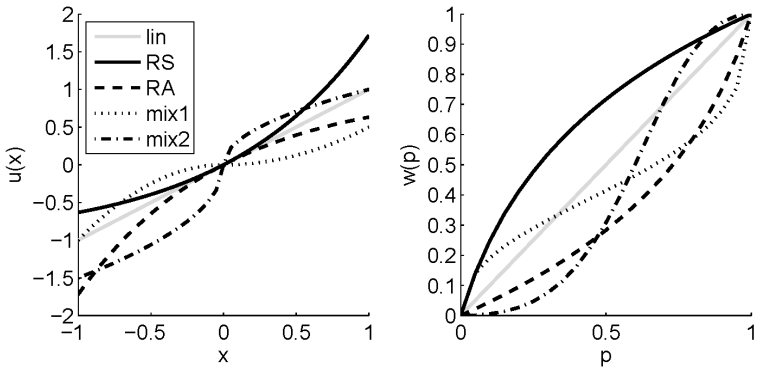


Figure 2.1: Shortfalls with different utility functions and induced subjective probabilities. (Left) utility functions defined as follows: lin : x; RS : $e^x - 1$; RA : $1 - e^{-x}$; mix1: $u_p(x)$ as defined in (2.16) with $k_+ = 0.5$, $l_+ = 2$, $k_- = 1$ and $l_- = 2$; mix2: same as mix1 but with $k_+ = 1$, $l_+ = 0.5$, $k_- = 1.5$ and $l_- = 0.5$. (Right) subjective probability functions calculated according to (2.17).

Let $v_1 = 1$ and $v_2 = -1$. Figure 2.1 (left) shows five different utility functions, one linear function "lin", one convex function "RS", one concave function "RA", and two mixed functions "mix1" and "mix2" (for details see caption). The corresponding subjective probabilities are shown in Figure 2.1 (right). Since the function "RA" is concave, the corresponding valuation function is risk-averse and therefore the probability of high-reward event is always underestimated. For the case of the convex function "RS", the probability of high-reward event is always overestimated. However, since the "mix1" function is convex on $[0,\infty)$ but concave on $(-\infty,0]$, high probabilities are underestimated while low probabilities are overestimated, which replicates very well the probability weighting function applied in prospect theory for gains (cf. Tversky and Kahneman, 1992, Figure 1). Conversely, the "mix2" function, which is concave on $[0,\infty)$ and convex on $(-\infty,0]$, corresponds to the overestimation of high probabilities and the underestimation of low probabilities. This corresponds to the weighting function used for losses in prospect theory (cf. Tversky and Kahneman, 1992, Figure 2).

We will see in Section 6.3 that the advantage of using the utility-based shortfall is that we can derive iterating learning algorithms for the estimation of the subjective valuations, whereas it is difficult to derive such algorithms in the framework

of prospect theory.

2.4.6 Optimized certainty equivalence

Let $u : \mathbb{R} \to \mathbb{R}$ be a continuous non-decreasing function satisfying there exists an $m \in \mathbb{R}$ such that

$$u(m) = m \text{ and } u(x) \leq x, \forall x \in \mathbb{R}. \tag{2.18}$$

Then

$$\rho(v, \mu) := \sup_{m \in \mathbb{R}} \{\mathbb{E}^{\mu}[u(v - m)] + m\} \tag{2.19}$$

defines an *optimized certainty equivalence*, which was initially introduced by Ben-Tal and Teboulle in Ben-Tal and Teboulle (1987) on finite spaces and extended to general spaces in Ben-Tal and Teboulle (2007); Schied (2007). It is easy to check that ρ defined in the above equation satisfies the three axioms of valuation functions.

If v is bounded and u is concave, then it has a nice dual representation (for a proof see, e.g., Ben-Tal and Teboulle (2007), mainly due to Young's inequality and Legendre-Fenchel transformation)

$$\rho(v, \mu) = \inf_{\nu \ll \mu} \left(\mathbb{E}^{\nu}[v] + \mathbb{E}^{\mu}\left[g\left(\frac{d\nu}{d\mu} \right) \right] \right) \tag{2.20}$$

where $g : [0, \infty) \to (-\infty, \infty]$ is defined as

$$g(z) := \sup_{x \in \mathbb{R}} (u(x) - xz) .$$

The second item, $\mathbb{E}^{\mu}\left[g\left(\frac{d\nu}{d\mu} \right) \right]$, on the right-hand side of (2.20) is also called *g-divergence* in statistics literature Csiszar (1967).

Remark 2.18. The constraint we set in (2.18) is to ensure that ρ is centralized, i.e., $\rho(0, \mu) = 0$, which is not required in the its original definition in Ben-Tal and Teboulle (2007).

Examples

The optimized certainty equivalence covers several important examples.

1) $u(x) = x$ gives the standard expectation $\rho(v, \mu) = \mathbb{E}^{\mu}[v]$.

2) $u(x) = \min(\frac{x}{\lambda}, 0)$, $\lambda \in (0, 1)$, yields a coherent valuation function, called *average value at risk, expected shortfall, conditional value at risk or tail value at risk* in finance literature (see e.g. Rockafellar and Uryasev (2000); Schied et al. (2009) and references therein):

$$-\rho(v, \mu) = -\sup_{m \in \mathbb{R}} \{\mathbb{E}^{\mu}[u(v - m)] + m\} = \frac{1}{\lambda} \inf_{m \in \mathbb{R}} (\mathbb{E}^{\mu}[(m - v)^{+}] - \lambda m) .$$

Note that for any valuation function ρ, its associated risk measure is $-\rho$. See remarks in Section 2.3.2.

3) $u(x) = -\frac{1}{\lambda}e^{-\lambda x} + \frac{1}{\lambda}$, $\lambda > 0$, gives again the entropic measure defined in Section 2.4.2. To see this, first, it is easy to check that u satisfies (2.18). Since u is differentiable, with the derivative $u'(x) = e^{-\lambda x}$, the optimal m^* to (2.19) satisfies $\mathbb{E}^\mu[u'(v - m^*)] = 1$, viz.,

$$\mathbb{E}^\mu[e^{-\lambda(v-m^*)}] = 1 \quad \Rightarrow \quad m^* = -\frac{1}{\lambda}\mathbb{E}^\mu[e^{-\lambda v}].$$

Hence, $\rho(v,\mu) = -\frac{1}{\lambda}\mathbb{E}^\mu[e^{-\lambda(v-m^*)}] + \frac{1}{\lambda} + m^* = -\frac{1}{\lambda}\mathbb{E}^\mu[e^{-\lambda v}]$.

Optimality conditions

Let us consider only concave u which induces risk-averse behaviors. Assume further that u is differentiable and denote by u' its first derivative. Then the optimal m^* to the optimization problem in the definition of optimized certainty equivalence (2.19) satisfies

$$\mathbb{E}^\mu\left[u'(v - m^*)\right] = 1,$$

which is an stochastic equation. Hence, similar to the utility-based shortfall introduced in the last subsection, we can apply stochastic approximation algorithms to solve the above equation and meanwhile calculate the original expectation

$$\mathbb{E}^\mu\left[u(v - m^*)\right] + m^*.$$

For more details see Hamm et al. (2013).

2.4.7 Mean-semideviation trade-off

As we have discussed in Section 2.4.1, the mean-variance trade-off is not a valid valuation function, since it violates the axiom of monotonicity. To amend this problem, Ogryczak and Ruszczyński (1999) (see also Ruszczynski and Shapiro, 2006) introduced the concept of mean-semideviation trade-off:

$$\rho(v,\mu) := \mathbb{E}^\mu[v] - \lambda\left[\mathbb{E}^\mu(\mathbb{E}^\mu[v] - v)_+^r\right]^{1/r}. \tag{2.21}$$

Here $r \geq 1$, $(x)_+ := \max(x, 0)$, and $\lambda \in [0, 1]$ controls how risk-averse the agent is. Comparing with the mean-variance trade-off, the risk is measured by the semideviation $\left[\mathbb{E}^\mu(\mathbb{E}^\mu[v] - v)_+^r\right]^{1/r}$ instead of the standard deviation $\left[\mathbb{E}^\mu|\mathbb{E}^\mu[v] - v|^r\right]^{1/r}$. It is called "semi", because only those events whose outcomes below the mean $\mathbb{E}^\mu[v]$ are regarded as "risky", whereas those events with outcomes higher than the mean are viewed as "safe" events. Hence, using variance or standard deviation as a risk measure punishes both positive and negative sides (taking the mean as the reference point), while the semideviation punishes only the negative side.

For more detailed comparison, we refer to Ogryczak and Ruszczyński (1999). In addition, it is also easy to check that $\rho(v,\mu)$ is concave with respect to v and is therefore risk-averse. This categorization (due to its concavity) is consistent with the intuition that risk, which is measured by the semideviation, is punished.

Monotonicity

It is easy to check that ρ defined in (2.21) satisfies the axioms of translation invariance and centralization. Below we show that it satisfies also the axiom of monotonicity. It is remarkable that our proof is based on the calculation of gradients of ρ with respect to v, which will be used in Section 4.3.1 as well. For its original proof, see Proposition 2 of Ogryczak and Ruszczyński (1999).

Fix a probability measure $\mu \in \mathscr{P}$ and let $L^r(\mu)$ be the space of all real-valued $\mathcal{F}$-measurable functions with a finite rth moment, i.e., $\mathbb{E}^\mu[|v|^r] < \infty$.

Proposition **2.19.** *Let ρ be defined in (2.21) with some $\lambda \in [0,1]$ and $r \geq 1$. Then, for each probability measure $\mu \in \mathscr{P}$ and $f \geq g \in L^r(\mu)$, $\rho(f,\mu) \geq \rho(g,\mu)$.*

Proof. We consider v as a differentiable function of $t \in [0,1]$, i.e., $v : [0,1] \to L^r(\mu)$, satisfying $v(0) = f$ and $v(1) = g$. Then, it is sufficient to show that $\frac{d}{dt}\rho(v,\mu) \geq 0$, provided $\dot{v}(t) \geq 0$ for each $t \in [0,1]$.

Let $u(x) = (x)_+$ and therefore its right derivative is

$$u'(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

which implies that $u(x)^p u'(x) = u(x)^p$ holds for all $x \in \mathbb{R}$ and $p > 0$.

If $v(t)$ is a constant function, then $\frac{d}{dt}\rho(v,\mu) = \mathbb{E}^\mu[\dot{v}] \geq 0$ already holds. Otherwise, we consider first the case (i) $r > 1$. Since $\dot{v} \geq 0$ and $\lambda \in [0,1]$, we have

$$\begin{aligned}
\frac{d}{dt}\rho(v,\mu) =& \mathbb{E}^\mu[\dot{v}] - \lambda \frac{\mathbb{E}^\mu\left[[u(\mathbb{E}^\mu[v]-v)]^{r-1} u'(\mathbb{E}^\mu[v]-v)\,(E^\mu[\dot{v}]-\dot{v})\right]}{\left(\mathbb{E}^\mu\left[u(\mathbb{E}^\mu[v]-v)\right]^r\right)^{1-1/r}} \\
=& \mathbb{E}^\mu[\dot{v}] - \lambda \frac{\mathbb{E}^\mu\left[[u(\mathbb{E}^\mu[v]-v)]^{r-1} (E^\mu[\dot{v}]-\dot{v})\right]}{\left(\mathbb{E}^\mu\left[u(\mathbb{E}^\mu[v]-v)\right]^r\right)^{1-1/r}} \qquad (2.22) \\
\geq& \mathbb{E}^\mu[\dot{v}]\left(1 - \lambda \frac{\mathbb{E}^\mu\left[[u(\mathbb{E}^\mu[v]-v)]^{r-1}\right]}{\left(\mathbb{E}^\mu\left[u(\mathbb{E}^\mu[v]-v)\right]^r\right)^{\frac{r-1}{r}}}\right).
\end{aligned}$$

Finally, due to Hölder's inequality, we have

$$\left(\mathbb{E}^\mu\left[[u(\mathbb{E}^\mu[v]-v)]^{r-1}\right]\right)^{\frac{1}{r-1}} \leq \left(\mathbb{E}^\mu\left[u(\mathbb{E}^\mu[v]-v)\right]^r\right)^{\frac{1}{r}},$$

which implies that $\frac{d}{dt}\rho(v,\mu) \geq \mathbb{E}^\mu[\dot{v}](1-\lambda) \geq 0$.

(ii) We consider now the case $r = 1$. Then,

$$\begin{aligned}\frac{d}{dt}\rho(v,\mu) =& \mathbb{E}^{\mu}[\dot{v}] - \lambda\mathbb{E}^{\mu}\left[u'(\mathbb{E}^{\mu}[v] - v)\left(\mathbb{E}^{\mu}[\dot{v}] - \dot{v}\right)\right] \\ \geq& \mathbb{E}^{\mu}[\dot{v}]\left(1 - \lambda\mathbb{E}^{\mu}\left[u'(\mathbb{E}^{\mu}[v] - v)\right]\right) \geq 0.\end{aligned}$$

Combining (i) and (ii), our claim follows. □

2.5 Summary

Risk is caused by the uncertainty of future events and their associated outcomes. A risky choice consists of probabilities and outcomes of future events. Different ways of evaluating risky choices result in different risk preferences. We therefore propose in this chapter a normative axiomatic framework of valuation functions as an indirect measure of risk, in the sense that given a valuation function, we can quantify its induced risk preference. Here, the risk preference is not necessarily uniform, i.e., one can be risk-averse in some situations while being risk-seeking in other situations, as observed in studies of human behavior like the prospect theory. We have also shown that most of the examples in risk-related literature of various fields can be covered by our framework. This framework makes, therefore, a broad and solid premise for our further development of models on risk-sensitive sequential decision-making problems in the following chapters.

Mathematical properties of each example has been briefly studied and compared. In particular, we find that a rich family of valuation functions, called utility-based shortfall, can replicate the key features predicted by prospect theory, if the utility function is appropriately chosen. Furthermore, the derived optimization problem reduces to solving a stochastic equation. These two features will play essential roles when we develop risk-sensitive reinforcement learning algorithms to quantify human behaviors (see Chapter 6 and 7).

3

Valuation Maps

Life can only be understood backwards;
but it must be lived forwards.
— Søren A. Kierkegaard

Précis From now on we assume that the underlying stochastic process is Markovian. For simplicity, we consider in this chapter only Markov chains without control variables. The extension to Markov decision/control problems will be investigated in Chapter 5. Applying the theory developed in the last chapter to Markov chains, we ought to answer the following two questions:

1. How to "customize" valuation functions in order to remain important properties, e.g., the Markov property?
2. Is our definition consistent with the originally more general definition in the sense that convexity (respectively concavity) induces risk-seeking (respectively risk-averse) behavior?

Our answer to these questions is to apply a *constructive* approach by introducing the concept of *valuation maps* (see Section 3.2 below). This approach admits a backward induction (or dynamic programming) for calculating the subjective valuation of additive rewards on Markov chains. In Section 3.3, we will also show that our approach is compatible with the literature (see e.g. Cheridito et al., 2006; Ruszczyński, 2010) where time-consistent dynamic risk measures are applied.

3.1 Markov property

Let $\{X_t, t = 0, 1, \ldots\}$ be a Markov chain on a measurable space $(\mathsf{X}, \mathcal{B})$. Given a family of real-valued $\mathcal{B}$-measurable functions $\{r_t\}_{t=0}^{T}$, we consider the following sum

$$S_T := \sum_{t=0}^{T} r_t(X_t), \tag{3.1}$$

and apply valuation functions defined in the last chapter to this sum.

To accomplish this task, the first step is to define the Markov chain in Kolmogorov's style (see, e.g., Section 2.2 of Hernández-Lerma and Lasserre, 2003). That means, let $(\Omega, \mathcal{F})$ be the (canonical) sample space with $\Omega := \mathsf{X}^\infty$ and $\mathcal{F}$ being the associated product σ-algebra. An element $\omega \in \Omega$ is a sequence $(x_0, x_1, \ldots)$ with components $x_t \in \mathsf{X}$. Let ν be a probability measure on $\mathcal{B}$, and for each $t = 0, 1, \ldots$, let $X_t : \Omega \to \mathsf{X}$ be the projection $\omega \mapsto X_t(\omega) := x_t$. Then by the well-known theorem of I. Ionescu Tulcea, there is a probability measure $\mathbb{P}_\mu$ on $\mathcal{F}$ such that $\mathbb{P}_\mu(X_0 \in \mathsf{B}) = \mu(\mathsf{B}), \forall \mathsf{B} \in \mathcal{B}$, and, moreover, for every $t = 0, 1, \ldots, x \in \mathsf{X}$ and $\mathsf{B} \in \mathcal{B}$, $\mathbb{P}_\mu(X_{t+1} \in \mathsf{B} | X_t = x) =: P^{(t)}(x, \mathsf{B})$. Note that if the Markov chain is *time-homogeneous*, then $P^{(t)} \equiv P$ is independent of time t. For $t = 0, 1, \ldots$, $P^{(t)}(x, \mathsf{B})$ is a *stochastic kernel* on X, which is defined as follows

Definition **3.1.** *A mapping $P : \mathsf{X} \times \mathcal{B} \to [0, 1]$ is called a* stochastic kernel *if*

(i) $P(x, \cdot)$ is a probability measure on B for each fixed $x \in \mathsf{X}$, and

(ii) $P(\cdot, \mathsf{B})$ is a measurable function on X for each fixed $\mathsf{B} \in \mathcal{B}$.

We also write $P(x, \mathsf{B})$ as $P_x(\mathsf{B})$ or $P(\mathsf{B}|x)$ in different contexts.

Note that since $P_x(\cdot)$ is a probability measure, we can define the following *conditional expectation* for any real-valued $\mathcal{B}$-measurable function f

$$\mathbb{E}^{P_x}[f] = P_x(f) := \int f(y) P(dy|x).$$

As a special example of valuation functions, we apply the standard expectation to the sum S_T defined in (3.1). By the Markov property, it can be expanded as follows,

$$\begin{aligned}\mathbb{E}^{\mathbb{P}_\mu}[S_T] =& \mathbb{E}^{\mu}\left[r_0(X_0) + \mathbb{E}^{P^{(0)}_{X_0}}\left[r_1(X_1) + \ldots + \mathbb{E}^{P^{(T-1)}_{X_{T-1}}}\left[r_T(X_T)\right]\ldots\right]\right] \\ =& \mu\left[r_0 + P^{(0)}\left[r_1 + \ldots + P^{(T-1)}[r_T]\ldots\right]\right].\end{aligned}$$

Moreover, if applying the Dirac measure at $x \in \mathsf{X}$, we denote $\mathbb{P}_\mu$ by $\mathbb{P}_x$. Then the expectation from each start point x becomes

$$\mathbb{E}^{\mathbb{P}_x}[S_T] = r_0(x) + P^{(0)}_x\left[r_1 + \ldots + P^{(T)}[r_T]\ldots\right] \tag{3.2}$$

Hence, one can further calculate the expectation $\mathbb{E}^{\mathbb{P}_x}[S_T]$ iteratively with a family of operators $\mathcal{T}_n : \mathscr{L} \to \mathscr{L}$, $n = 0, 1, 2, \ldots$, defined as

$$\mathcal{T}_n(f) := r_n + P^{(n)}[f]. \tag{3.3}$$

We consider the following iteration,

$$f_T := r_T, f_n := \mathcal{T}_n(f_{n+1}), n = T-1, \ldots, 0. \tag{3.4}$$

Then, it is easy to verify that $\mathbb{E}^{\mathbb{P}_x}[S_T] = f_0(x)$.

The special feature of additive rewards This iterative view reveals the most important feature of the expectation of additive rewards with the Markov property: although the sum S_T depends on the whole history $(X_0, X_1, \ldots, X_T) \in \mathsf{X}^T$, one needs merely to store the function f^n on X and propagate information iteratively as in (3.4). Computationally, if the size of the state space X is N, then the iterative approach reduces the complexity enormously from N^T to $N \times T$. This is the exact "Markov property" that we are going to to keep in our generalization.

3.2 Definition

Hence, we define the new objective function directly from the iterative representation (3.2)

$$\rho[S_T|X_0 = x] := r_0(x) + \mathcal{U}_x^{(0)}\left[r_1 + \ldots + \mathcal{U}^{(T)}[r_T]\ldots\right], \tag{3.5}$$

where the valuation map $\mathcal{U}$ that generalizes the conditional expectation is formally defined as follows with two steps. First, we define the valuation map in a general setting and then restrict to the Markovian setting

Definition **3.2.** *A mapping $\mathcal{U}(x,(v,\mu)) : \mathsf{X}\times\mathscr{L}\times\mathscr{P} \to \mathbb{R}$ is said to be a* valuation map*, if*

(i) for each $x \in \mathsf{X}$, $\mathcal{U}(x,(\cdot,\cdot))$ is a valuation function; and

(ii) $\mathcal{U}(\cdot,(v,\mu))$ is $\mathcal{B}(\mathsf{X})$-measurable for each $(v,\mu) \in \mathscr{L}\times\mathscr{P}$.

Definition **3.3.** *A mapping $\mathcal{U}(x,v) : \mathsf{X}\times\mathscr{L} \to \mathbb{R}$ is said to be a* valuation map *on a stochastic kernel P, if there exists a valuation map $\tilde{\mathcal{U}}$ satisfying*

$$\mathcal{U}(x,v) = \tilde{\mathcal{U}}(x,(v,P_x)) \quad \forall x \in \mathsf{X}, v \in \mathscr{L}.$$

Furthermore, we write $\mathcal{U}(x,v)$ also as $\mathcal{U}_x(v)$ or $\mathcal{U}(v|x)$ depending on different contexts.

Examples Note that all examples valuation functions we have presented in Section 2.4.2–2.4.7 can be easily extended to valuation maps correspondingly by replacing μ with some transition kernel P. For instance, the entropic measure $\rho^\lambda(v,\mu) = \frac{1}{\lambda}\log\left\{\int_\Omega e^{\lambda v}d\mu\right\}$ defined in Section 2.4.2 can be extended to a valuation map, which is said to be an *entropic map*, as follows,

$$\mathcal{U}_x(v) := \frac{1}{\lambda}\log\left\{\int_{\mathsf{X}} e^{\lambda v(y)}P_x(dy)\right\}, \lambda \neq 0.$$

Backward induction For a Markov chain with transition kernels $\{P^{(t)}, t = 0, 1, \ldots\}$, we can define correspondingly a series of valuation maps $\{\mathcal{U}^{(t)}\}$ and generalize the family of operators $\{\mathcal{T}_t\}$ defined in (3.3) to be

$$\mathcal{T}_t(f) := r_t + \mathcal{U}^{(t)}[f]. \tag{3.6}$$

The T-stage valuation function in (3.5) has, therefore, the following iterative representation

$$\rho[S_T|X_0 = x] = f_0(x), \text{ where } f_T := r_T, f_t := \mathcal{T}_t(f_{t+1}), t = T-1, \ldots, 0. \tag{3.7}$$

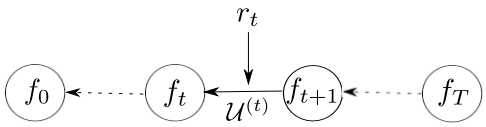


Figure 3.1: Illustration of the backward induction in (3.6).

The backward induction procedure (see also Figure 3.1) can be interpreted as follows. Suppose at time $t+1$, the subjective valuation of all positive future rewards at time $t+2, t+3, \ldots$, is $f_{t+1}(X_{t+1})$ which depends on the state at time $t+1$. Taking one step backwards, i.e., at time t, given the current state $X_t = x$, the successive state X_{t+1} is uncertain, whose law is govern by the transition kernel $P_x^{(t)}$. Facing this uncertainty or risk, agents recalculate their subjective valuation based on the current $X_t = x$ with $f_t(x) = \mathcal{U}_x^{(t)}\left(r(X_{t+1}) + f_{t+1}(X_{t+1})\right)$.

3.3 Time consistency

This backward induction is closely related to the concept of *time-consistency* in the literature of dynamic risk measures (see e.g. Riedel, 2004; Detlefsen and Scandolo, 2005; Ruszczyński, 2010 and references therein). In this section, we follow mostly the notations from Ruszczyński (2010).

Let $(\Omega, \mathcal{F})$ be a Borel space with a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}$. In other words, $\mathcal{F}_t$ denotes the space of all possible sequences of states up to time t. Let $\mathscr{L}_t$ be the space of all real-valued functions which are $\mathcal{F}_t$-measurable, $t = 0, 1, 2, \ldots$. Let $\mathscr{L}_{t,T} := \mathscr{L}_t \times \mathscr{L}_{t+1} \times \ldots \times \mathscr{L}_T$. For instance, by definition, each $R_t := r_t(X_t) \in \mathscr{L}_t$. Hence, $(R_t, R_{t+1}, \ldots, R_T) \in \mathscr{L}_{t,T}$.

Definition **3.4.** *A dynamic valuation function is a family $\{\rho_{t,T}, t = 0, 1, \ldots, T\}$ such that for each $\mu \in \mathscr{P}$, and each $t = 0, 1, \ldots, T$, $\rho_{t,T} : \mathscr{L}_{t,T} \times \mathscr{P} \to \mathscr{L}_t$ satisfies*

(i) (monotonicity) $\rho_{t,T}(f, \mu) \leq \rho_{t,T}(g, \mu)$, $\forall f \leq g \in \mathscr{L}_{t,T}$

(ii) (translation invariance) $\rho_{t,T}(f + g, \mu) = \rho_{t,T}(f, \mu) + g$, $\forall f \in \mathscr{L}_{t,T}$, $g \in \mathscr{L}_t$, and

(iii) (centralization) $\rho_{t,T}(0,\mu) = 0$.

Here, for each $t = 0,1,\ldots,T$, the conditional valuation function $\rho_{t,T}$ can be viewed as a subjective evaluation of future outcomes $R_{t:T} := (R_t, R_{t+1}, \ldots, R_T)$ up to time T. Now suppose there are two outcome functions $R_{0:T}$ and $R'_{0:T}$ such that they have the same outcome at time t, $R_t = R'_t$, and their subjective evaluation at time $t+1$ are also the same, $\rho_{t+1,T}(R_{t+1:T},\mu) = \rho_{t+1,T}(R'_{t+1:T},\mu)$. Then, taking one step backwards, their subjective evaluations at time t should be also identical, i.e., $\rho_{t,T}(R_{t:T},\mu) = \rho_{t,T}(R'_{t:T},\mu)$. This property is called *time consistency*, which is formally defined as follows.

Definition **3.5.** *A dynamic valuation function* $\{\rho_{t,T}, t = 0,1,\ldots,T\}$ *is said to be* time-consistent *if for each* $t = 0,1,\ldots,T-1$, $\mu \in \mathscr{P}$ *and* $R_t, R'_t \in \mathscr{L}_t$,

$$R_t = R'_t \text{ and } \rho_{t+1,T}(R_{t+1:T},\mu) = \rho_{t+1,T}(R'_{t+1:T},\mu)$$

implies that $\rho_{t,T}(R_{t:T},\mu) = \rho_{t,T}(R'_{t:T},\mu)$.

Remark 3.6. Time consistency is the essential property that allows us to apply dynamic programming to solve the optimization problem induced by valuation maps (see Chapter 5). It has been studied in various contexts (see Artzner et al., 2007; Detlefsen and Scandolo, 2005; Koopmans, 1960; Kreps and Porteus, 1978; Ruszczyński, 2010).

For a dynamic valuation function $\{\rho_{t,T}, t = 0,1,\ldots,T\}$ we define

$$\rho_{s,t}((R_s,\ldots,R_t),\mu) := \rho_{s,T}((R_s,\ldots,R_t,0,0,\ldots,0),\mu), 0 \le s \le t \le T.$$

Now we restate Theorem 1 of Ruszczyński (2010) as follows

Theorem **3.7.** *A dynamic valuation function* $\{\rho_{t,T}, t = 0,1,\ldots,T\}$ *is time-consistent if and only if for all* $0 \le s < t \le T$, $\mu \in \mathscr{P}$ *and* $\{R_t, t = 0,1,\ldots,T\} \in \mathscr{L}_{0,T}$ *the following equality holds*

$$\rho_{s,T}(R_{s:T},\mu) = \rho_{s,t}((R_{s:t-1}, \rho_{t,T}(R_{t:T},\mu)),\mu).$$

The above theorem implies immediately that for each $t = 0,1,\ldots,T$,

$$\rho_{t,T}(\sum_{s=t}^{T} R_s) = R_t + \rho_{t,t+1}(R_{t+1} + \rho_{t+1,t+2}(\ldots + \rho_{T-1,T}(R_T)\ldots)),$$

where we omit the fixed probability measure μ. Note that here $\rho_{t,t+1} : \mathscr{L}_{t+1} \times \mathscr{P} \to \mathscr{L}_t$ is the one-step conditional valuation function, which is coincident with our definition of valuation maps, if the underlying stochastic process is Markovian. In other words, given a time-consistent dynamic valuation function, $\{\rho_{t,T}\}$, when applying to the sum S_T defined in (3.1), one can always obtain a backward induction procedure as in (3.5). Hence, the *ad hoc* way that we apply valuation functions to the sum is in fact of no loss of generality, as long as the time consistency is required.

3.4 Time consistency of risk preferences

Before introducing the implied risk preferences, we first state the following definition.

Definition **3.8** (cf. Definition 2.7)**.** *A valuation map $\mathcal{U}$ is said to be* convex *(respectively* concave, homogeneous*) if $\mathcal{U}_x$ is convex (respectively* concave, homogeneous*), for all $x \in \mathsf{X}$.*

Similar to the analysis done in Section 2.3.1, we show by the following proposition that an agent with a convex (respectively concave) valuation map is risk-seeking (respectively risk-averse).

Proposition **3.9.** *Let $\{r_t\}$ and $\{r'_t\}$ be two sets of real-valued measurable functions and define*

$$S_T := \sum_{t=0}^{T} r_t(X_t) \text{ and } S'_T := \sum_{t=0}^{T} r'_t(X_t),$$

where $\{X_t\}$ is a Markov chain with a family of transition kernels $\{P^{(t)}\}$. If a valuation map $\mathcal{U}$ is concave (resp. convex), then

$$\rho(\alpha S_T + (1-\alpha)S'_T | X_0 = x) \geq \alpha\rho(S_T|X_0 = x) + (1-\alpha)\rho(S'_T|X_0 = x)$$
$$\text{resp. } \rho(\alpha S_T + (1-\alpha)S'_T | X_0 = x) \leq \alpha\rho(S_T|X_0 = x) + (1-\alpha)\rho(S'_T|X_0 = x)$$

holds for all $x \in \mathsf{X}$ and $\alpha \in [0,1]$, where ρ is defined in (3.5)*.*

Proof. We prove only the concave case. The convex case follows analogously. Define $r^\alpha_t := \alpha r_t + (1-\alpha)r'_t$. By (3.5),

$$\begin{aligned}&\rho(\alpha S_T + (1-\alpha)S'_T | X_0 = x)\\ =&r^\alpha_0(x) + \mathcal{U}^{(0)}_x[r^\alpha_1 + \ldots + \mathcal{U}^{(T)}[r^\alpha_T]\ldots]].\end{aligned}$$

By the concavity of $\mathcal{U}^{(T)}$, we have

$$r^\alpha_{T-1} + \mathcal{U}^{(T)}[r^\alpha_T] \geq \alpha r_{T-1} + (1-\alpha)r'_{T-1} + \alpha\mathcal{U}^{(T)}[r_T] + (1-\alpha)\mathcal{U}^{(T)}[r'_T].$$

By the monotonicity and again the concavity of $\mathcal{U}^{(T-1)}$, the above inequality implies

$$\begin{aligned}&\mathcal{U}^{(T-1)}\left[r^\alpha_{T-1} + \mathcal{U}^{(T)}[r^\alpha_T]\right]\\ \geq&\alpha\mathcal{U}^{(T-1)}\left[r_{T-1} + \mathcal{U}^{(T)}[r_T]\right] + (1-\alpha)\mathcal{U}^{(T-1)}\left[r'_{T-1} + \mathcal{U}^{(T)}[r'_T]\right].\end{aligned}$$

By induction, we obtain the required inequality. □

The above proposition shows that the convexity (respectively concavity) of $\mathcal{U}$ implies the convexity (respectively concavity) of the constructed valuation function ρ. This sheds light on how to choose $\mathcal{U}$ if we want uniform risk-averse (or risk-seeking) behaviors. If a mixed-preference is expected, one can apply $\mathcal{U}$ that is neither convex nor concave, e.g., 1) $\mathcal{U}_x$ is convex for some states and is concave for other states, or 2) $\mathcal{U}_x$ is neither convex nor concave for some states (for an example see Section 2.4.5).

4

Poisson Equation

Die Beschäftigung mit der Mathematik, sage ich,
ist das beste Mittel gegen die Kupidität.
— Thomas Mann, *Zauberberg*

Précis Given a reward function r and a valuation map on a time-homogeneous Markov chain $\mathcal{U}$, we investigate in this chapter the related Poisson equation:

$$\mathcal{T}(h) := r + \mathcal{U}(h) = \rho + h,$$

where its solution (ρ, h) is composed of a constant and a function on the state space X. We start with a review of the Lyapunov approach applied in literature of Markov chains on general state spaces with possibly unbounded reward functions. In the same general setting, we extend this approach to general valuation maps. Two types of conditions, i.e., 1) existence of a Lyapunov function and 2) Doeblin-like conditions, are stated to ensure not only the existence of a solution to the Poisson equation but also a geometric convergence of iterations to the solution. For the important type of valuation maps, the entropic map, we will show that the above conditions are satisfied, if 1) a Lyapunov function exists for the entropic map, 2) the local Doeblin's condition holds for the underlying Markov chain, and 3) a growth condition for reward functions. Finally, we investigate the same problem on finite state spaces. In this restrict setting, we state sufficient conditions for multistep contractions.

Publications related to this chapter Main results of Section 4.3 and 4.4 and have been contained in Shen et al., 2013, Section 3 and Shen et al., 2014b, Section 3 and 4.

Notations Let X denote the state space, which is a *Borel space*, i.e., a Borel subset of a complete separable metric space, and its Borel σ-algebra is denoted by $\mathcal{B}(\mathsf{X})$. Denote by $\mathscr{P}$ the space of all probability measures on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$, and by $\mathscr{L}$ a linear space of real-valued $\mathcal{B}(\mathsf{X})$-measurable functions containing all constant functions.

Hence, $\mathbb{R} \subset \mathscr{L}$. Finally, for arbitrary two functions f and g in $\mathscr{L}$, we say $f \leq g$, if $f(x) \leq g(x)$ for each $x \in \mathsf{X}$.

4.1 Motivation

In this chapter, we consider merely the time-homogeneous Markov chain with a transition kernel P and a reward function r. Let $\mathcal{U}$ be a valuation map on P. Hence, the operators defined in (3.3) is also time-homogeneous, i.e.,

$$\mathcal{T}_n(f) \equiv \mathcal{T}(f) = r + \mathcal{U}(f), n = 1, 2, \ldots$$

and therefore, $\rho[S_T] = \mathcal{T}^T(f)$, where the iteration is defined as

$$\mathcal{T}^0(f) := f, \quad \mathcal{T}^n(f) := \mathcal{T}(\mathcal{T}^{n-1}(f)), n = 1, 2, \ldots. \tag{4.1}$$

We now investigate the limit behavior of the *average valuation*, i.e.,

$$\lim_{n\to\infty} \frac{1}{n+1} \mathcal{T}^n(r).$$

We will show below that the the average valuation problem is strongly connected to the following *Poisson equation*:

$$r(x) + \mathcal{U}_x(h) = \rho + h(x), \forall x \in \mathsf{X}, \tag{4.2}$$

where $\rho \in \mathbb{R}$ and $h : \mathsf{X} \to \mathbb{R}$. If (4.2) holds, then (ρ, h) is said to be a solution to the Poisson equation.

Indeed, if (ρ, h) is a solution to the Poisson equation (4.2), we obtain $\mathcal{T}^n(h)(x) = n\rho + h(x), \forall x \in \mathsf{X}$, which implies that for each $x \in \mathsf{X}$, $\lim_{n\to\infty} \frac{1}{n+1} \mathcal{T}^n(h)(x) = \rho$. If we can furthermore show

$$\frac{1}{n+1} (\mathcal{T}^n(h) - \mathcal{T}^n(r)) \to 0,$$

under appropriately chosen norm, then the average valuation problem is solved:

$$\lim_{n\to\infty} \frac{1}{n+1} \mathcal{T}^n(r)(x) = \rho, \forall x \in \mathsf{X}.$$

This average valuation problem is also related to the *invariant valuation function* to be defined below.

Definition **4.1.** *A valuation function v is said to be an* invariant valuation function *of U on $\mathscr{L}$ if it satisfies $v(\mathcal{U}(f)) = v(f), \forall f \in \mathscr{L}$.*

Hence, by its definition, if there exists a unique invariant valuation function v, then the Poisson equation (4.2) implies also $v(r) = \rho$.

Note that ensuring a (unique) solution to the Poisson equation is not sufficient for our purpose, since we would like, furthermore, to obtain the solution (ρ, h) by

some iterative algorithm. In other words, we expect $\mathcal{T}^n(f) \to h$ under some distance measure and ideally, we can even quantify the convergence speed as a function of n. In the literature of Markov chains, the above requirements can be satisfied if the underlying Markov chain is *geometrically ergodic* (for details see the next section). This result will be generalized to general valuation maps in Section 4.3.

4.2 Lyapunov approach for Markov chains

In this section, we consider merely the special case where $\mathcal{U} = P$. In this case, the link between the Poisson equation (4.2) and the average valuation (reward) is well known. See e.g. Glynn and Meyn (1996) and Makowski and Shwartz (2002). Furthermore, in this case the invariant valuation function ν is called an *invariant measure* (see e.g. Hernández-Lerma and Lasserre, 2003), if we consider $\mathscr{L} = \mathscr{B}$, i.e., the space of all bounded real-valued $\mathcal{B}(\mathsf{X})$-measurable functions. In addition, ν is linear. This means also that P is *ergodic* with respect to ν. For the formal definition of ergodicity in the context of probability theory, we refer to Walters (2000).

A general condition dated back to Harris (1956) states that if a Markov chain admits a "small" set, then it is uniquely ergodic and therefore a solution to the Poisson equation is guaranteed. This is usually established by finding a Lyapunov function with "small" level sets (Meyn and Tweedie, 1993, Chapter 14). If the Lyapunov function is strong enough, the transition probabilities converge exponentially fast towards the unique invariant measure and the constant in front of the exponential rate is controlled by the Lyapunov function (Meyn and Tweedie, 1993, Chapter 15). There have been other variations which have made use of Poisson equations or worked at getting explicit constants (Kontoyiannis and Meyn, 2005; Douc et al., 2004; Del Moral et al., 2003). Hairer and Mattingly (2011) stated a simplified version of the conditions, which is the main approach that we will follow and extend in this thesis. We state first the following notations mainly taken from Hairer and Mattingly (2011).

4.2.1 Weighted norm

Let $w : \mathsf{X} \to [1, \infty)$ be a given real-valued $\mathcal{B}(\mathsf{X})$-measurable function. Consider the w-norm

$$\|u\|_w := \sup_{x \in \mathsf{X}} \frac{|u(x)|}{w(x)}.$$

Let $\mathscr{B}_w$ be the space of real-valued $\mathcal{B}(\mathsf{X})$ measurable functions with bounded w-norm. It is obvious that $\mathscr{B} \subset \mathscr{B}_w$, where $\mathscr{B}$ denotes the space of bounded $\mathcal{B}(\mathsf{X})$-measurable functions. Let μ be a signed measure on $\mathcal{B}(\mathsf{X})$. Define

$$\|\mu\|_w := \sup_{\|u\|_w \leq 1} \left| \int_{\mathsf{X}} u d\mu \right| = \int_{\mathsf{X}} w \mathrm{d}|\mu| \geq \|\mu\|_{TV},$$

where $\|\cdot\|_{TV}$ denotes the total variation norm of probability measures.

The *w-seminorm* is

$$\|v\|_{s,w} := \sup_{x\neq y} \frac{|v(x)-v(y)|}{d_w(x,y)}, \text{where } d_w(x,y) := \begin{cases} 0 & x = y \\ w(x)+w(y) & x \neq y \end{cases}.$$

This Lipschitz-type seminorm is originally used by Hairer and Mattingly (2011) to study the ergodicity of Markov chains. In particular, when restricting to the bounded space $\mathscr{B}$, i.e., setting $w \equiv 1$, the seminorm is called *span-norm* in Hernández-Lerma (1989) or *Hilbert seminorm* in Gaubert and Gunawardena (2004). In the following, we restate the Lemma 2.1 in Hairer and Mattingly (2011) and incorporate its proof for readers' convenience.

LEMMA **4.2.** $\|v\|_{s,w} = \min_{c\in\mathbb{R}} \|v+c\|_w, \forall v \in \mathscr{B}_w$.

Proof. It is obvious that $\|v\|_{s,w} \leq \|v\|_w$ and therefore

$$\|v\|_{s,w} \leq \inf_{c\in\mathbb{R}} \|v+c\|_w.$$

It remains to prove the reverse inequality. Given any $\|v\|_{s,w} \leq 1$, set

$$c := \inf_x \{w(x) - v(x)\}.$$

Note that for any x and y,

$$v(x) \leq |v(y)| + |v(x)-v(y)| \leq |v(y)| + w(x) + w(y).$$

Hence $w(x) - v(x) \geq -w(y) - |v(y)|$, which implies that c is bounded below and hence $|c| < \infty$. Observe that

$$\begin{aligned} v(x)+c &\leq v(x) + w(x) - v(x) \leq w(x) \qquad \text{and} \\ v(x)+c &= \inf_y \{v(x) + w(y) - v(y)\} \\ &\geq \inf_y \{w(y) - d_w(x,y)\|v\|_{s,w}\} \geq -w(x). \end{aligned}$$

Hence, $|v(x)+c| \leq w(x)$ as required. □

4.2.2 Ergodicity conditions

Hairer and Mattingly (2011) state the following two conditions.

Assumption 4.3 (Assumption 1 and 2 in Hairer and Mattingly, 2011). (i) There exists a function $w : \mathsf{X} \to [0,\infty)$, which is $\mathcal{B}(\mathsf{X})$-measurable, and constants $K \geq 0$ and $\gamma \in (0,1)$ such that

$$P_x(w) \leq \gamma w(x) + K, \forall x \in \mathsf{X}$$

(ii) There exists a constant $\alpha \in (0,1)$ and a probability measure ν so that

$$\inf_{x\in \mathsf{B}} P(x,\mathsf{C}) \geq \alpha \nu(\mathsf{C}), \mathsf{C} \in \mathcal{B}(\mathsf{X})$$

with $\mathsf{B} := \{x \in \mathsf{X} : w(x) \leq R\}$ for some $R > \frac{2K}{1-\gamma}$.

The first condition guarantees that there exists a Lyapunov function that controls the growth of iterations. More specifically, suppose that $f \in \mathscr{B}_{1+\beta w}$ with some $\beta > 0$. Then,

$$P(f) \leq \|f\|_{1+\beta w} P(1+\beta w) \leq \|f\|_{1+w}(1+\beta\gamma w + \beta K).$$

By iteration and the linearity of P, we have

$$P^n(f) \leq \|f\|_{1+\beta w}(1+\beta\gamma^n w + \beta\frac{1-\gamma^n}{1-\gamma}K),$$

from which we can conclude that $\|P^n(f)\|_{1+\beta w}$ is globally upper bounded. Recall that $\mathcal{T}(f) = r + \mathcal{U}(f) = r + P(f)$. By the linearity of P and (4.3), the above inequality implies that for all $f, g \in \mathscr{B}_{1+\beta w}$,

$$\frac{1}{n}\|\mathcal{T}^n(f) - \mathcal{T}^n(g)\|_{1+\beta w} = \frac{1}{n}\|P^n(f) - P^n(g)\|_{1+\beta w} \to 0, \tag{4.3}$$

under the $(1+\beta w)$-norm.

The second condition in Assumption 4.3 ensures that there exists an ergodic "small" set B. This condition can be also viewed as a variant of the Doeblin's condition (see e.g. Doob, 1953).

We now state the contraction result obtained by Hairer and Mattingly (2011). We will show in Section 4.3 that this contraction under an appropriately chosen weighted seminorm is the most important property for proving the existence of a solution to the Poisson equation .

Lemma **4.4** (Theorem 3.1 in Hairer and Mattingly, 2011)**.** *Suppose Assumption 4.3 holds. There exist constants $\bar{\alpha} \in (0,1)$ and $\beta > 0$, both of which depend on γ, K and α, such that*

$$\|P(f)\|_{s,1+\beta w} \leq \bar{\alpha}\|f\|_{s,1+\beta w}, \forall f \in \mathscr{B}_{1+\beta w}.$$

Proof. See the proof of Theorem 3.1 in Hairer and Mattingly (2011). We refer to also the proof of Lemma 4.9 in this chapter, which is in fact a generalized version of this lemma. □

An example

We state one example of Markov chains which satisfies Assumption 4.3. Consider a 1-dimensional simple autoregressive model

$$X_{t+1} = \delta X_t + \sigma N_t \tag{4.4}$$

with some $\delta \in (-1,1)$, $\sigma > 0$ and N_t being standard is i.i.d. white noise. The transition kernel is then

$$P(dy|x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\delta x)^2}{2\sigma^2})dy.$$

This model is a special case of *autoregressive moving average* (ARMA) time series (see e.g. Hamilton, 1994), which are widely applied in econometrics. In fact, the property to be introduced below can be applied to general ARMA models with slight modifications.

Lyapunov function We consider $w(x) = e^{\epsilon x^2}$ with a coefficient $\epsilon > 0$ to be specified later. We first assume $1 - 2\epsilon\sigma^2 > 0$. Applying w, we have

$$\begin{aligned} P_x(w) =& \frac{1}{\sqrt{2\pi}\sigma} \int e^{-\frac{(y-\delta x)^2}{2\sigma^2} + \epsilon y^2} dy \\ =& \frac{1}{\sqrt{2\pi}\sigma} \int e^{-(\frac{1}{2\sigma^2} - \epsilon) y^2 + \frac{2\delta y x}{2\sigma^2} - \frac{\delta^2}{2\sigma^2} x^2} dy \\ =& \frac{1}{\sqrt{1-2\epsilon\sigma^2}} e^{\frac{\delta^2 \epsilon}{1-2\epsilon\sigma^2} x^2} \end{aligned}$$

Let $\gamma := \frac{\delta^2}{1-2\epsilon\sigma^2}$ and $C := \frac{1}{\sqrt{1-2\epsilon\sigma^2}}$. Note that since $|\delta| < 1$, we can always select ϵ such that $\gamma \in [0,1)$. This yields $P_x(w) \leq Cw^\gamma(x)$. Finally, for each $\gamma \in [0,1)$ and $C > 0$, we can always select sufficiently large K such that $\gamma x + K \geq Cx^\gamma, \forall x \geq 0$, which implies that

$$P_x(w) \leq Cw^\gamma(x) \leq \gamma w(x) + K, \forall x \in \mathsf{X}.$$

Doeblin's condition Given the Lyapunov function $w(x) = e^{\epsilon x^2}$, the level set is then of the form $\{x \in \mathbb{R} || x| \leq \tilde{R}\}$ with some constant $\tilde{R}$. We show below for any positive $\tilde{R}$, the required constant and probability measure exist. Indeed, by $(y-\delta x)^2 \leq 2y^2 + \delta^2 x^2$, we can deduce

$$e^{-\frac{(y-\delta x)^2}{2\sigma^2}} \geq e^{-\frac{y^2}{\sigma^2} - \frac{\delta^2 x^2}{\sigma^2}} \geq e^{-\frac{\delta^2 \tilde{R}^2}{\sigma^2}} e^{-\frac{y^2}{\sigma^2}},$$

which yields the required constant $\alpha = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\delta^2 \tilde{R}^2}{\sigma^2}} (\int_{\mathbb{R}} e^{-\frac{y^2}{\sigma^2}} dy)$ and the required probability measure $\mu(dy) = (\int_{\mathbb{R}} e^{-\frac{y^2}{\sigma^2}} dy)^{-1} e^{-\frac{y^2}{\sigma^2}} dy$.

Remark 4.5. We can choose in fact Lyapunov functions which "grow slowly", while the Doeblin's condition remains satisfied. Examples are $w(x) = \epsilon x^2$, or $w(x) = e^{\epsilon |x|^p}$ with $p \in (0,2)$ and some appropriately chosen positive constant ϵ. To see this, we refer to Section 4.4.1 below.

4.3 General theory

Recall that under some Lyapunov assumptions (see Assumption 4.3), we obtain the following two inequalities (see Lemma 4.4 and (4.3)) for Markov chains, i.e., $\mathcal{U} = P$,

$$\|\mathcal{T}(f) - \mathcal{T}(g)\|_{s,1+\beta w} \leq \tilde{\alpha}\|f - g\|_{s,1+\beta w}, \forall f, g \in \mathscr{B}_{1+\beta w}. \tag{4.5}$$

and

$$\frac{1}{n}\|\mathcal{T}^n(f) - \mathcal{T}^n(g)\|_{1+\beta w} \to 0. \tag{4.6}$$

Let $w' := 1 + \beta w$. Based on these two conditions, we state the following result for arbitrary valuation maps.

THEOREM **4.6.** *Suppose the operator $\mathcal{T}(\cdot) = r + \mathcal{U}(\cdot)$ satisfies (4.5) and (4.6). Then there exist*

(i) a solution $(\rho, h) \in \mathbb{R} \times \mathscr{B}_{w'}$ to the Poisson equation (4.2), where ρ is unique and

(ii) a valuation function v satisfying

$$v(r + \mathcal{U}(f)) = v(f) + \rho, \forall f \in \mathscr{B}_{w'}.$$

Proof. (i) Let $\tilde{\mathscr{B}}_{w'} = \mathscr{B}_{w'}/\sim$ be the quotient space, which is induced by the equivalence relation $\sim$ on $\mathscr{B}_{w'}$ defined by $f \sim g$ if and only if there exists some constant $A \in \mathbb{R}$ such that $f(x) - g(x) = A \forall x \in \mathsf{X}$, endowed with the quotient norm induced by the weighted seminorm. Starting from any $v \in \mathscr{B}_{w'}$, $\{v_n := \mathcal{T}^n(v)\}$ is a Cauchy sequence in $\tilde{\mathscr{B}}_{w'}$ under the w'-seminorm due to (4.5). Then by the fixed point argument w.r.t. the w'-seminorm as applied in p. 321 of Arapostathis et al. (1993), there exists a fixed point $h \in \mathscr{B}_{w'}$ such that $\|\mathcal{T}(h) - h\|_{s,w'} = 0$. Hence, there exists a constant $\rho \in \mathbb{R}$ such that $\mathcal{T}(h) = r + \mathcal{U}(h) = h + \rho$.

Uniqueness of ρ. Suppose there are two solutions (ρ, h) and (ρ', h') in $\mathbb{R} \times \mathscr{B}_w$. Then, $\mathcal{T}^n(h) = h + n\rho$ and $\mathcal{T}^n(h') = h' + n\rho'$. By (4.6),

$$\frac{1}{n}\|\mathcal{T}^n(h) - \mathcal{T}^n(h')\|_w = \frac{1}{n}\|h - h' + n(\rho - \rho')\|_w \to 0$$

as $n \to \infty$ implies that $\rho = \rho'$.

(ii) Let $\mu_0 \in \mathscr{M}_{w'}$ be a probability measure and h be one solution in $\mathscr{B}_{w'}$. We show first that

$$\lim_{m\to\infty} \sup_{n \geq m} |\mu_0[\mathcal{T}^n(v) - \mathcal{T}^n(h)] - \mu_0[\mathcal{T}^m(v) - \mathcal{T}^m(h)]| = 0, \forall v \in \mathscr{B}_{w'}. \tag{4.7}$$

Indeed, define $v_n := \mathcal{T}^n(v)$ and $h_n := \mathcal{T}_c^n(h), n = 1, 2, \ldots$ and we have

$$\begin{aligned}
&\sup_{\|v-h\|_{s,w'} \le A} |\mu_0[v_n - h_n] - \mu_0[v_m - h_m]| \\
\le &\sup_{\|v_1-h_1\|_{s,w'} \le \bar{\alpha} A} \left|\mu_0\left[\mathcal{T}^{n-1}(v_1) - \mathcal{T}^{n-1}(h_1)\right] - \mu_0\left[\mathcal{T}^{m-1}(v_1) - \mathcal{T}^{m-1}(h_1)\right]\right| \\
\le &\sup_{\|v_m-h_m\|_{s,w'} \le \bar{\alpha}^m A} |\mu_0[\mathcal{T}^{n-m}(v_m) - \mathcal{T}^{n-m}(h_m)] - \mu_0[v_m - h_m]|.
\end{aligned}$$

by which (4.7) follows immediately.

Define $\mathcal{D}(\cdot) := \mathcal{T}(\cdot) - \rho$ and $\mu_n(\cdot) := \mu_0(\mathcal{D}^n(\cdot))$. (4.7) is equivalent to

$$\lim_{m\to\infty} \sup_{n\ge m} |\mu_0[\mathcal{D}^n(v) - \mathcal{D}^m(v)]| = \lim_{m\to\infty} \sup_{n\ge m} |\mu_n(v) - \mu_m(v)| = 0, \forall v \in \mathscr{B}_{w'}.$$

Hence, μ_n converges to a mapping $\mu_\infty : \mathscr{B}_{w'} \to \mathbb{R}$ satisfying

$$\mu_\infty(\mathcal{D}(v)) = \mu_\infty(v), \forall v \in \mathscr{B}_{w'}.$$

On the other hand, for each n, μ_n satisfies the axioms of valuation functions except the axiom of centralization. Hence, μ_∞ preserves two axioms of valuation functions and by setting $\nu(\cdot) := \mu_\infty(\cdot) - \mu_\infty(0)$ we obtain the required valuation function. □

Extension I: uniformly bounded iterations For some valuation maps, for instance the entropic map (see Section 4.4), the contraction parameter α in (4.5) might depend on also the distance between f and g, $\|f - g\|_{s,w'}$ and $\alpha \to 1$ as $\|f - g\|_{s,w'} \to \infty$. Then 4.5 does not hold. Nevertheless, if we can show that starting with any f satisfying $\|f\|_{s,w'} \le C$, the iteration $\{\mathcal{T}^n(f), n = 1, 2, \ldots\}$ remains bounded by C under the same seminorm, the results stated above still holds by restricting to the bounded subset. More details are referred to Section 4.3.2.

Extension II: multistep contraction Comparing with the Doeblin's condition stated in Assumption 4.3 (ii), a more general condition is to assume that there exists some positive integer n_0 such that $P^{n_0}(\cdot) \ge \alpha\mu(\cdot)$. This yields the multistep contraction: $\|\mathcal{T}^{n_0}(f) - \mathcal{T}^{n_0}(g)\|_{s,w'} \le \bar{\alpha}\|f - g\|_{s,w'}$ for the linear case $\mathcal{U} \equiv P$. One can check that replacing (4.5) with the above multistep contraction, the statement of Theorem 4.6 still holds. More details about this extension will be discussed in Section 4.5, where properties of valuation maps on finite state spaces are investigated under the assumption that the underlying Markov chain satisfies the multistep Doeblin's condition.

By Theorem 4.6, it is sufficient to investigate the conditions, under which the contraction stated in (4.5) and the boundedness stated in (4.6) hold. We first study in Section 4.3.1 a special case where the valuation map $\mathcal{U}$ is convex and homogeneous and then investigate general cases in Section 4.3.2.

4.3.1 Convex and homogeneous valuation maps

We state the following assumption that generalizes Assumption 4.3.

Assumption 4.7. Let $\mathcal{U}$ be a convex and homogeneous valuation map. There exists a function $w : \mathsf{X} \to [0, \infty)$ which is $\mathcal{B}(\mathsf{X})$-measurable, constants $K \geq 0$, $\gamma \in (0,1)$, $\alpha \in (0,1)$, and a probability measure ν such that (i)

$$\mathcal{U}_x(w) \leq \gamma w(x) + K, \forall x \in \mathsf{X} \tag{4.8}$$

and (ii)

$$\inf_{x \in \mathsf{B}} \{\mathcal{U}_x(v) - \alpha\nu(v) - \mathcal{U}_x(u) + \alpha\nu(u)\} \geq 0 \tag{4.9}$$

whenever $v \geq u \in \mathscr{B}_{1+w}$, where $\mathsf{B} := \{x \in \mathsf{X} : w(x) \leq R\} \in \mathcal{B}(\mathsf{X})$ for some $R > 2K/(1-\gamma)$.

Comparing with Assumption 4.3, the Doeblin's condition (ii) is generalized here with every partially ordered pairs $v \geq u$ in $\mathscr{B}_{1+w}$ which is in fact equivalent to $\mathscr{B}_{1+\beta w}$ with any constant $\beta > 0$.

In the following, we connect the generalized Doeblin's condition in (4.9) with the classical one stated in Assumption 4.3(ii) by *subgradients* of valuation maps (see also e.g. Svindland, 2009b). Define the subgradient at state $x \in \mathsf{X}$ and function $u \in \mathscr{B}_{1+w}$ for a real-valued valuation map $\mathcal{U}$ as follows,

$$\delta\mathcal{U}_x(u) := \left\{ g \;\middle|\; \begin{array}{l} g \text{ is } \mathcal{B}(\mathsf{X})\text{-measurable and } \int |g|(w+1)dP_x < \infty, \\ \mathcal{U}_x(v) \geq \mathcal{U}_x(u) + \int g(v-u)dP_x, \forall v \geq u \in \mathscr{B}_{1+w} \end{array} \right\}.$$

Proposition **4.8.** *Suppose the transition kernel P satisfies Assumption 4.3(ii) with some constant $\beta > 0$, set $\mathsf{B} \in \mathcal{B}(\mathsf{X})$ and probability measure μ, i.e.*

$$\inf_{x \in \mathsf{B}} \{P_x(\mathsf{A}) - \beta\mu(\mathsf{A})\} \geq 0, \forall \mathsf{A} \in \mathcal{B}(\mathsf{X}).$$

Assume further that there exists $g(x,u) \in \delta\mathcal{U}_x(u)$ and positive constant $\epsilon > 0$ such that $g(x,u) \geq \epsilon$ for all $x \in \mathsf{B}$ and $u \in \mathscr{B}_w$. Then (4.9) holds for $\nu = \mu$ and $\alpha = \epsilon\beta$.

Proof. By definition, we have for each $x \in B$ and $u \in \mathscr{B}_w$

$$\mathcal{U}_x(v) \geq \mathcal{U}_x(u) + \int g(x,u)(v-u)dP_x \geq \mathcal{U}_x(u) + \epsilon\beta\mu(v-u).$$

Then setting $\nu = \mu$ and $\alpha = \epsilon\beta$, (4.9) holds. □

Now we state the contraction property under weighted seminorm.

Lemma **4.9.** *Suppose Assumption 4.7 holds. Then there exist constants $\bar{\alpha} \in (0,1)$ and $\beta > 0$ such that*

$$\|\mathcal{U}(v) - \mathcal{U}(u)\|_{s,1+\beta w} \leq \bar{\alpha}\|v - u\|_{s,1+\beta w},$$

for all v and u in $\mathscr{B}_{1+\beta w}$.

Proof. Let $w' := 1 + \beta w$ with some constant $\beta > 0$ which will be specified later. Clearly, the assertion is equivalent to $\|\mathcal{U}(v+u) - \mathcal{U}(u)\|_{s,w'} \le \bar{\alpha}\|v\|_{s,w'}$, $\forall v, u \in \mathscr{B}_{w'}$. Suppose $\|v\|_{s,w'} = C$. Lemma 4.2 suggests that we can always find a real value c such that $\|v + c\|_{w'} = C$. Since adding any constant to v will not change the values of both sides of the required inequality, without loss of generality, we assume $\|v\|_{w'} = C$. Hence, $|v(x)| \le \|v\|_{w'} w'(x) = Cw'(x)$, $\forall x \in \mathsf{X}$. Note that $\mathcal{U}$ is convex and homogeneous. Then by Proposition 2.9 and the monotonicity of $\mathcal{U}$, we have $\forall x \in \mathsf{X}$,

$$\mathcal{U}_x(v+u) - \mathcal{U}_x(u) \le \mathcal{U}_x(v) \le \mathcal{U}_x(|v|).$$

Switching $v + u$ and u, we obtain $\mathcal{U}_x(v) - \mathcal{U}_x(v+u) \le \mathcal{U}_x(|v|)$. Hence,

$$|\mathcal{U}_x(v+u) - \mathcal{U}_x(u)| \le \mathcal{U}_x(|v|) = \|v\|_{w'}\mathcal{U}_x(w') = C\left(1 + \beta\mathcal{U}_x(w)\right), \tag{4.10}$$

where the equalities are due to the homogeneity of $\mathcal{U}$.

We first assume $w(x) + w(y) \ge R$ and set $\gamma_0 := \gamma + \frac{2K}{R} < 1$ and $\gamma_1 := \frac{2+\beta R\gamma_0}{2+\beta R} \in (\gamma_0, 1)$. (4.10) yields

$$\begin{aligned}
&|\mathcal{U}_x(v+u) - \mathcal{U}_x(u) - \mathcal{U}_y(v+u) + \mathcal{U}_y(u)| \\
\le&|\mathcal{U}_x(v+u) - \mathcal{U}_x(u)| + |\mathcal{U}_y(v+u) - \mathcal{U}_y(u)| \\
\le&C\left(2 + \beta\mathcal{U}_x(w) + \beta\mathcal{U}_y(w)\right) \le C\left(2 + \beta\gamma w(x) + \beta\gamma w(y) + 2\beta K\right) \\
\le&C\left(2 + \beta\gamma_0 w(x) + \beta\gamma_0 w(y)\right) \le C\left(2\gamma_1 + \beta\gamma_1 w(x) + \beta\gamma_1 w(y)\right) \\
=&C\gamma_1 d_\beta(x,y),
\end{aligned} \tag{4.11}$$

where the last inequality is due to fact that $\frac{2(1-\gamma_1)}{\beta(\gamma_1-\gamma_0)} = R \le w(x) + w(y)$.

Now consider $w(x) + w(y) \le R$. Thus $x, y \in \mathsf{B}$. Define a new valuation map $\tilde{\mathcal{U}}_x(v) := \frac{1}{1-\alpha}\mathcal{U}_x(v) - \frac{\alpha}{1-\alpha}\nu(v)$. It is easy to verify that $\tilde{\mathcal{U}}$ is valid valuation map on B. In fact, the monotonicity is guaranteed by Assumption 4.7 (ii). Furthermore, $\tilde{\mathcal{U}}$ is also convex and homogeneous. Hence, by replacing $\mathcal{U}$ with $\tilde{\mathcal{U}}$ (4.10) holds for all $x, y \in \mathsf{B}$, which yields

$$\begin{aligned}
&|\mathcal{U}_x(v+u) - \mathcal{U}_x(u) - \mathcal{U}_y(v+u) + \mathcal{U}_y(u)| \\
=&(1-\alpha)|\tilde{\mathcal{U}}_x(v+u) - \tilde{\mathcal{U}}_x(u) - \tilde{\mathcal{U}}_y(v+u) + \tilde{\mathcal{U}}_y(u)| \\
\le&(1-\alpha)C\left(2 + \beta\tilde{\mathcal{U}}_x(w) + \beta\tilde{\mathcal{U}}_y(w)\right).
\end{aligned}$$

On the other hand $\tilde{\mathcal{U}}_x(w) \le (1-\alpha)^{-1}\mathcal{U}_x(w)$, since $w \ge 0$. Hence,

$$\begin{aligned}
(1-\alpha)C(2 + \beta\tilde{\mathcal{U}}_x(w) + \beta\tilde{\mathcal{U}}_y(w)) \le&2(1-\alpha)C + \beta C\left(2 + \beta\mathcal{U}_x(w) + \beta\mathcal{U}_y(w)\right) \\
\le&2(1-\alpha)C + \beta C(\gamma w(x) + \gamma w(y) + 2K).
\end{aligned}$$

Since $\beta = \alpha_0/K$ for some $\alpha_0 \in (0, \alpha)$, setting $\gamma_2 := (1 - \alpha + \alpha_0) \vee \gamma \in (0,1)$ yields,

$$\begin{aligned}
&|\mathcal{U}_x(v+u) - \mathcal{U}_x(u) - \mathcal{U}_y(v+u) + \mathcal{R}_y(u)| \\
\le&2C(1 - \alpha + \alpha_0) + C\gamma\beta(w(x) + w(y)) \le C\gamma_2 d_\beta(x,y).
\end{aligned} \tag{4.12}$$

Hence, setting $\bar{\alpha} := \gamma_1 \vee \gamma_2 < 1$, (4.11) and (4.12) imply,

$$|\mathcal{U}_x(v+u) - \mathcal{U}_x(u) - \mathcal{U}_y(v+u) + \mathcal{U}_y(u)| \leq \|v\|_{s,w'} \bar{\alpha} d_\beta(x,y),$$

by which our claim follows. □

Remark 4.10. If the valuation map $\mathcal{U}$ is concave and homogeneous, then $\tilde{\mathcal{U}}(\cdot) := -\mathcal{U}(-\cdot)$ is convex and homogeneous. Hence, Lemma 4.9 holds if $\tilde{\mathcal{U}}$ satisfies Assumption 4.7.

Proposition **4.11.** *Let $\mathcal{U}$ be convex and homogeneous valuation map. Suppose Assumption 4.7(i) holds. Then (4.6) holds with any constant $\beta > 0$.*

Proof. For any $f, g \in \mathscr{B}_{1+\beta w}$, by the convexity and homogeneity of $\mathcal{U}$, we have

$$|\mathcal{U}(f) - \mathcal{U}(g)| \leq \mathcal{U}(|f-g|) \leq \|f-g\|_{w'} \mathcal{U}(w') \leq \|f-g\|_{w'}(1 + \beta\gamma w + \beta K),$$

where the last inequality is due to Assumption 4.7 (i). Iterating the above inequality yields

$$|\mathcal{T}^n(f) - \mathcal{T}^n(g)| \leq \|f-g\|_{w'} \left(1 + \beta\gamma^n w + \beta \frac{1-\gamma^{n-1}}{1-\gamma} K\right),$$

which implies (4.6). □

An example with the mean-semideviation trade-off

We consider again the simple autoregressive process considered in (4.4), i.e.,

$$X_{t+1} = \delta X_t + \sigma N_t \tag{4.13}$$

with some $\delta \in (-1,1)$, $\sigma > 0$ and N_t being standard is i.i.d. white noise. We apply then the mean-semideviation introduced in Section 2.4.7, which is,

$$\mathcal{U}_x(f) := P_x(f) - \lambda \sqrt{P_x[P_x(f) - f]_+^2},$$

where $\lambda \in [0,1)$ controls how risk-averse the agent is. We have shown in Section 2.4.7, this map is concave and homogeneous. As we have commented in Remark 4.10, to check the Lyapunov conditions, it is equivalent to considering its convex counterpart,

$$\tilde{\mathcal{U}}_x(f) = -\mathcal{U}_x(-f) = P_x(f) + \lambda \sqrt{P_x[f - P_x(f)]_+^2}.$$

Lyapunov function We consider $w(x) = x^2$. Note that given the current state $X_t = x$, the successive state X_{t+1} is drawn, in fact, from a Gaussian distribution $\mathcal{N}(\delta x, \sigma^2)$ with mean δx and variance σ^2. We have then

$$\begin{aligned} P_x(w) = \delta^2 x^2 + \sigma^2, P_x(w^2) = \delta^4 x^4 + 6\delta^2 x^2 \sigma^2 + 3\sigma^4, \\ \text{and } P_x(w - P_x(w))^2 = P_x(w^2) - (P_x(w))^2 = 4\delta^2 x^2 \sigma^2 + 2\sigma^4, \end{aligned}$$

which yield

$$\begin{aligned} \tilde{\mathcal{U}}_x(w) =& P_x[w] + \lambda \sqrt{P_x[w - P_x(w)]_+^2} \\ \leq& P_x[w] + \lambda \sqrt{P_x[w - P_x(w)]^2} \\ =& \sigma^2 + \delta^2 x^2 + \lambda \sqrt{4\delta^2 x^2 \sigma^2 + 2\sigma^4}. \end{aligned}$$

Note that since for any $\delta^2 \in (0,1)$, σ and λ, fixing one $\gamma \in (\delta^2, 1)$, there always exists one sufficiently large constant $K > 0$ such that

$$\sigma^2 + \delta^2 x^2 + \lambda \sqrt{4\delta^2 x^2 \sigma^2 + 2\sigma^4} \leq \gamma x^2 + K,$$

the first condition in 4.7 holds.

Doeblin's condition Note that if $v \in \mathscr{B}_{1+w}$, then $P_x[v^2] \leq \|v\|_{1+w}^2 P_x[1+w]^2 < \infty$ for each k. Hence, $L^2(P_x(\cdot)) \supset \mathscr{B}_{1+w}$. By (2.22) (see also Svindland, 2009b, Section 6.3),

$$g(x,u) = \begin{cases} 1 & \text{if } u \text{ is constant} \\ 1 - \lambda \dfrac{P_x(u - P_x(u))_+ - (u - P_x(u))_+}{\sqrt{P_x(u - P_x(u))_+^2}} & \text{otherwise} \end{cases}$$

is one subgradient defined on $L^2(P_x(\cdot))$ for $u \in L^2(P_x(\cdot)) \supset \mathscr{B}_{1+w}$. Note that

$$P_x(u - P_x(u))_+ - (u - P_x(u))_+ \leq P_x(u - P_x(u))_+ \leq \sqrt{P_x(u - P_x(u))_+^2},$$

implies

$$g(k,u) \geq 1 - \lambda > 0, \lambda \in (0,1). \tag{4.14}$$

We now check that $P_x(|g(k,u)|(w+1)) < \infty$ as required in the definition of subgradients on $\mathscr{B}_{1+w}$. Indeed, for each x and $u \in \mathscr{B}_{1+w}$, we have $P_x(|g(k,u)|) = P_x(g(k,u)) = 1$ and

$$\begin{aligned} P_x(|g(k,u)|w) =& P_x(g(k,u)w) \\ \leq& P_x(w) + \lambda P_x\left(\frac{(u - P_x(u))_+}{\sqrt{P_x(u - P_x(u))_+^2}} w\right) \\ \leq& P_x(w) + \lambda \sqrt{P_x\left(\frac{(u - P_x(u))_+}{\sqrt{P_x(u - P_x(u))_+^2}}\right)^2 P_x(w^2)} \\ =& P_x(w) + \lambda \sqrt{P_x(w^2)} < \infty, \end{aligned}$$

where the second inequality is due to the Hölder's inequality. Hence $g(x,u) \in \delta\tilde{\mathcal{U}}$, $\forall x \in \mathsf{X}, u \in \mathscr{B}_w$. Finally, we have already shown in page 40 (see also Meyn and Tweedie, 1993, Page 380) that for the Markov chain defined in (4.4), the classical Doeblin's condition holds for any closed level set $\left\{x \in \mathsf{X} \mid |x| \leq \tilde{R}\right\}$ with $\tilde{R} > 0$. Hence, by Proposition 4.8, the generalized Doeblin's condition required by Assumption 4.7(ii) holds.

4.3.2 General valuation maps

One essential property of homogeneous valuation maps is that once $\mathcal{U}(|r|) < \infty$ with some reward function r, then $\mathcal{U}(k|r|) = k\mathcal{U}(|r|) < \infty$ holds for any positive constant k. This property, however, does not hold for convex valuation maps, for instant, the entropic map with $\lambda = 1$ defined as

$$\mathcal{U}(f) := \log\left[P_x(e^f)\right].$$

In fact, $\mathcal{U}(|r|) < \infty$ equals to $P_x(e^{|r|}) < \infty$, but $P_x(e^{|r|}) < \infty$ does not imply $P_x(e^{k|r|}) < \infty$ for all $k > 0$. Hence, during the iteration $\mathcal{T}^n(v)$, where $\mathcal{T}(\cdot) = r + \mathcal{U}(\cdot)$, both the reward function r and the initial function v being inside $\mathscr{B}_w$ with some weight function $w \geq 1$ does not guarantee that $\mathcal{T}^n(v) \in \mathscr{B}_w$ for each $n \in \mathbb{N}$. In other words, the iteration might explode. Hence, we need further control on the growth of r for general valuation maps like the entropic map.

Bounded forward invariant subset

We state first a set of sufficient conditions that guarantee the existence of a bounded (under w-seminorm) forward invariant subset covering the whole sequence of $\{\mathcal{T}^n(v)\}$. More specifically, we consider subspaces of the following form

$$\mathscr{B}_w^{(C)} := \{v \in \mathscr{B}_w \mid \|v\|_{s,w} \leq C\}. \tag{4.15}$$

Here, we choose w-seminorm, since we shall prove the contraction property under the w-seminorm as in (4.5).

Assumption 4.12. There exist a $\mathcal{B}(\mathsf{X})$-measurable function $w_0 : \mathsf{X} \to [0,\infty)$ and constants $\gamma_0 \in (0,1)$, $K_0 > 0$ and $\tilde{K}_0 > K_0$ such that

(i) for each $x \in \mathsf{X}$

$$(r(x) + \mathcal{U}_x(w_0)) \vee (-r(x) - \mathcal{U}_x(-w_0)) \leq \gamma_0 w_0(x) + K_0, \tag{4.16}$$

(ii) and for all $x, y \in \mathsf{B} := \{x \in \mathsf{X} | w_0(x) \leq R_0 := \frac{2K_0}{1-\gamma_0}\}$, the following inequality

$$\mathcal{U}_x(v) - \mathcal{U}_y(v) \leq 2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0) \tag{4.17}$$

holds for all v satisfying $|v| \leq w_0 + \tilde{K}_0$.

Remark 4.13. (a) If $\mathcal{U}$ is a concave, which induces risk-averse behavior, a sufficient condition to guarantee the assumption (i) is

$$|r| - \mathcal{U}(-w_0) \le \gamma_0 w_0 + K_0, \tag{4.18}$$

since by Proposition 2.8, $-\mathcal{U}(-w_0) \ge \mathcal{U}(w_0)$.

(b) The assumption (i) can be replaced by two conditions for the reward function r and valuation map $\mathcal{U}$ separately,

1) w_0 is a Lyapunov function satisfying

$$\mathcal{U}(w_0) \vee (-\mathcal{U}(-w_0)) \le \hat{\gamma}_0 w_0 + \hat{K}_0,$$

2) and $|r| \le \tilde{\gamma}_0 w_0 + C_0$ with some constants $\tilde{\gamma}_0 \in (0, 1 - \hat{\gamma}_0)$ and $C_0 > 0$.

(c) The assumption (ii) is more general than the Doeblin's condition in Assumption 4.7(ii). Indeed, Assumption 4.7(ii) implies that for all $\tilde{K}_0 > 0$, there exists a constant $\alpha \in (0,1)$ such that

$$\mathcal{U}_x(v) - \mathcal{U}_y(v) \le 2(1-\alpha)\tilde{K}_0 + \mathcal{U}_x(w_0) + \mathcal{U}_y(w_0),$$

which implies (4.17).

(d) Applying entropic maps, we will show in Section 4.4 some sufficient conditions ensuring (ii) based on the properties of the underlying Markov chain.

THEOREM **4.14.** *Suppose Assumption 4.12 holds. Then*

$$\|r + \mathcal{U}(v)\|_{s,1+\beta_0 w_0} \le \beta_0^{-1},$$

whenever $\|v\|_{s,1+\beta_0 w_0} \le \beta_0^{-1}$ *with* $\beta_0 := \tilde{K}_0^{-1}$.

Proof. Note that adding a constant to v will not change the required inequality. Due to Lemma 4.2, we assume that $|v| \le \beta_0^{-1} + w_0$. By the definition of w-seminorm, the task is to prove

$$|r(x) + \mathcal{U}_x(v) - r(y) - \mathcal{U}_y(v)| \le 2\beta_0^{-1} + w_0(x) + w_0(y), \forall x \ne y \in \mathsf{X}.$$

Note that since switching x and y will not change the right side of the inequality, it is sufficient to show

$$r(x) + \mathcal{U}_x(v) - r(y) - \mathcal{U}_y(v) \le 2\beta_0^{-1} + w_0(x) + w_0(y), \forall x \ne y \in \mathsf{X}. \tag{4.19}$$

We consider the following two cases. Case I: $w_0(x) + w_0(y) \ge R_0$. By (4.16), we have for all $\beta_0 > 0$,

$$\begin{aligned} r(x) + \mathcal{U}_x(v) &\le r(x) + \beta_0^{-1} + \mathcal{U}_x(w_0) \le \beta_0^{-1} + \gamma_0 w_0(x) + K_0, \text{ and} \\ -r(y) - \mathcal{U}_x(v) &\le -r(y) + \beta_0^{-1} - \mathcal{U}_y(w_0) \le \beta_0^{-1} + \gamma_0 w_0(y) + K_0. \end{aligned}$$

By the choice of R_0,

$$2\beta_0^{-1} + \gamma_0(w_0(x) + w_0(y)) + 2K_0 \leq 2\beta_0^{-1} + w_0(x) + w_0(y)$$

holds. Hence, (4.19) holds for this case.

Case II: $w_0(x) + w_0(y) \leq R_0$. Then both x and y are in the subset B. By (4.17),

$$\begin{aligned}
&r(x) + \mathcal{U}_x(v) - r(y) - \mathcal{U}_y(v)\\
\leq& r(x) - r(y) + 2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0)\\
\leq& 2(\tilde{K}_0 - K_0) + \gamma_0 w_0(x) + \gamma_0 w_0(y) + 2K_0\\
\leq& 2\beta_0^{-1} + w_0(x) + w_0(y).
\end{aligned}$$

Combining I and II, we obtain the required inequality. □

Geometric contraction

Given a valuation map satisfying Assumption 4.12, we can then restrict ourselves to the invariant subset $\mathscr{B}_w^{(C)}$ (with $C = \tilde{K}_0$) rather than the whole set $\mathscr{B}_w$.

Definition **4.15.** *A convex homogeneous valuation function $\bar{v}^{(w,C)}$ is said to be an* upper envelope *of a valuation function v given a bound $C \in \mathbb{R}_+$, if the following inequality holds*

$$v(v) - v(u) \leq \bar{v}^{(w,C)}(v - u), \forall v, u \in \mathscr{B}_w^{(C)}. \tag{4.20}$$

Analogously, a convex homogeneous valuation map $\bar{\mathcal{U}}^{(w,C)}$ is said to be an upper envelope *of a valuation map $\mathcal{U}$ given a bound $C \in \mathbb{R}_+$, if for all $v, u \in \mathscr{B}_w^{(C)}$,*

$$\mathcal{U}_x(v) - \mathcal{U}_x(u) \leq \bar{\mathcal{U}}_x^{(w,C)}(v - u), \forall x \in \mathsf{X}.$$

Remark 4.16. Apparently, if v (resp. $\mathcal{U}$) is convex and homogeneous, then v (resp. $\mathcal{U}$) is an upper envelope of itself for all bounds $C > 0$, due to its sublinearity (see Proposition 2.9).

We now prove the contraction property based on the following assumption.

Assumption 4.17. There exist two real-valued $\mathcal{B}(\mathsf{X})$-measurable functions,

$$w_0 : \mathsf{X} \to [0, \infty) \text{ and } w : \mathsf{X} \to [1, \infty)$$

satisfying that

(i) $\mathscr{B}_{1+w_0} = \mathscr{B}_w$;

(ii) there exist constants $\gamma \in (0, 1)$, $K > 0$ and an upper envelope $\bar{\mathcal{U}}^{(w,C)}$ such that

$$\bar{\mathcal{U}}^{(w,C)}(w_0) \leq \gamma w_0 + K;$$

(iii) for all $v \geq u \in \mathscr{B}_{1+w_0}$, there exist a constant $\alpha \in (0,1)$ and a probability measure μ on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ such that

$$\bar{\mathcal{U}}_x^{(w,C)}(v) - \bar{\mathcal{U}}_x^{(w,C)}(u) \geq \alpha \int (v(x) - u(x))\, \mu(dx), \forall x \in \mathsf{B},$$

where $\mathsf{B} := \{x \in \mathsf{X} | w_0(x) \leq R\}$ for some $R > \frac{2K}{1-\gamma}$.

LEMMA **4.18.** *Suppose Assumption 4.17 holds. Then there exists a constant $\bar{\alpha} \in (0,1)$ and $\beta > 0$ such that*

$$\|\mathcal{U}(v) - \mathcal{U}(u)\|_{s,1+\beta w_0} \leq \bar{\alpha} \|v - u\|_{s,1+\beta w_0}, \forall v, u \in \mathscr{B}_w^{(C)}.$$

Proof. Define $w' := 1 + \beta w_0$ for some $\beta \in \mathbb{R}_+$, whose value will be specified later. Suppose $\|v - u\|_{s,w'} = A \in \mathbb{R}_+$. Due to Lemma 4.2 and the fact that adding any constant to v and u will not change the values of both sides of the required inequality, we may assume that $\|v - u\|_{w'} = A$.

By the definition of upper envelope, we have then

$$\mathcal{U}_x(v) - \mathcal{U}_x(u) \leq \bar{\mathcal{U}}_x^{(w,C)}(v - u) \leq \bar{\mathcal{U}}_x^{(w,C)}(|v - u|), \forall x \in \mathsf{X},$$

where the last inequality is due to Proposition 2.8. Switching v and u, we obtain

$$|\mathcal{U}_x(v) - \mathcal{U}_x(u)| \leq \bar{\mathcal{U}}_x^{(w,C)}(|v - u|) \leq \|v - u\|_{w'} \bar{\mathcal{U}}_x^{(w,C)}(w'), \forall x \in \mathsf{X}. \tag{4.21}$$

Case I: $w_0(x) + w_0(y) \geq R$ and set $\gamma_0 := \gamma + \frac{2K}{R} < 0$ and $\gamma_1 := \frac{2+\beta R \gamma_0}{2+\beta R}$ for some $\beta > 0$. It is easy to verify that $\gamma_1 \in (0,1)$. Then (4.21) yields

$$\begin{aligned}
&|\mathcal{U}_x(v) - \mathcal{U}_x(u) - \mathcal{U}_y(v) + \mathcal{U}_y(u)| \\
\leq &|\mathcal{U}_x(v) - \mathcal{U}_x(u)| + |\mathcal{U}_y(v) - \mathcal{U}_y(u)| \\
\leq &A(2 + \beta \bar{\mathcal{U}}_x^{(w,C)}(w_0) + \beta \bar{\mathcal{U}}_y^{(w,C)}(w_0)) \leq A(2 + \beta\gamma w_0(x) + \beta\gamma w_0(y) + 2\beta K) \\
\leq &A(2 + \beta\gamma_0 w_0(x) + \gamma_0 w_0(y)) \leq A\gamma_1(w'(x) + w'(y)).
\end{aligned} \tag{4.22}$$

Case II: $w_0(x) + w_0(y) \leq R$. Hence both x and y are in the subset B. We define for all $x \in \mathsf{B}$,

$$\begin{aligned}
\tilde{\mathcal{U}}_x(v) &:= \frac{1}{1-\alpha}\mathcal{U}_x(v) - \frac{\alpha}{1-\alpha}\mu(v), \quad \text{and} \\
\tilde{\bar{\mathcal{U}}}_x^{(w,C)}(v) &:= \frac{1}{1-\alpha}\bar{\mathcal{U}}_x^{(w,C)}(v) - \frac{\alpha}{1-\alpha}\mu(v).
\end{aligned}$$

It is easy to verify that $\tilde{\bar{\mathcal{U}}}_x^{(w,C)}$ is a valid convex and homogeneous valuation function on $\mathscr{B}_{1+\beta w_0} = \mathscr{B}_{1+w_0} = \mathscr{B}_w$ for all $x \in \mathsf{B}$. Indeed, the monotonicity is satisfied

due to Assumption 4.17(iii). Hence, $\tilde{\mathcal{U}}_x(v) - \tilde{\mathcal{U}}_x(u) \le \bar{\tilde{\mathcal{U}}}_x^{(w,C)}(v-u)$ indicates that $\bar{\tilde{\mathcal{U}}}_x^{(w,C)}$ is an upper envelope of $\tilde{\mathcal{U}}_x$ for all $x \in \mathsf{B}$. Hence,

$$\begin{aligned}
&|\mathcal{U}_x(v) - \mathcal{U}_x(u) - \mathcal{U}_y(v) + \mathcal{U}_y(u)| \\
=&(1-\alpha)|\tilde{\mathcal{U}}_x(v) - \tilde{\mathcal{U}}_x(u) - \tilde{\mathcal{U}}_y(v) + \tilde{\mathcal{U}}_y(u)| \\
\le&(1-\alpha)|\tilde{\mathcal{U}}_x(v) - \tilde{\mathcal{U}}_x(u)| + (1-\alpha)|\tilde{\mathcal{U}}_y(v) - \tilde{\mathcal{U}}_y(u)| \\
\le&(1-\alpha)\bar{\tilde{\mathcal{U}}}_x^{(w,C)}(|v-u|) + (1-\alpha)\bar{\tilde{\mathcal{U}}}_x^{(w,C)}(|v-u|) \\
\le&2A(1-\alpha) + A(1-\alpha)\beta\left(\bar{\tilde{\mathcal{U}}}_x^{(w,C)}(w_0) + \bar{\tilde{\mathcal{U}}}_y^{(w,C)}(w_0)\right).
\end{aligned}$$

Note that since $(1-\alpha)\bar{\tilde{\mathcal{U}}}_x^{(w,C)}(w_0) \le \bar{\mathcal{U}}_x^{(w,C)}(w_0)$ holds for all $x \in \mathsf{B}$, we obtain

$$\begin{aligned}
&|\mathcal{U}_x(v) - \mathcal{U}_x(u) - \mathcal{U}_y(v) + \mathcal{U}_y(u)| \\
\le&2A(1-\alpha) + A\beta\left(\bar{\mathcal{U}}_x^{(w,C)}(w_0) + \bar{\mathcal{U}}_y^{(w,C)}(w_0)\right) \\
\le&2A(1-\alpha) + A\beta(w_0(x) + w_0(y) + 2K).
\end{aligned} \tag{4.23}$$

We select $\beta := \frac{\alpha_0}{K}$ for some $\alpha_0 \in (0,\alpha)$. Setting $\gamma_2 := (1-\alpha+\alpha_0) \vee \gamma \in (0,1)$ yields for all $x \ne y$

$$\begin{aligned}
&|\mathcal{U}_x(v) - \mathcal{U}_x(u) - \mathcal{U}_y(v) + \mathcal{U}_y(u)| \\
\le&2A(1-\alpha+\alpha_0) + A\gamma\beta(w_0(x) + w_0(y)) \le A\gamma_2(w'(x) + w'(y)).
\end{aligned} \tag{4.24}$$

Hence, setting $\bar{\alpha} := \gamma_1 \vee \gamma_2 < 1$, (4.22) and (4.24) imply for all $x \ne y$

$$|\mathcal{U}_x(v) - \mathcal{U}_x(u) - \mathcal{U}_y(v) + \mathcal{U}_y(u)| \le \|v-u\|_{s,w'}\bar{\alpha}(w'(x) + w'(y)),$$

the required inequality. □

Poisson equation

We set $w' = 1 + \beta w_0$ as in Lemma 4.18, $w = 1 + \tilde{K}_0^{-1} w_0$ and $C = \tilde{K}_0$ as in Theorem 4.14. Hence, apparently $\mathscr{B}_{w'} = \mathscr{B}_w$.

Lemma **4.19.** *Suppose Assumption 4.12 and 4.17 hold. Then*

$$\lim_{n\to\infty} \frac{1}{n}\|\mathcal{T}^n(v) - \mathcal{T}^n(u)\|_w = 0, \forall v, u \in \mathscr{B}_w^{(C)}.$$

Proof. It is sufficient to show that $\|\mathcal{T}^n(v) - \mathcal{T}^n(u)\|_w$ is uniformly bounded, which is equivalent to requiring that $\|\mathcal{T}^n(v) - \mathcal{T}^n(u)\|_{w'}$ is uniformly bounded.

Indeed, by Assumption 4.17(ii), setting $K' := \beta K + 1 - \gamma$, we have

$$|\mathcal{T}(v) - \mathcal{T}(u)| \le \bar{\mathcal{U}}^{(w,C)}(|v-u|) \le \|v-u\|_{w'}(\gamma w' + K')$$

where the first inequality is due to Proposition 2.8.

In addition, by Theorem 4.14, $\|\mathcal{T}^n(v)\|_{s,w} \leq C$ holds for all $n \in \mathbb{N}_+$. Hence, by induction w.r.t. n, we have for $n = 2, 3, \ldots$

$$\begin{aligned}|\mathcal{T}^n(v) - \mathcal{T}^n(u)| \leq & \bar{\mathcal{U}}^{(w,C)}(|\mathcal{T}^{n-1}(v) - \mathcal{T}^{n-1}(u)|) \\ \leq & \|v-u\|_w \bar{\mathcal{U}}^{(w,C)}\left(\gamma^{n-1}w' + K'\sum_{k=0}^{n-2}\gamma^k\right) \\ \leq & \|v-u\|_w\left(\gamma^n w' + K'\sum_{k=0}^{n-1}\gamma^k\right),\end{aligned}$$

which implies that $\|\mathcal{T}^n(v) - \mathcal{T}^n(u)\|_{w'} \leq \frac{K'}{1-\gamma}$. □

THEOREM **4.20.** *Suppose Assumption 4.12 and 4.17 hold. Then there exist*

(i) a solution $(\rho, h) \in \mathbb{R} \times \mathscr{B}_w$ to the Poisson equation $r + \mathcal{U}(h) = r + h$, where ρ is unique and

(ii) a valuation function v satisfying $v(r + \mathcal{U}(v)) = v(v) + \rho, \forall v \in \mathscr{B}_w^{(C)}$.

Proof. Restricting to the bounded forward invariant subspace $\mathscr{B}_w^{(C)}$, (4.5) and (4.6) are satisfied due to Lemma 4.18 and Lemma 4.19. Then the proof follows the same line as in the proof of Theorem 4.6 by restricting to $\mathscr{B}_w^{(C)}$. □

Remark 4.21. If $\mathcal{U}$ is convex and homogeneous, its upper envelope $\bar{\mathcal{U}}_w^{(C)}$ becomes $\mathcal{U}$ itself. In this case, Assumption 4.12 is no longer needed to determine *a priori* the size of the bounded forward invariant subset, C. Moreover, 1) Assumption 4.17(iii) implies Assumption 4.12(ii) due to (4.23), and 2) Theorem 4.20(ii) holds for all $v \in \mathscr{B}_w$.

An example with the utility-based shortfall

We consider the AR1 process defined in (4.4) and the utility based shortfall defined in (2.14)

$$\mathcal{U}_x(v) = \sup\left\{m \in \mathbb{R} \mid \int_{\mathsf{X}} u(v(y) - m)P_x(dy) \geq 0\right\}, \tag{4.25}$$

under the assumption that there exist constants l and R satisfying $0 < l \leq 1 \leq L < \infty$ and

$$l \leq \frac{u(x) - u(y)}{x - y} \leq L, \forall x, y \in \mathbb{R}. \tag{4.26}$$

In other words, for each $x, y \in \mathbb{R}$, we have $u(x) - u(y) = \delta(x,y)(x-y)$, with some $\delta(x,y) \in [l, L]$.

Remark 4.22. Note that u is not required to be convex, nor concave. Hence, the induced risk preference can be mixed as we have shown in Section 2.4.5. One example of u that satisfies the assumption we made above is a piecewise linear function with slopes upper bounded by L and lower bounded by l.

By Proposition 2.15, if the utility function u is strictly increasing, then the optimal m^* is obtained when the equality holds, i.e., for each $x \in \mathsf{X}$,

$$m^*(x) := \mathcal{U}_x(v), \int u(v(y) - m^*(x))P_x(dy) = 0,$$

Let $m = \mathcal{U}(v)$ and $m' = \mathcal{U}(v')$. Hence, for each $x \in \mathsf{X}$,

$$\int u(v(y) - m(x))P_x(dy) = \int u(v'(y) - m'(x))P_x(dy) = 0.$$

We have then

$$\begin{aligned} 0 &= \int u(v(y) - m(x))P_x(dy) - \int u(v'(y) - m'(x))P_x(dy) \\ &\leq \int \delta(v, v', x, y)(v(y) - v'(y) - m(x) + m'(x))P_x(dy), \end{aligned}$$

which implies that

(4.27)
$$(m(x) - m'(x)) \int \delta(v, v', x, y)P_x(dy) \leq \int \delta(v, v', x, y)(v(y) - v'(y))P_x(dy).$$

Verification of Assumption 4.12 Let $w(x) = e^{\epsilon x^2}$ be the Lyapunov function defined in Page 40 and we have shown that it satisfies $P_x(w) \leq C(w(x))^\gamma$, for some $C > 0$ and $\gamma \in (0,1)$. First, taking $v = w$ and $v' = 0$ in (4.27), we have $m' = 0$ and $\mathcal{U}_x(w) = m(x) \leq \frac{L}{l}P_x(w) \leq \frac{L}{l}C(w(x))^\gamma$. Second, taking $v = 0$ and $v' = -w$ in (4.27), we have $m = 0$ and

$$-\mathcal{U}_x(-w) = -m'(x) \leq \frac{L}{l}P_x(w) \leq \frac{L}{l}C(w(x))^\gamma.$$

Hence, given $r \in \mathscr{B}_{1+w^{\gamma'}}$ with some $\gamma' \in (0,1)$, we can always find a sufficiently large K_0 such that

$$(r(x) + \mathcal{U}_x(w)) \vee (-r(x) - \mathcal{U}_x(-w)) \leq \gamma_0 w(x) + K_0, \forall x \in \mathsf{X}. \tag{4.28}$$

This ensures Assumption 4.12 (i).

Taking $v' = w + \tilde{K}_0$ in (4.27), where $\tilde{K}_0$ will be specified later, due to the assumption $v \le w + \tilde{K}_0$, we have $m \le m'$ and

$$\begin{aligned} L(m(x) - m(x')) \le & (m(x) - m'(x)) \int \delta(v, v', x, y) P_x(dy) \\ \le & \int \delta(v, v', x, y)(v(y) - v'(y)) P_x(dy) \\ \le & l \int (v(y) - v'(y)) P_x(dy) \\ = & l \int (v(y) - w(y) - \tilde{K}_0) P_x(dy) \\ \Rightarrow \quad & \mathcal{U}_x(v) - \mathcal{U}_x(w + \tilde{K}_0) \le \frac{l}{L} \int (v(y) - w(y) - \tilde{K}_0) P_x(dy). \end{aligned}$$

Analogously we obtain

$$\mathcal{U}_y(-w - \tilde{K}_0) - \mathcal{U}_y(v) \le \frac{l}{L} \int (-w - \tilde{K}_0 - v(y')) P_y(dy').$$

On the other hand, we have $P_x(\cdot) \ge \alpha\mu(\cdot)$ with some probability measure μ and $\alpha \in (0,1)$ for x in any bounded level-set. Hence, we have

$$\frac{l}{L} \int (w(y) + \tilde{K}_0 - v(y)) P_x(dy) + \frac{l}{L} \int (w(y) + \tilde{K}_0 + v(y)) P_{x'}(dy) \ge \frac{l}{L} 2\alpha \tilde{K}_0,$$

which implies that

$$\mathcal{U}_x(v) - \mathcal{U}_x(w + \tilde{K}_0) + \mathcal{U}_y(-w - \tilde{K}_0) - \mathcal{U}_y(v) \le -2\frac{\alpha l}{L} \tilde{K}_0.$$

Therefore, taking $\tilde{K}_0 := \frac{L}{\alpha l} K_0$, the Assumption 4.12 (ii) holds.

Verification of Assumption 4.17 By (4.27), we have

$$\begin{aligned} \mathcal{U}_x(v) - \mathcal{U}_x(v') \le & \frac{\int \delta(v, v', x, y) P_x(dy)(v(y) - v'(y))}{\int \delta(v, v', x, y) P_x(dy)} \\ \le & \sup_{\delta : l \le \delta(x,y) \le L} \frac{\int \delta(x, y) P_x(dy)(v(y) - v'(y))}{\int \delta(x, y) P_x(dy)} =: \bar{\mathcal{U}}_x(v - v') \end{aligned} \tag{4.29}$$

It is easy to see that $\bar{\mathcal{U}}$ is a convex and homogeneous valuation map. Note that

$$\bar{\mathcal{U}}_x(w) = \sup_{\delta : l \le \delta(x,y) \le L} \frac{\int \delta(x, y) P_x(dy)(w(y))}{\int \delta(x, y) P_x(dy)} \le \frac{L}{l} P_x(w),$$

which implies that w is a Lyapunov function satisfying Assumption 4.17(ii) satisfying with constants γ_0 and $K > 0$. On the other hand, for any $v \geq v'$,

$$\sup_{\delta: l \leq \delta(x,y) \leq L} \frac{\int \delta(x,y) P_x(dy)(v(y))}{\int \delta(x,y) P(dy)} - \sup_{\delta: l \leq \delta(x,y) \leq L} \frac{\int \delta(x,y) P_x(dy)(v'(y))}{\int \delta(x,y) P(dy)}$$

(4.30)

$$\geq \inf_{\delta: l \leq \delta(x,y) \leq L} \frac{\int \delta(x,y) P_x(dy)(v(y) - v'(y))}{\int \delta(x,y) P_x(dy)} \geq \frac{l}{L} P_x(v - v') \geq \frac{\alpha l}{L} \mu(v - v'),$$

holds for all x in any bounded level-set. Hence, Assumption 4.17(iii) holds.

4.4 The entropic map

Recall that, given a Markov transition kernel P, the entropic map is defined as

$$\mathcal{U}_x(v) := \frac{1}{\lambda} \log \left(\int e^{\lambda v} dP_x \right), \lambda > 0. \tag{4.31}$$

Without loss of generality, in the remaining part of this paper, we set $\lambda = 1$. This map has been widely used in the literature of risk-sensitive Markov control/decision processes, see e.g. Howard and Matheson, 1972; Chung and Sobel, 1987; Avila-Godoy and Fernández-Gaucherand, 1998; Borkar and Meyn, 2002; Di Masi and Stettner, 2008; Coraluppi and Marcus, 2000; Fleming and Hernández-Hernández, 1997; Hernández-Hernández and Marcus, 1996; Marcus et al., 1997; Cavazos-Cadena, 2010.

Upper envelope

We now derive the upper envelope for entropic measures.

Proposition **4.23.** *Let $\nu(v) := \log \left(\int e^v d\mu \right)$ with a probability measure μ on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$. Suppose that for all $v \in \mathscr{B}_w$, $\int e^{|v|} d\mu < \infty$ holds. Then (i) $\nu(v) \leq \frac{\mu(e^v v)}{\mu(e^v)}$, and (ii)*

$$\bar{\nu}^{(w,C)}(u) := \sup_{v \in \mathscr{B}_w^{(C)}} \frac{\int e^v u d\mu}{\int e^v d\mu}$$

is an upper envelope for ν given C.

Proof. Given any two $u, v \in \mathscr{B}_w$, we obtain

$$\nu(v) - \nu(u) = \log \frac{\mu(e^v)}{\mu(e^u)} = \log \frac{\mu(e^u e^{v-u})}{\mu(e^u)} \geq \frac{\mu(e^u (v - u))}{\mu(e^u)}, \tag{4.32}$$

where the last inequality is due to Jensen's inequality. Hence,

$$\log(\mu e^v) \geq \frac{\mu(e^u v)}{\mu e^u} - \mu\left(\frac{e^u}{\mu(e^u)}(u - \log(\mu e^u))\right), \forall u, v \in \mathscr{B}_w.$$

Restricting u and v to be in the subset $\mathscr{B}_w^{(C)}$, the above inequality yields

$$\log(\mu e^v) \geq \sup_{\xi=\frac{e^u}{\mu(e^u)}, u\in\mathscr{B}_w^{(C)}} \mu(\xi v) - \mu(\xi\log(\xi)).$$

Since the equality holds by taking $\xi^* := \frac{e^v}{\mu(e^v)}$, we obtain

$$\log(\mu e^v) = \sup_{\xi=\frac{e^u}{\mu(e^u)}, u\in\mathscr{B}_w^{(C)}} \mu(\xi v) - \mu(\xi\log(\xi)). \tag{4.33}$$

The second term $\mu(\xi\log(\xi))$ on the right-hand side of the above equation is the *relative entropy* and is always nonnegative (for proof see, e.g., (Ledoux, 2001, Section 5.1)). Hence, we obtain (i). Finally, (ii) is followed by

$$\log(\mu e^v) - \log(\mu e^u) \leq \sup_{\xi=\frac{e^f}{\mu(e^f)}, f\in\mathscr{B}_w^{(C)}} \mu(\xi(v-u)) = \sup_{f\in\mathscr{B}_w^{(C)}} \frac{\int e^f(v-u)d\mu}{\int e^f d\mu}.$$

and it is easy to verify that $\bar{v}^{(w,C)}(u) = \sup_{f\in\mathscr{B}_w^{(C)}} \frac{\int e^f u d\mu}{\int e^f d\mu}$ is a valid convex and homogeneous valuation function. □

Remark 4.24. The inequality in (4.33) is similar to the dual representation of convex risk measures on L^∞ (Föllmer and Schied, 2002, 2004) or on more general spaces such as *Orlicz hearts* (Cheridito and Li, 2009). However, since we consider a different functional space, i.e., the weighted norm space, the existing result cannot be directly applied here. On the other hand, for other types of convex valuation functions, their dual representation provide us a generic approach to calculate their upper envelopes, as shown in the above proposition.

By Proposition 4.23, we obtain one upper envelope for the entropic map:

$$\bar{\mathcal{U}}_x^{(C)}(u) = \sup_{f\in\mathscr{B}_w:\|f\|_{s,w}\leq C} \frac{\int e^f u dP_x}{\int e^f dP_x}, \tag{4.34}$$

provided that $P_x(e^f) < \infty$ holds for all $f \in \mathscr{B}_w$ and $x \in \mathsf{X}$.

4.4.1 Lyapunov functions

Now we investigate properties of *Lyapunov functions* w.r.t. the entropic map.

Definition **4.25.** *A function w is said to be a* Lyapunov function *w.r.t. a valuation map $\mathcal{U}$, if*

(i) $w : \mathsf{X} \to [0,\infty)$ is $\mathcal{B}(\mathsf{X})$-measurable and unbounded from above, and

(ii) there exist constants $\gamma \in (0,1)$ and $K > 0$ satisfying $\mathcal{U}_x(w) \leq \gamma w(x) + K, \forall x \in \mathsf{X}$.

We also introduce the following notation of *level-sets*. For any unbounded nonnegative $\mathcal{B}(\mathsf{X})$-measurable function w and any real number $R \in \mathbb{R}$, we define

$$\mathsf{B}_w(R) := \{x \in \mathsf{X} | w(x) \leq R\}$$

and $\mathsf{B}_w^c(R)$ its complementary set. We then make the following assumption.

Assumption 4.26. There exists a Lyapunov function $w_1 \geq 1$ w.r.t. the entropic map $\mathcal{U}$, with constants $\gamma_1 \in (0,1)$ and $K_1 > 0$.

If the above assumption holds and setting $w_0 := w_1^p$ with any $p \in (0,1)$, then for all $f \in \mathscr{B}_{w_0}$, there exists a constant K_f (depending on p and $\|f\|_{w_0}$) satisfying

$$|f(x)| \leq \|f\|_{w_0} w_0(x) \leq w_1(x) + K_f, \forall x \in \mathsf{X}.$$

We immediately have

$$P_x(e^f) \leq P_x(e^{w_1 + K_f}) \leq e^{K_f} e^{\gamma_1 w_1 + K_1} < \infty, \forall x \in \mathsf{X}$$

and therefore, the upper envelope for the entropic map in (4.34) is well defined. In the following theorem, we show that if w_1 is a Lyapunov function w.r.t. $\mathcal{U}$, then $w_0 = w_1^p$ with any $p \in (0,1)$ is a Lyapunov function w.r.t. the upper envelope of $\mathcal{U}$.

THEOREM **4.27.** *Suppose that Assumption 4.26 holds. Let $w_0 := w_1^p$ with $p \in (0,1)$. Then, for any constant $C > 0$, there exist constants $\gamma_2 \in (0,1)$ (depending only on p and γ_1) and $K_2 > 0$ (depending on p, C, γ_1 and K_1) such that*

$$\sup_{f : f \in \mathscr{B}_{w_0}, |f| \leq w_0 + C} \frac{P_x(e^f w_0)}{P_x(e^f)} \leq \gamma_2 w_0(x) + K_2, \forall x \in \mathsf{X}.$$

Proof. Due to Assumption 4.26, for any $\lambda \in (\gamma_1, 1)$, we have

$$\mathcal{U}_x(w_1) \leq \lambda w_1(x), \forall x \in \mathsf{B}_{w_1}^c(A), A := \frac{K_1}{\lambda - \gamma_1}.$$

It implies that for all $x \in \mathsf{B}_{w_1}^c(A)$,

$$\begin{aligned} &\int_{\mathsf{B}_{w_1}^c(\lambda w_1(x))} P_x(dy) \left(e^{w_1(y) - \lambda w_1(x)} - 1\right) \\ &\leq \int_{\mathsf{B}_{w_1}(\lambda w_1(x))} P_x(dy) \left(1 - e^{w_1(y) - \lambda w_1(x)}\right). \end{aligned} \tag{4.35}$$

Taking some $\gamma_2 \in (\lambda^p, 1)$, by the definition of w_0, we have then

$$\mathsf{B}_{w_0}^c(\gamma_2 w_0(x)) \subset \mathsf{B}_{w_1}^c(\lambda w_1(x)), \forall x \in \mathsf{X}. \tag{4.36}$$

Indeed, for any $y \in \mathsf{B}_{w_0}^c(\gamma_2 w_0(x))$, it satisfies $w_0(y) > \gamma_2 w_0(x)$, which is equivalent to $w(y) > (\gamma_2)^{1/p} w_1(x) > \lambda w_1(x)$. Hence, $y \in \mathsf{B}_{w_1}^c(\lambda w_1(x))$ as well.

LEMMA **4.28.** *For any $\eta \in (0, 1-\lambda)$, $p \in (0,1)$ and $\gamma_2 \in ((\lambda+\eta)^p, 1)$, there exists a constant $R_1 > 0$ such that for all $y \in \mathsf{B}^c_{w_0}(\gamma_2 w_0(x))$, $x \in \mathsf{B}^c_{w_1}(R)$ and $R \geq R_1$,*

$$e^{w_0(y)+C+\eta w_1(x)}\left(w_0(y) - \gamma_2 w_0(x)\right) \leq e^{w_1(y)-\lambda w_1(x)} - 1.$$

Proof. It is sufficient to show that there exists a constant $R_1 > 0$ satisfying

$$w_0(y) + \log w_0(y) + C + \log 2 + \eta w_1(x) \leq w_1(y) - \lambda w_1(x) \tag{4.37}$$

for all $y \in \mathsf{B}^c_{w_0}(\gamma_2 w_0(x))$, $x \in \mathsf{B}^c_{w_1}(R)$ and $R \geq R_1$. Note that for any $p \in (0,1)$ and $\epsilon \in (0,1)$, there exists a constant D (depending on p and ϵ) satisfying

$$x^p + p\log x \leq \epsilon x + D, \forall x \geq 1,$$

which implies that $w_0(x) + \log w_0(x) \leq \epsilon w_1(x) + D, \forall x \in \mathsf{X}$. Hence, for all $y \in \mathsf{B}^c_{w_0}(\gamma_2 w_0(x))$, we have

$$\begin{aligned}
&w_1(y) - w_0(y) - \log w_0(y) - (\lambda+\eta) w_1(x) \\
\geq &(1-\epsilon) w_1(y) - (\lambda+\eta) w_1(x) - D \geq \left((1-\epsilon)\gamma_2^{1/p} - \lambda - \eta\right) w_1(x) - D.
\end{aligned}$$

Choosing $\gamma_2 \in ((\lambda+\eta)^p, 1)$, $\epsilon < 1 - \frac{\lambda+\eta}{\gamma_2^{1/p}}$ and $R_1 := \frac{D+C+\log 2}{(1-\epsilon)\gamma_2^{1/p} - \lambda - \eta}$, (4.37) holds for all $y \in \mathsf{B}^c_{w_0}(\gamma_2 w_0(x))$, $x \in \mathsf{B}^c_{w}(R)$ and $R \geq R_1$. □

LEMMA **4.29.** *For any $\eta > 0$, $p \in (0,1)$ and $C \geq 0$, there exists a constant R_2 such that for all $y \in \mathsf{B}_{w_1}(\lambda w_1(x))$, $x \in \mathsf{B}^c_{w_1}(R)$ and $R \geq R_2$,*

$$e^{-w_0(y)+\eta w_1(x)-C}\left(\gamma_2 w_0(x) - w_0(y)\right) \geq 1 - e^{w_1(y)-\lambda w_1(x)}. \tag{4.38}$$

Proof. It is sufficient to show that $e^{-w_0(y)+\eta w_1(x)-C}\left(\gamma_2 w_0(x) - w_0(y)\right) \geq 1$ under the same condition. Note that there exists a constant $D > 0$ such that

$$\frac{\gamma_2}{\eta} x^p \leq x + D, \forall x \geq 1,$$

which yields $-w_0(y) + \eta w_1(x) - C \geq -w_0(y) + \gamma_2 w_0(x) - C - D$ and hence,

$$e^{-w_0(y)+\eta w_1(x)-C}\left(\gamma_2 w_0(x) - w_0(y)\right) \geq e^{\gamma_2 w_0(x) - w_0(y) - C - D}\left(\gamma_2 w_0(x) - w_0(y)\right).$$

For all $y \in \mathsf{B}_{w_1}(\lambda w_1(x))$, we have $\gamma_2 w_0(x) - w_0(y) \geq (\gamma_2 - \lambda^p) w_0(x)$. Hence,

$$e^{\gamma_2 w_0(x) - w_0(y) - C - D}\left(\gamma_2 w_0(x) - w_0(y)\right) \geq e^{(\gamma_2 - \lambda^p) w_0(x) - C - D}(\gamma_2 - \lambda^p) w_0(x).$$

Due to the fact that $g(x) = e^x \cdot x$ is an increasing function on $\mathbb{R}_+$, we can choose $\tilde{R}_2 > 0$ such that $e^{\tilde{R}_2} \cdot \tilde{R}_2 = e^{C+D}$. Hence, we have for all $y \in \mathsf{B}_{w_1}(\lambda w_1(x))$, $x \in \mathsf{B}^c_{w_0}(\tilde{R})$ and $\tilde{R} \geq \tilde{R}_2$, $e^{-w_0(y)+\eta w_1(x)-C}\left(\gamma_2 w_0(x) - w_0(y)\right) \geq 1$ holds. Finally, setting $R_2 = \tilde{R}_2^{1/p}$, the assertion is obtained. □

Hence, by Lemma 4.28 and 4.29, for all $x \in \mathsf{B}^c_{w_1}(R_1 \vee R_2 \vee A)$,

$$\begin{aligned}
&\int_{\mathsf{B}_{w_0}(\gamma_2 w_0(x))} P_x(dy) e^{w_0(y)+C+\eta w_1(x)} \left(w_0(y) - \gamma_2 w_0(x)\right) \\
\text{(Lemma 4.28)} \quad &\leq \int_{\mathsf{B}_{w_0}(\gamma_2 w_0(x))} P_x(dy) \left(e^{w_1(y)-\lambda w_1(x)} - 1\right) \\
(4.36) \quad &\leq \int_{\mathsf{B}_{w_1}(\lambda w_1(x))} P_x(dy) \left(e^{w_1(y)-\lambda w_1(x)} - 1\right) \\
(4.35) \quad &\leq \int_{\mathsf{B}^c_{w_1}(\lambda w_1(x))} P_x(dy) \left(1 - e^{w_1(y)-\lambda w_1(x)}\right) \\
\text{(Lemma 4.29)} \quad &\leq \int_{\mathsf{B}^c_{w_1}(\lambda w_1(x))} P_x(dy) e^{-w_0(y)+\eta w_1(x)-C} \left(\gamma_2 w_0(x) - w_0(y)\right) \\
(4.36) \quad &\leq \int_{\mathsf{B}^c_{w_0}(\gamma_2 w_0(x))} P_x(dy) e^{-w_0(y)+\eta w_1(x)-C} \left(\gamma_2 w_0(x) - w_0(y)\right),
\end{aligned}$$

which implies that for all $f \in \mathscr{B}_{w_0}$ satisfying $|f| \leq w_0 + C$,

$$\int P_x(dy) e^{f(y)} \left(w_0(y) - \gamma_2 w_0(x)\right) \leq 0, \forall x \in \mathsf{B}^c_{w_1}(R_1 \vee R_2 \vee A). \tag{4.39}$$

Finally, for all $x \in \mathsf{B}_{w_1}(R_1 \vee R_2 \vee A)$ and $f \in \mathscr{B}_{w_0}$ satisfying $|f| \leq w_0 + C$,

$$\frac{P_x(e^f w_0)}{P_x(e^f)} \leq \frac{P_x(e^{w_0+C} w_0)}{P_x(e^{-w_0-C})} \leq e^{2C} P_x(e^{w_0} w_0) \cdot P_x(e^{w_0})$$

Using the fact that there exists some constant $D > 0$ satisfying

$$x^p + p \log x \leq x + D, \forall x \geq 1,$$

we obtain that $P_x(e^{w_0} w_0) \leq e^D P_x(e^{w_1})$ which is upper bounded on $\mathsf{B}_{w_1}(R_1 \vee R_2 \vee A)$. Hence, there exists a $K_2 > 0$ such that for all $f \in \mathscr{B}_{w_0}$ satisfying $|f| \leq w_0 + C$,

$$\frac{P_x(e^f w_0)}{P_x(e^f)} \leq K_2, \forall x \in \mathsf{B}_{w_1}(R_1 \vee R_2 \vee A),$$

which together with (4.39) implies the required inequality. □

Remark 4.30. The statement of Theorem 4.27 can be easily generalized to: for any positive C and A, there exist constants $\gamma_2 \in (0,1)$ and $K_2 \in \mathbb{R}_+$ such that

$$\sup_{f: f \in \mathscr{B}_{w_0}, |f| \leq A w_0 + C} \frac{P_x(e^f w_0)}{P_x(e^f)} \leq \gamma_2 w_0(x) + K_2.$$

Corollary **4.31.** *Suppose that Assumption 4.26 holds. Then, for any $p \in (0,1)$, $w_0 := w_1^p$, there exist constants $\hat{\gamma}_0 \in (0,1)$ (depending on p and γ_1) and $\hat{K}_0$ (depending on p, γ_1 and K_1) satisfying $\mathcal{U}_x(w_0) \leq \hat{\gamma}_0 w_0(x) + \hat{K}_0, \forall x \in \mathsf{X}$.*

Proof. By Proposition 4.23(i), $\mathcal{U}_x(w_0) = \log P_x(e^{w_0}) \leq \frac{P_x(e^{w_0} w_0)}{P_x(e^{w_0})}, \forall x \in \mathsf{X}$. Then, by Theorem 4.27, there exist constants $\hat{\gamma}_0 \in (0,1)$ (depending on p and γ_1) and $\hat{K}_0 > 0$ (depending on p, γ_1 and K_1) such that

$$\frac{P_x(e^{w_0} w_0)}{P_x(e^{w_0})} \leq \sup_{f \in \mathscr{B}_{w_0} : |f| \leq w_0} \frac{P_x(e^f) w_0}{P_x(e^f)} \leq \hat{\gamma}_0 w_0 + \hat{K}_0,$$

which yields the required inequality. □

In summary, if Assumption 4.26 holds, then

a) by Corollary 4.31, w_0 is a Lyapunov function w.r.t. the entropic map with constants $\hat{\gamma}_0$ and $\hat{K}_0$;

b) by Theorem 4.27, the same w_0 is also a Lyapunov function with constants γ_2 and K_2, which satisfies satisfying Assumption 4.17(i) (see Remark 4.13(a) and (b)) if the cost function c satisfies $|c| \leq \tilde{\gamma}_0 w_0 + C_0$ with some $\tilde{\gamma}_0 \in (0, 1-\gamma_2)$ and $C_0 > 0$;

c) combining (a) and (b), Assumption 4.17(i) holds also.

4.4.2 Minorization properties

We investigate now the properties of the entropic map restricted to bounded level-sets. We introduce first the *local Doeblin's condition* (see Douc et al. (2009) and references therein) as follows.

Assumption 4.32. Let $w_0 : \mathsf{X} \to [0, \infty)$ be a $\mathcal{B}(\mathsf{X})$-measurable function. For any level-set $\mathsf{C} := \mathsf{B}_{w_0}(R)$, $R > 0$, there exist a measure μ_{C} and constants $\lambda_{\mathsf{C}}^+, \lambda_{\mathsf{C}}^- > 0$ such that $\mu_{\mathsf{C}}(\mathsf{C}) > 0$ and

$$\lambda_{\mathsf{C}}^- \mu_{\mathsf{C}}(\mathsf{A} \cap \mathsf{C}) \leq P_x(\mathsf{A} \cap \mathsf{C}) \leq \lambda_{\mathsf{C}}^+ \mu_{\mathsf{C}}(\mathsf{A} \cap \mathsf{C}), \forall x \in \mathsf{C}, \mathsf{A} \in \mathcal{B}(\mathsf{X}).$$

The following proposition indicates the connection to the standard Doeblin's condition.

Proposition **4.33.** *The following two conditions are equivalent:*

(i) there exist a measure μ_{C} and a constant $\lambda_{\mathsf{C}}^- > 0$ such that $\mu_{\mathsf{C}}(\mathsf{C}) > 0$ and

$$P_x(\mathsf{A} \cap \mathsf{C}) \geq \lambda_{\mathsf{C}}^- \mu_{\mathsf{C}}(\mathsf{A} \cap \mathsf{C}), \forall x \in \mathsf{C}, \mathsf{A} \in \mathcal{B}(\mathsf{X}). \tag{4.40}$$

(ii) there exist a probability measure μ and a constant $\alpha > 0$ such that

$$P_x(\mathsf{A}) \geq \alpha \mu(\mathsf{A}), \forall x \in \mathsf{C}, \mathsf{A} \in \mathcal{B}(\mathsf{X}). \tag{4.41}$$

Proof. First, it is clear that (4.41) implies (4.40). Second, assume that (4.40) holds. Then, $\mu(\cdot) := \frac{\mu_{\mathsf{C}}(\mathsf{C} \cap \cdot)}{\mu_{\mathsf{C}}(\mathsf{C})}$, and $\alpha := \lambda_{\mathsf{C}}^- \mu_C(\mathsf{C})$ satisfy (4.41). □

THEOREM **4.34.** *Suppose Assumption 4.32 and Assumption 4.26 hold. Let w_1 be the Lyapunov function and $\mathsf{B} = \mathsf{B}_{w_0}(R_0)$ be a bounded levels-set with some $R_0 > 0$, where $w_0 := w_1^p$, $p \in (0,1)$. Then for any positive constant $K_0 > 0$, there exists a positive constant $\tilde{K}_0 > 0$ such that for all $v \in \mathscr{B}_{w_0}$ satisfying $|v| \le w_0 + \tilde{K}_0$, the following inequality holds*

$$\mathcal{U}_x(v) - \mathcal{U}_y(v) \le 2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0), \forall x, y \in \mathsf{B}.$$

Proof. Let $\mathsf{C} := \mathsf{B}_{w_0}(R) \supset B = \mathsf{B}_{w_0}(R_0)$ with $R > R_0$. Then

$$\frac{P_x(e^v)}{P_y(e^v)} = \frac{P_x(e^v \mathbf{1}_{\mathsf{C}}) + P_x(e^v \mathbf{1}_{\mathsf{C}^c})}{P_y(e^v \mathbf{1}_{\mathsf{C}}) + P_y(e^v \mathbf{1}_{\mathsf{C}^c})} \le \frac{P_x(e^v \mathbf{1}_{\mathsf{C}}) + P_x(e^v \mathbf{1}_{\mathsf{C}^c})}{P_y(e^v \mathbf{1}_{\mathsf{C}})}. \tag{4.42}$$

We first consider the second quotient. By $|v| \le \tilde{K}_0 + w_0$, we obtain

$$\frac{P_x(e^v \mathbf{1}_{\mathsf{C}^c})}{P_y(e^v \mathbf{1}_{\mathsf{C}})} \le e^{2\tilde{K}_0} \frac{P_x(e^{w_0} \mathbf{1}_{\mathsf{C}^c})}{P_y(e^{-w_0} \mathbf{1}_{\mathsf{C}})} = e^{2\tilde{K}_0} \frac{\theta(x, \mathsf{C}) P_x(e^{w_0})}{\theta'(y, \mathsf{C}) P_y(e^{-w_0})}$$

where we define

$$\theta(x, \mathsf{C}) := \frac{P_x(e^{w_0} \mathbf{1}_{\mathsf{C}^c})}{P_x(e^{w_0})} \text{ and } \theta'(y, \mathsf{C}) := \frac{P_y(e^{-w_0} \mathbf{1}_{\mathsf{C}})}{P_y(e^{-w_0})}.$$

By Theorem 4.27, there exist some constants $\gamma_2 \in (0,1)$ and $K_2 > 0$ such that

$$\begin{aligned} \theta(x, \mathsf{C}) &\le \|\mathbf{1}_{\mathsf{C}^c}\|_{w_0} \frac{P_x(e^{w_0} w_0)}{P_x e^{w_0}} \\ &\le \|\mathbf{1}_{\mathsf{C}^c}\|_{w_0} \sup_{|v| \le w_0} \frac{P_x(e^v w_0)}{P_x e^v} \le \|\mathbf{1}_{\mathsf{C}^c}\|_{w_0} (\gamma_2 w_0(x) + K_2). \end{aligned}$$

Hence,

$$\theta(x, \mathsf{C}) \le \|\mathbf{1}_{\mathsf{C}^c}\|_{w_0} \sup_{x \in \mathsf{B}} (\gamma_2 w_0(x) + K_2) \le \frac{\gamma_2 R_0 + K_2}{R}.$$

Similarly, we have

$$\theta'(y, \mathsf{C}) = 1 - \frac{P_y(e^{-w_0} \mathbf{1}_{\mathsf{C}^c})}{P_y(e^{-w_0})} \ge 1 - \frac{\gamma_2 R_0 + K_2}{R}$$

Hence, $\sup_{x,y \in \mathsf{B}} \frac{\theta(x,\mathsf{C})}{\theta'(y,\mathsf{C})} \to 0$ as $R \to \infty$, which implies that for any $K_0 > 0$, we can select sufficiently large R such that

$$\log \frac{\theta(x, \mathsf{C})}{\theta'(y, \mathsf{C})} \le -2K_0 - \log 2, \forall x, y \in \mathsf{B}. \tag{4.43}$$

Thus for any positive $\tilde{K}_0$ and K_0, there exists a sufficiently large R (depending on K_0) such that

$$\frac{P_x(e^v \mathbf{1}_{\mathsf{C}^c})}{P_y(e^v \mathbf{1}_{\mathsf{C}})} \le e^{2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0) - \log 2}. \tag{4.44}$$

Now we consider the first quotient in (4.42). By Assumption 4.32, we immediately have $\frac{P_x(e^v \mathbf{1}_C)}{P_y(e^v \mathbf{1}_C)} \le \frac{\lambda_C^+}{\lambda_C^-}$. Hence, setting

$$\tilde{K}_0 := K_0 + \frac{1}{2}\log 2 + \log(\frac{\lambda_C^+}{\lambda_C^-}), \tag{4.45}$$

we obtain

$$\frac{P_x(e^v \mathbf{1}_C)}{P_y(e^v \mathbf{1}_C)} \le e^{2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0) - \log 2}.$$

Together with (4.44), it yields the required inequality:

$$\frac{P_x(e^v)}{P_y(e^v)} \le e^{2(\tilde{K}_0 - K_0) + \mathcal{U}_x(w_0) - \mathcal{U}_y(-w_0)},$$

where $\tilde{K}_0$ is chosen according to (4.45), while R is determined by (4.43). □

We investigate now the minorization property of the upper envelope $\bar{\mathcal{U}}^{(w,C)}$ of the entropic map $\mathcal{U}$, which is required by Assumption 4.17(iii).

Proposition **4.35.** *Let $w : \mathsf{X} \to [1,\infty)$ be a $\mathcal{B}(\mathsf{X})$-measurable function and $\mathsf{B} := \mathsf{B}_w(R)$ with some $R > 0$. Suppose Assumption 4.32 holds. Assume further that $\bar{\mathcal{U}}_x^{(w,C)}(w_0) < \infty$ for all $x \in \mathsf{B}$. Then there exist a constant $\alpha \in (0,1)$ and a probability measure on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ satisfying*

$$\bar{\mathcal{U}}_x^{(w,C)}(v) - \bar{\mathcal{U}}_x^{(w,C)}(u) \ge \alpha\mu(v-u), \forall x \in \mathsf{B}, v \ge u \in \mathscr{B}_{1+w_0}.$$

Proof. Note that since $\bar{\mathcal{U}}_x^{(w,C)}(w_0) < \infty$, we have for all $v \in \mathscr{B}_{1+w_0}$ and $x \in \mathsf{B}$,

$$|\bar{\mathcal{U}}_x^{(w,C)}(v)| \le \bar{\mathcal{U}}_x^{(w,C)}(|v|) \le \|v\|_{1+w_0} \bar{\mathcal{U}}_x^{(w,C)}(1+w_0) < \infty.$$

By (4.34), we have for all $x \in \mathsf{B}$ and $v \ge u \in \mathscr{B}_{1+w_0}$,

$$\begin{aligned}\bar{\mathcal{U}}_x^{(w,C)}(v) - \bar{\mathcal{U}}_x^{(w,C)}(u) &= \sup_{h \in \mathscr{B}_w^{(C)}} \frac{P_x(e^h v)}{P_x(e^h)} - \sup_{h' \in \mathscr{B}_w^{(C)}} \frac{P_x(e^{h'}u)}{P_x(e^{h'})} \\ &= \inf_{h' \in \mathscr{B}_w^{(C)}} \left(\sup_{h \in \mathscr{B}_w^{(C)}} \frac{P_x(e^h v)}{P_x(e^h)} - \frac{P_x(e^{h'}u)}{P_x(e^{h'})} \right) \ge \inf_{h' \in \mathscr{B}_w^{(C)}} \frac{P_x\left(e^{h'}(v-u)\right)}{P_x(e^{h'})}.\end{aligned}$$

By Proposition 4.33, Assumption 4.32 implies that there exist a probability measure μ_B and α_B such that $P_x(v) \ge \alpha_\mathsf{B}\mu_\mathsf{B}(v)$ for all nonnegative measurable function v. Hence, for all $x \in \mathsf{B}$ and $h' \in \mathscr{B}_w^{(C)}$, we have

$$\begin{aligned}\frac{P_x\left(e^{h'}(v-u)\right)}{P_x(e^{h'})} &\ge \frac{\alpha_\mathsf{B}\mu_\mathsf{B}\left(e^{h'}(v-u)\right)}{P_x(e^{Cw})} \ge \frac{\alpha_\mathsf{B}\mu_\mathsf{B}\left(e^{-Cw}(v-u)\right)}{\max_{x\in\mathsf{B}} P_x(e^{Cw})} \\ &= \frac{\alpha_\mathsf{B}\mu_\mathsf{B}(e^{-Cw})}{\max_{x\in\mathsf{B}} P_x(e^{Cw})} \frac{\mu_\mathsf{B}\left(e^{-Cw}(v-u)\right)}{\mu_\mathsf{B}\left(e^{-Cw}\right)}\end{aligned}$$

Hence, $\alpha := \frac{\alpha_\mathsf{B}\mu_\mathsf{B}(e^{-Cw})}{\max_{x\in\mathsf{B}} P_x(e^{Cw})}$ and the probability measure $d\mu := \frac{e^{-Cw}d\mu_\mathsf{B}}{\int e^{-Cw}d\mu_\mathsf{B}}$ are the required constant and probability measure respectively. □

The following theorem shows that applying the entropic map, together with an additional growth condition for cost functions (see (4.46) below), Assumption 4.26 and 4.32 are sufficient for Assumption 4.12 and 4.17.

THEOREM **4.36.** *Let $\mathcal{U}$ be the entropic map with $\lambda = 1$. Suppose Assumption 4.26 and 4.32 hold and w_1 is the Lyapunov function. If the reward function r satisfies*

$$r \in \mathscr{B}_{w_1^q} \text{ with some } q \in (0,1), \tag{4.46}$$

then Assumption 4.12 holds with $w_0 = w_1^p$ for any $p \in (q,1)$, and some $\tilde{K}_0 > 0$, and Assumption 4.17 holds with w_0 and $w = 1 + \tilde{K}_0^{-1} w_0$.

Proof. Fix one $p \in (q,1)$ and let $w_0 = w_1^p$. Then by Corollary 4.31, there exists $\hat{\gamma}_0 \in (0,1)$ and $\hat{K}_0 > 0$ satisfying $\mathcal{U}_x(w_0) \leq \hat{\gamma}_0 w_0(x) + \hat{K}_0$. By assumption, there exists some $C > 0$ and $q \in (0,1)$ such that $|c| \leq C w_1^q$. Choosing one $\gamma_0^{(c)} \in (0, 1 - \hat{\gamma}_0)$, there exists a constant $K_0^{(c)} > 0$ satisfying $C w_1^q(x) \leq \gamma_0^{(c)} w_0(x) + K_0^{(c)}$. Hence, Assumption 4.12(i) holds with $\gamma_0 := \hat{\gamma}_0 + \gamma_0^{(c)} \in (0,1)$ and $K_0 := \hat{K}_0 + K_0^{(c)}$. Due to Proposition 4.33, Assumption 4.12(ii) holds with some constant $\tilde{K}_0 > 0$. Next, by Theorem 4.14 and 4.27, Assumption 4.17(i) and (ii) hold with $w := 1 + \tilde{K}_0^{-1} w_0$ and $C := \tilde{K}_0^{-1}$. Assumption 4.17(iii) holds due to Proposition 4.35. □

Comparison with literature

Hence, for the entropic map, the required sufficient conditions in Assumption 4.12 and 4.17 can replaced by the existence of Lyapunov function in Assumption 4.26, the local Doeblin's condition in Assumption 4.32 and the growth condition for the cost function in (4.46). We compare our results with the mostly related literature Kontoyiannis and Meyn (2005).

Among others, Kontoyiannis and Meyn (2005) developed (see also their earlier work on the same topic: Kontoyiannis and Meyn, 2003) a spectral theory of multiplicative Markov processes, where the Poisson equation w.r.t. the entropic map (called multiplicative Poisson equation in Kontoyiannis and Meyn, 2005) plays the central role. Though our assumptions are less general than the assumptions stated in Kontoyiannis and Meyn (2005, 2003), our proof that generalizes the Hairer-Mattingly approach (Hairer and Mattingly, 2011) is conceptually simpler than the one provided in Kontoyiannis and Meyn (2005, 2003), and can also be applied to other types of valuation maps. Note again that, in our approach, the convergence rate of iterations towards the solution to the Poisson equation is explicitly specified by $\bar{\alpha}$ in Lemma 4.18 under the chosen seminorm.

4.4.3 An example with AR1 processes

Consider again the 1-dimensional simple autoregressive model (cf. (4.4))

$$X_{t+1} = \delta X_t + \sigma N_t$$

with some $\delta \in (-1,1)$, $\sigma > 0$ and N_t being standard is i.i.d. white noise. For one entropy map

$$\mathcal{U}^{(\lambda)}(v) = \frac{1}{\lambda}\log\left(P_x(e^{\lambda v})\right)$$

with $\lambda \neq 0$, it is sufficient to consider its convex counterpart, i.e., $\mathcal{U}^{(|\lambda|)}(v)$ for checking conditions assumed in this section. Furthermore, without loss of generality, we assume $\lambda = 1$.

On the one hand, we have already shown in Page 40 that $w_1 = 1 + \epsilon x^2$ is a Lyapunov function satisfying

$$\log(P_x(e^{w_1})) \leq \gamma w_1(x) + K, \forall x \in \mathsf{X}$$

with some constants $\gamma \in (0,1)$ and $K > 0$. Hence, Assumption 4.26 holds.

One the other hand, the transition kernel of the underlying Markov chain is

$$P(dy|x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y-\delta x)^2}{2\sigma^2})dy.$$

Hence, the density function (with respect to the Lebesgue measure) exists

$$p(y|x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(y-\delta x)^2}{2\sigma^2})$$

which is apparently upper bounded and lower bounded away from 0 on any bounded level set $(x,y) \in \mathsf{C}^2$, where $\mathsf{C} := \{x \in \mathbb{R} | |x| \leq R\}$ with $R > 0$. Hence, Assumption 4.32 holds.

Therefore, by Theorem 4.36, the existence of solutions to Poisson equation is guaranteed, if the reward function $r \in \mathscr{B}_{w_1^q}$ for any $q \in (0,1)$.

4.5 Finite state spaces

In this section, we consider the case of finite state spaces, which is the most widely used setting in the field of machine learning. It implies immediately that rewards are bounded, i.e., $\max_{x\in\mathsf{X}}|r(x)| < \infty$. Therefore, the weight function considered in this section is $w \equiv 1$. For simplicity, the seminorm $\|\cdot\|_{s,1}$ is written as $\|\cdot\|_s$ and the norm $\|\cdot\|_1$ as its conventional notation $\|\cdot\|_\infty$. Note that the case of finite state spaces can be considered as a special case of bounded rewards. Hence, for the sake of generality, we consider still general state spaces with, however, bounded rewards. We first show that each map $\mathcal{U}$ is nonexpansive under the norm and also the seminorm.

Nonexpansiveness

Proposition **4.37** (cf. Proposition 1.1 in Gunawardena and Keane, 1995)**.** *For all* $f, g \in \mathscr{B}$,

(i) $\|\mathcal{U}(f) - \mathcal{U}(g)\|_s \leq \|f - g\|_s$,

(ii) $\|\mathcal{U}(f) - \mathcal{U}(g)\|_\infty \leq \|f - g\|_\infty$.

Proof. It is sufficient to show that

$$\begin{aligned} &\sup_{x\in\mathsf{X}} (\mathcal{U}_x(f) - \mathcal{U}_x(g)) \leq \sup_{x\in\mathsf{X}} (f(x) - g(x)) \\ \text{and} \quad &\inf_{x\in\mathsf{X}} (\mathcal{U}_x(f) - \mathcal{U}_x(g)) \geq \inf_{x\in\mathsf{X}} (f(x) - g(x)) \end{aligned}$$

Below, we show only the first inequality. The second one can be obtained analogously. By

$$f(x) - g(x) \leq \sup_{x\in\mathsf{X}} (f(x) - g(x)) \, \forall x \in \mathsf{X}$$

and the monotonicity and the translation invariance of $\mathcal{U}$, we obtain for each $x \in \mathsf{X}$

$$\mathcal{U}_x(f) \leq \mathcal{U}_x(g + \sup_{x\in\mathsf{X}} (f(x) - g(x))) = \mathcal{U}_x(g) + \sup_{x\in\mathsf{X}} (f(x) - g(x)) \,,$$

which implies the required inequality. □

This proposition implies immediately the operator $\mathcal{T}(f) := r + \mathcal{U}(f)$ is also nonexpansive under both the norm and the seminorm for any $r, f \in \mathscr{B}$. Hence, we have for any $f, g \in \mathscr{B}$,

$$\frac{1}{n}\|\mathcal{T}^n(f) - \mathcal{T}^n(g)\|_\infty \leq \frac{1}{n}\|f - g\|_\infty \to 0.$$

Therefore, to ensure the existence of solutions to the Poisson equation, it remains the verify the contraction property.

Multistep contraction

For Markov chains with finite state spaces, an equivalent condition to ensure the ergodicity is (see e.g. Häggström, 2002, Corollary 4.1): there exist constants $n_0 \in \mathbb{N}$, $\alpha \in (0,1)$ such that $\mathbb{P}(X_{n_0} = y | X_0 = x) \geq \alpha, \forall x, y \in \mathsf{X}$. In other words, for any two states x and y, there always exists a non-zero chance such that y is visited after n_0 steps when starting from x. For general state spaces, this condition can be reformulated as:

Assumption 4.38. There exist constants $n_0 \in \mathbb{N}$, $\alpha \in (0,1]$ and a probability measure μ on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ such that for each $\mathsf{B} \in \mathcal{B}(\mathsf{X})$ and $x \in \mathsf{X}$,

$$\mathbb{P}(X_{n_0} \in \mathsf{B} | X_0 = x) = P_x^{n_0}(\mathsf{B}) \geq \alpha\mu(\mathsf{B}).$$

This assumption can be viewed as a generalized multistep version of the Doeblin's condition in Assumption 4.3(ii). Inspired by Assumption 4.38, we may generalize the conditions stated in Theorem 4.6 to the multistep setting.

THEOREM **4.39.** *Suppose the operator $\mathcal{T}(\cdot) = r + \mathcal{U}(\cdot)$, $r \in \mathscr{B}$, satisfies*

$$\|\mathcal{T}^{n_0}(f) - \mathcal{T}^{n_0}(g)\|_s \leq \bar{\alpha}\|f - g\|_s, \forall f, g \in \mathscr{B} \tag{4.47}$$

with some positive constants $n_0 \in \mathbb{N}$ and $\bar{\alpha} \in \mathbb{R}$. Then there exist

(i) a solution $(\rho, h) \in \mathbb{R} \times \mathscr{B}$ to the Poisson equation $r + \mathcal{U}(h) = \rho + h$, where ρ is unique and

(ii) a valuation function ν satisfying

$$\nu(r + \mathcal{U}(f)) = \nu(f) + \rho, \forall f \in \mathscr{B}.$$

Proof. Note that $\{\mathcal{T}^{in_0}(f), i = 1, 2, \ldots\}$ is a Cauchy sequence in the quotient space $\tilde{\mathscr{B}}$ implies that $\{\mathcal{T}^{i}(f), i = 1, 2, \ldots\}$ is also a Cauchy sequence. The rest of the proof follows the same line of the proof of Theorem 4.6 and is therefore omitted. □

For sufficient conditions, under which the multistep contraction (4.47) holds, we consider first a special case, convex and homogeneous maps and then followed by a treatment with general maps.

Convex and homogeneous maps

LEMMA **4.40.** *Suppose $\mathcal{U}$ is a convex homogeneous valuation map such that there exist a positive constant $\alpha \in (0, 1)$ and a probability measure $\nu \in \mathscr{M}$ satisfying*

$$\mathcal{U}_x^{n_0}(v) - \alpha\nu(v) - \mathcal{U}_x^{n_0}(u) + \alpha\nu(u) \geq 0, \forall x \in \mathsf{X}, \forall v \geq u \in \mathscr{B}. \tag{4.48}$$

Then

$$\|\mathcal{T}^{n_0}(v) - \mathcal{T}^{n_0}(u)\|_s \leq (1 - \alpha)\|v - u\|_s,$$

for all v and u in $\mathscr{B}$.

Proof. Define a new valuation map $\tilde{\mathcal{U}}_x(v) := \frac{1}{1-\alpha}\mathcal{U}_x^{n_0}(v) - \frac{\alpha}{1-\alpha}\nu(v)$. It is easy to verify that $\tilde{\mathcal{U}}$ is valid valuation map. In fact, the monotonicity is guaranteed by (4.48). Furthermore, it is also easy to check that $\tilde{\mathcal{U}}$ is also convex and homogeneous.

Let $\mathcal{F}(\cdot) := \mathcal{T}^{n_0}(\cdot) - \mathcal{T}^{n_0}(0)$. Then the assertion is equivalent to $\|\mathcal{F}(v) - \mathcal{F}(u)\|_s \leq (1 - \alpha)\|v - u\|_s$. Suppose $\|v - u\|_s = C$. Lemma 4.2 suggests that we can always find a real value c such that $\|v - u + c\|_\infty = C$. Since adding any constant to $v - u$ will not change the values of both sides of the required inequality, without loss of generality, we assume $\|v - u\|_\infty = C$.

By Proposition 2.9(ii) and the monotonicity of $\mathcal{U}$, we have

$$\mathcal{F}_x(v) - \mathcal{F}_x(u) \leq \mathcal{U}_x^{n_0}(v - u),$$

which implies that for all $x \in \mathsf{X}$

$$\begin{aligned}&\frac{1}{1-\alpha}\left(\mathcal{F}_x(v) - \mathcal{F}_x(u) - \alpha v(v-u)\right)\\ \leq &\frac{1}{1-\alpha}\mathcal{U}_x^{n_0}(v-u) - \frac{\alpha}{1-\alpha}v(v-u) = \tilde{\mathcal{U}}_x(v-u) \leq \tilde{\mathcal{U}}_x(|v-u|).\end{aligned}$$

Switching v and u, we obtain immediately

$$\frac{1}{1-\alpha}\left|\mathcal{F}_x(v) - \mathcal{F}_x(u) - \alpha v(v-u)\right| \leq \tilde{\mathcal{U}}_x(|v-u|) \leq C.$$

Hence, we have

$$\begin{aligned}&\left|\mathcal{F}_x(v) - \mathcal{F}_x(u) - \mathcal{F}_y(v) + \mathcal{F}_y(u)\right|\\ =&\left|\mathcal{F}_x(v) - \alpha v(v) - \mathcal{F}_x(u) + \alpha v(u) - \mathcal{F}_y(v) + \alpha v(v) + \mathcal{F}_y(u) - \alpha v(u)\right|\\ \leq&\left|\mathcal{F}_x(v) - \mathcal{F}_x(u) - \alpha v(v-u)\right| + \left|\mathcal{F}_y(v) - \mathcal{F}_y(u) - \alpha v(v-u)\right|\\ \leq&2(1-\alpha)C,\end{aligned}$$

by which the required inequality follows. □

Proposition **4.41.** *Suppose the transition kernel P satisfies Assumption 4.38 with some positive constants $n_0 \in \mathbb{N}$, $\beta \in \mathbb{R}$, and probability measure μ, i.e.*

$$P_x^{n_0}(\mathsf{A}) \geq \beta\mu(\mathsf{A}), \forall x \in \mathsf{X}, \forall \mathsf{A} \in \mathcal{B}(\mathsf{X}).$$

Assume further that there exists $g(x,u) \in \delta\mathcal{U}_x(u)$ and positive constant $\epsilon > 0$ such that $g(x,u) \geq \epsilon$ for all $x \in \mathsf{B}$ and $u \in \mathscr{B}_w$. Then (4.48) holds for $\nu = \mu$ and $\alpha = \epsilon^{n_0}\beta$.

Proof. The proof is similar to the proof of Proposition 4.8 and is therefore ignored. □

An example Let $\{X_t\}$ be a time-homogeneous Markov chain satisfying Assumption 4.38. Then applying the semi-deviation map introduced in Section 2.4.7, which is,

$$\mathcal{U}_x(f) := P_x(f) - \lambda\sqrt{P_x[P_x(f) - f]_+^2},$$

where $\lambda \in [0,1)$ controls how risk-averse the agent is. Following the same line as in the similar example discussed in Section 4.3.1, it is easy to verify that the assumption of Proposition 4.41 holds and therefore Lemma 4.40 holds, which, by Theorem 4.39, ensures the existence of solutions to the Poisson equation and the existence of invariant valuation functions as well.

General maps

Recall that the subspace $\mathscr{B}_w^{(C)}$ of size $C > 0$ is defined as:

$$\mathscr{B}_w^{(C)} := \{v \in \mathscr{B}_w | \|v\|_{s,w} \leq C\}.$$

and a convex homogeneous valuation map $\bar{\mathcal{U}}^{(w,C)}$ is said to be an upper envelope of a valuation map $\mathcal{U}$ given a bound $C > 0$, if for all $v, u \in \mathscr{B}_w^{(C)}$,

$$\mathcal{U}_x(v) - \mathcal{U}_x(u) \leq \bar{\mathcal{U}}_x^{(w,C)}(v-u), \forall x \in \mathsf{X}.$$

For $w \equiv 1$, we write $\mathscr{B}_1^{(C)}$ and $\bar{\mathcal{U}}^{(1,C)}$ simply as $\mathscr{B}^{(C)}$ and $\bar{\mathcal{U}}^{(C)}$ respectively.

Assumption 4.42. (i) There exist positive constants $n_0 \in \mathbb{N}$ and $\tilde{K}_0 \in \mathbb{R}_+$ such that $\|\mathcal{T}^{n_0}(v)\|_s \leq \tilde{K}_0$, for all $v \in \mathscr{B}^{(\tilde{K}_0)}$.

(ii) There exist a constant $\alpha \in (0,1)$ and a probability measure μ on $(\mathsf{X}, \mathcal{B}(\mathsf{X}))$ such that

$$(\bar{\mathcal{U}}_x^{(\tilde{K}_0')})^{n_0}(v) - (\bar{\mathcal{U}}_x^{(\tilde{K}_0')})^{n_0}(u) \geq \alpha\mu(v-u), \forall x \in \mathsf{X},$$

holds for all $v \geq u \in \mathscr{B}^{(\tilde{K}_0')}$, where $\tilde{K}_0' = \tilde{K}_0 + (n_0 - 1)\|r\|_s$.

Lemma **4.43.** *Suppose Assumption 4.42 holds. Then*

$$\|\mathcal{T}^{n_0}(v) - \mathcal{T}^{n_0}(u)\|_s \leq (1-\alpha)\|v-u\|_s, \forall v, u \in \mathscr{B}_w^{(\tilde{K}_0)}.$$

Proof. By Assumption 4.42(i), the iteration $\{\mathcal{T}^{in_0}(v), i = 1,2,\ldots\}$ remains in $\mathscr{B}_w^{(\tilde{K}_0)}$, provided that we start with an element in $\mathscr{B}_w^{(\tilde{K}_0)}$. For $j = 1,2,\ldots,n_0 - 1$, we have

$$\begin{aligned}\|\mathcal{T}^j(v)\|_s \leq& \|\mathcal{T}^j(v) - \mathcal{T}^j(0)\|_s + \|\mathcal{T}^j(0)\|_s \\ \leq& \|v\|_s + j\|r\|_s \leq \tilde{K}_0 + (n_0 - 1)K = \tilde{K}_0'.\end{aligned}$$

Hence, starting with $v \in \mathscr{B}^{(\tilde{K}_0)}$, the iteration $\{\mathcal{T}^n(v), n = 1,2,\ldots\}$ remains in $\mathscr{B}_w^{(\tilde{K}_0')}$. Hence, we have for all $v, u \in \mathscr{B}_w^{(\tilde{K}_0')}$

$$\mathcal{T}_x^{n_0}(v) - \mathcal{T}_x^{n_0}(u) \leq \bar{\mathcal{U}}_x^{(\tilde{K}_0')}(\mathcal{T}_x^{n_0-1}(v) - \mathcal{T}_x^{n_0-1}(u)) \leq \ldots \leq (\bar{\mathcal{U}}_x^{(\tilde{K}_0')})^{n_0}(v-u).$$

The rest of the proof follows the same line as in the proof of Lemma 4.40 and is therefore omitted. □

By Theorem 4.39, this lemma immediately ensures the existence of solutions to the Poisson equation and invariant valuation functions (restricted to the subspace $\mathscr{B}_w^{(\tilde{K}_0)}$).

An example Let $\{X_t\}$ be a time-homogeneous Markov chain satisfying Assumption 4.38. Consider the utility-based shortfall defined in (4.25) satisfying (4.26). Let $\bar{\mathcal{U}}$ be the upper envelope defined in (4.29). Then $\bar{\mathcal{U}}^{n_0}$ is apparently an upper envelope of $\mathcal{R}(\cdot) := \mathcal{T}^{n_0}(\cdot) - \mathcal{T}^{n_0}(0)$. By (4.30), we obtain $\bar{\mathcal{U}}^{n_0}(v) - \bar{\mathcal{U}}^{n_0}(v') \geq \alpha \left(\frac{l}{L}\right)^{n_0} \mu(v - v')$ whenever $v \geq v' \in \mathscr{B}$. This verifies Assumption 4.42(ii). In addition, following the same line in the proof of Lemma 4.40, we obtain $\|\mathcal{R}(v)\|_s = \|\mathcal{T}^{n_0}(v) - \mathcal{T}^{n_0}(0)\|_s \leq (1 - \alpha \left(\frac{l}{L}\right)^{n_0})\|v\|_s$. Hence, provided $K_0 = \|r\|_s$, $\tilde{K}_0 = \frac{n_0 K_0}{\alpha} \left(\frac{L}{l}\right)^{n_0}$ verifies Assumption 4.42(i).

Conditions for the entropic map

Assumption 4.44. There exists a measure μ and positive constants $\lambda^+ > \lambda^- > 0$, $n_0 \in \mathbb{N}$ such that

$$\lambda^- \mu(\mathsf{A}) \leq P_x^{n_0}(\mathsf{A}) \leq \lambda^+ \mu(\mathsf{A}), \forall \mathsf{A} \in \mathcal{B}(\mathsf{X}).$$

Proposition **4.45.** *Suppose Assumption 4.44 holds. Then Assumption 4.42 holds for the entropic map with $\lambda = 1$.*

Proof. First, we show that Assumption 4.42(i) holds. Assume $\|r\|_s = K$. Define $\tilde{P}_x(dy) := e^{r(x)} P_x(dy)$ and $\tilde{P}_x(v) := \int v(y) \tilde{P}_x(dy)$. We have then $\frac{\tilde{P}_x(e^v)}{\tilde{P}_y(e^v)} \leq e^K \frac{P_x(e^v)}{P_y(e^v)}$. Hence,

$$e^{\mathcal{T}_x^{n_0}(v) - \mathcal{T}_y^{n_0}(v)} = \frac{\tilde{P}_x^{n_0}(e^v)}{\tilde{P}_y^{n}(e^v)} \leq e^{2n_0 K} \frac{P_x^{n_0}(e^v)}{P_y^{n_0}(e^v)} \leq e^{2n_0 K} \frac{\lambda^+}{\lambda^-},$$

where the last inequality is due to Assumption 4.44. Taking

$$\tilde{K}_0 := n_0 K + \frac{1}{2} \log \frac{\lambda^+}{\lambda^-},$$

we obtain the required inequality for Assumption 4.42(i).

Second, Assumption 4.42(ii) can be obtained by applying Proposition 4.35 with $w_0 \equiv 1$. □

Example 4.46. Let $\{X_t\}$ be a time-homogeneous ergodic Markov chain with a finite state space, i.e., it satisfies that $P^{n_0}(y|x) \geq \epsilon$ holds for some positive $n_0 \in \mathbb{N}$ and for all $x, y \in \mathsf{X}$. It yields also that $P^{n_0}(y|x) \leq 1 - \epsilon$. Without loss of generality, we may assume $\epsilon \in (0, \frac{1}{N})$, where N denotes the cardinality of X. Then, Assumption 4.44 holds with $\lambda^- = N\epsilon$, $\lambda^+ = N(1 - \epsilon)$ and μ being the uniform distribution on X.

Comparing with literature Cavazos-Cadena and Hernández-Hernández (2009) stated a set of necessary and sufficient conditions for a solution to the Poisson equation with the entropic map on finite state spaces: (a) the underlying Markov chain is a unichain, i.e., it contains a unique recurrent class $\mathsf{R} \subset \mathsf{X}$ and (b) there exists a positive integer n_0 such that $\mathbb{P}[T_{\mathsf{R}} \leq n_0 | X_0 = x] = 1$, $\forall x \in \mathsf{X}$, where $T_{\mathsf{R}} := \min\{n \geq 1 \mid X_n \in \mathsf{R}\}$ denotes the first hitting time. Our condition is obviously stronger than the above conditions. Nevertheless, under our assumptions, a geometric convergence under the semi-norm can be obtained by Lemma 4.43, whereas under the necessary and sufficient conditions, the convergence of such iterations $\{\mathcal{T}^n, n = 1, 2, \ldots\}$ is not guaranteed. For an example of a Markov chain satisfying the above two conditions with, however, non-converged iterations, see Example 8.5.1 of Puterman (1994).

4.6 Summary and discussion

In this chapter, we investigated the conditions under which

1) there exists a (unique) solution to the Poisson equation with arbitrary valuation maps and

2) a convergence rate can be quantified for the associated iterations.

To this end, we generalized the Lyapunov approach that was applied by Hairer and Mattingly (2011) to ensure the geometric ergodicity for Markov chains. Our assumptions and their variants for different types of valuations maps are composed of two conditions,

a) the existence of a Lyapunov function to control the growth of iterations, and

b) a Doeblin-like condition for local contraction.

In particular, for the same problem on finite state spaces, the above two conditions can be reduced to one condition, that is, a (possibly) multistep Doeblin-like condition for global contraction.

Our motivation of studying the nonlinear Poisson equation is to understand the limit behavior of average valuations, which plays a crucial role in the average Markov decision/control processes to be introduced in the next chapter. Due to its generality, the nonlinear Poisson equation may have other applications in mathematics. One example is that, as pointed out by Gaubert and Gunawardena (2004), it is closely linked to the nonlinear Perron-Frobenius theory (for a comprehensive introduction see Lemmens and Nussbaum, 2012) with nonexpansive and homogeneous maps. Another example is that, the upper envelope for the entropic map defined in (4.34) is closed related to the calculation of posterior distributions in Bayesian statistics, especially in a hidden Markov model (HMM, see e.g. Cappé et al., 2005). Its ergodicity property can be very useful for understanding the stability of the underlying HMM. For details see Douc et al. (2009).

We end this theoretical chapter with a final remark. In several places of this chapter, we started from the assumption that the underlying Markov chain is (geometrically) ergodic and then stated the sufficient conditions under which the valuation maps with this Markov chain are ergodic as well (see e.g. Proposition 4.8 for homogeneous maps and Theorem 4.34 for the entropic map). This means, the ergodicity of the underlying Markov chain does not necessarily imply the ergodicity of valuation maps. It is now remarkable that the opposite direction does not holds either: a non-ergodic objective transition kernel may allow an ergodic valuation map. For instance, consider the following 1-dimensional autoregressive model, $X_{t+1} = bX_t + N_t$, where N_t is standard white noise and assume that $1 < |b| \leq C < \infty$. Then, the transition kernel is

$$P(dy|x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y - bx)^2}{2}\right) dy$$

By assumption, $|b| > 1$, this Markov chain is transient. Now, we let $g(y) = -\frac{C}{2}y^2$ and define $\mathcal{U}_x(f) := \frac{P_x[e^g f]}{P_x[e^g]}$, which is a valid *linear* valuation map. Then it is easy to check that

$$\mathcal{U}_x(f) = \frac{\sqrt{1+C}}{\sqrt{2\pi}} \int f(y) \exp\left(-(C+1)\frac{(y - \frac{b}{1+C}x)^2}{2}\right) dy$$

Or in other words, the corresponding new transition kernel is

$$\tilde{P}(dy|x) = \frac{\sqrt{1+C}}{\sqrt{2\pi}} \exp\left(-(C+1)\frac{(y - \frac{b}{1+C}x)^2}{2}\right) dy,$$

which is w-geometric ergodic (for a proof, see e.g. Section 16.5.1 in Meyn and Tweedie, 1993). Hence, in summary, the valuation maps may *effectively* change the dynamics of the underlying Markov chain.

5

Risk-sensitive Markov Decision Processes

Nature has placed mankind under the governance of two sovereign masters, pain and pleasure. It is for them alone to point out what we ought to do, as well as to determine what we shall do.
— Jeremy Bentham (1789)

Précis We introduce a unified framework for measuring risk in the context of Markov decision processes with valuation maps on general Borel spaces. Within the framework, applying weighted norm spaces to incorporate also unbounded rewards, we study two types of infinite-horizon risk-sensitive criteria, discounted and average valuation, and solve the associated optimization problems by value iteration. For the discounted case, we propose a new discount scheme, which is different from the conventional form but consistent with existing literature, while for the average criterion, we state Lyapunov-type stability conditions that generalize known conditions for Markov chains to ensure the existence of solutions to the optimality equation and a geometric convergence rate for the value iteration.

Publications related to this chapter The main results of this chapter has been published in Shen et al., 2013, Section 4, Shen et al., 2014b, Section 5 and 6 and Shen et al., 2014c.

5.1 Introduction

Markov decision processes (MDPs, see e.g. Puterman, 1994; White, 1993 and Hernández-Lerma and Lasserre, 1996; 1999 under the name *Markov control processes*) are widely applied to model sequential decision making problems of agents. The induced optimal control problem is to find the best policy that maximizes the expected total rewards. The core of the MDP-framework consists of two *objective* descriptions of some mechanism of the environment *transition probabilities*

of switching states when performing actions, and immediate *outcomes* (rewards or costs) obtained at states by executing actions. Facing the same environment, however, different agents might have different policies. Therefore, in many applications, it is important to also incorporate the *subjective* perceptions of an agent into the MDP-framework. The subjective outcomes are usually modeled by *utility functions* (see e.g. Gollier, 2004), which can be easily incorporated by simply replacing the immediate outcome with its utility, whereas the subjective transition probabilities require a more sophisticated mathematical framework. They are commonly incorporated in the *risk*, which is caused by an uncertain environment.

Coherent/convex risk measures (CRMs) (Artzner et al., 1999; Föllmer and Schied, 2002) have been widely employed to model subjective probabilities in mathematical finance since the last one and half decades. Several works (see e.g. Roorda et al., 2005; Föllmer and Penner, 2006; Cheridito et al., 2006; Ruszczyński and Shapiro, 2006; Cheridito and Kupper, 2011 and references therein) extend CRMs to temporal structures in various setups, where they consider mainly finite-horizon problems. On the contrary, in the literature of MDPs, while the infinite-horizon risk-sensitive optimal decision-making or control problems are studied, they apply merely the *entropic map* (Chung and Sobel, 1987; Hernández-Hernández and Marcus, 1996; Fleming and Hernández-Hernández, 1997; Marcus et al., 1997; Avila-Godoy and Fernández-Gaucherand, 1998; Coraluppi and Marcus, 2000; Cavazos-Cadena, 2010; Borkar and Meyn, 2002; Di Masi and Stettner, 2008), which is convex and in fact a special type of CRM. All risk measures mentioned in the above literature are coherent/convex based on the assumption that the agent is supposed to be economically rational and therefore *risk-averse*. This limits applications in the fields of decision-making under risk and behavioral economics, where more general risk measures (see e.g. Savage, 1972; Chateauneuf and Cohen, 2008; Tversky and Kahneman, 1992 and references therein) are applied, since human beings are not always risk-averse. However, the models in these fields can only be applied to one-step decision making problems.

To overcome the limitations mentioned above, we have already extended the definition of CRMs in Chapter 2 (see Section 2.3), to include valuation functions considered also in behavioral economics. In this chapter, we will apply the constructive approach introduced Chapter 3 to the MDP framework. We have already shown that this approach maintains the Markov property, which ensures, therefore, the existence of stationary optimal policies for two infinite-horizon objectives, namely, the discounted and average criteria. With the generalized valuation functions and constructed valuation maps, we provide in this chapter a unified treatment in the context of MDPs to infinite-horizon risk-sensitive optimal control problems considered in various fields. Using weighted norm spaces, we can incorporate unbounded rewards in risk-sensitive MDPs also. We prove that two types of objectives, the discounted and the average valuations, can be optimized with *dynamic programming* algorithms under proper assumptions. For the case of discounted criterion, we apply a new discount scheme which is different from the conventional form but consistent with the one applied in Ruszczyński (2010)

where coherent risk measures are considered. For the average case, we state sufficient conditions, which generalize Lyapunov-type conditions from the literature of Markov chains (see e.g. Meyn and Tweedie, 1993), to ensure the existence of solutions to the associated optimality equation.

5.2 Markov decision processes

Notations A *Borel space* is a Borel subset of a complete separable metric space. If X is a Borel space, its Borel σ-algebra is denoted by $\mathcal{B}(\mathsf{X})$. Let X and Y be two Borel spaces. A *stochastic kernel on* X *given* Y is a function $\psi(B|y), B \in \mathcal{B}(\mathsf{X}), y \in \mathsf{Y}$ such that i) $\psi(\cdot|y)$ is a probability measure on $\mathcal{B}(\mathsf{X})$ for every fixed $y \in \mathsf{Y}$, and ii) $\psi(B|\cdot)$ is a measurable function on Y for every fixed $B \in \mathcal{B}(\mathsf{X})$.

Definition A Markov decision process (MDP, see e.g., White, 1993; Puterman, 1994),

$$(\mathsf{X}, \mathsf{A}, \{\mathsf{A}(x)|x \in \mathsf{X}\}, \mathcal{P}, r),$$

consists of the following components:

- *state space* X and *action space* A, which are Borel spaces;
- the feasible action set $\mathsf{A}(x)$, which is a nonempty Borel space of A, for a given state $x \in \mathsf{X}$;
- the *transition model* $\mathcal{P}(\mathsf{B}|x,a), \mathsf{B} \in \mathcal{B}(\mathsf{X}), (x,a) \in \mathsf{K}$: a *stochastic kernel* on X given K, where K denotes the set of feasible state-action pairs $\mathsf{K} := \{(x,a)|x \in \mathsf{X}, a \in \mathsf{A}(x)\}$, which is a Borel subset of $\mathsf{X} \times \mathsf{A}$; and
- the *reward function* r: $\mathsf{K} \to \mathbb{R}$, $\mathcal{B}(\mathsf{K})$-measurable.

Random variables are denoted by capital letters, e.g. X_t and A_t, whereas realizations of the random variables are denoted by normal letters, e.g. x_t and a_t.

Remark 5.1. The notations used in this chapter are mostly taken from Hernández-Lerma and Lasserre (1999, 1996), except that we consider rewards rather than costs. In the literature of optimal control, the similar framework with cost functions is correspondingly called *Markov control processes* (MCPs). In fact, given an MDP, setting $c := -r$, we immediately obtain the corresponding MCP.

Policy We consider here only *Markov policies*:

$$\boldsymbol{\pi} = [\pi_0, \pi_1, \pi_2, \ldots],$$

where each *single-step policy* $\pi_t(\cdot|x_t)$, which denotes the probability of choosing action a_t at x_t, $(x_t, a_t) \in \mathsf{K}$, is Markov (independent of the states and actions before t) and, therefore, a stochastic kernel on A given X. We use the boldface to represent a sequence of policies while using normal typeface for a single-step

policy. Let Δ denote the set of all stochastic kernels on A given X, μ, such that $\mu(\mathsf{A}(x)|x) = 1$ and Π_M denotes the set of all Markov policies. Thus $\Pi_M = \Delta^\infty$. A policy $f \in \Delta$ is *deterministic* if for each $x \in \mathsf{X}$, there exists some $a \in \mathsf{A}(x)$ such that $f(\{a\}|x) = 1$. Let $\Delta_D \subset \Delta$ denote the set of all deterministic single-step policies. A policy $\boldsymbol{\pi}$ is said to be *stationary*, if $\boldsymbol{\pi} = \pi^\infty$ for some $\pi \in \Delta$. For each $x \in \mathsf{X}$ and single-step policy $\pi \in \Delta$, define

$$r^\pi(x) := \int_{\mathsf{A}(x)} r(x,a)\pi(\mathrm{d}a|x) \tag{5.1}$$

$$P^\pi(\mathsf{B}|x) := \int_{\mathsf{A}(x)} \mathcal{P}(\mathsf{B}|x,a)\pi(\mathrm{d}a|x), \mathsf{B} \in \mathcal{B}(\mathsf{X}). \tag{5.2}$$

Objective There are usually three types of objectives used in the literature of MDPs: finite-stage, discounted and average rewards, depicted as

$$S_T := \sum_{t=0}^{T} r(X_t, A_t), \quad S_\gamma := \sum_{t=0}^{\infty} \gamma^t r(X_t, A_t), \text{ and } S_A := \liminf_{T\to\infty} \frac{1}{T} S_T \tag{5.3}$$

where $\gamma \in [0,1)$ denotes the discount factor. Suppose we start from one given state $X_0 = x$. The optimization problem is then to maximize the expected objective

$$\sup_{\boldsymbol{\pi}\in\Pi_M} \mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x] \tag{5.4}$$

by selecting a policy $\boldsymbol{\pi}$, where $\mathcal{S}$ is S_T, S_γ or S_A.

Technical considerations

In some applications of MDPs, the reward can be also "noisy", i.e., the reward can be decomposed into two parts: $R = r + n$, where r denotes the reward function as in the framework applied in this chapter, and n denotes some additive noise (real-valued random variable), whose distribution might be dependent on the state-action pair (s,a). This more general setting is especially popular in reinforcement learning (Sutton and Barto, 1998), which is a stochastic approximation (Kushner and Yin, 2003) approach for solving the optimization induced MDPs. This topic will be dealt with in details in Chapter 6.

Other two generalizations of the reward function are to

1. allow rewards to depend also on the successive state x' (Bertsekas and Tsitsiklis, 1996), i.e., $r(x,a,x')$, and/or
2. allow rewards to be time-dependent.

For the first generalization, in the standard MDP theory (see e.g., Puterman, 1994, Section 2.1.3), it is equivalent to using

$$r(x,a) := \int r(x,a,x')\mathcal{P}(dx'|x,a)$$

as in our framework. However, it is remarkable that in the risk-sensitive MDP framework to be developed below this equivalence does not hold. Nevertheless, for simplicity, we restrict ourselves in this thesis to reward functions depending merely on the current state-action pair (x, a), since this setup covers already numerous applications (see e.g., Puterman, 1994, Chapter 1 and White, 1993, Chapter 8). For the second generalization, we can use time-dependent reward functions for the finite-stage problem. However, for the discounted and average valuation, the time-dependent setting will make the problem computational infeasible. Hence, we assume in our framework that the reward function is dependent of time.

In the classical MDP theory, one can show (see e.g., Puterman, 1994, Section 4.1.3 for the finite case, Section 6.2 for the discounted valuation and Section 8.4.3 for the average valuation) that among the whole policy set containing those policies that depend on the whole history, there exists always a Markov policy that optimizes the objective (5.4). In the risk sensitive cases, this statement, however, no longer holds, due to the added nonlinearity. Hence, we restrict ourselves merely to the set of Markov policies.

Another objective that is not covered by our framework is the *total reward*, $S := \lim_{T\to\infty} S_T$. This objective is usually studied under the assumption that the underlying MDP contains at least one 0-reward absorbing state (see e.g., Puterman, 1994, Chapter 5, Altman, 1999, Hernández-Lerma and Lasserre, 1999, Chapter 9 and Bertsekas and Tsitsiklis, 1996). In fact, the discounted valuation can be reformulated as an transient MDP with one additional "dummy" state. For a nice treatment, see Altman, 1999, Chapter 10. In this thesis, however, we focus on recurrent MDPs (for definition, see e.g., Hernández-Lerma and Lasserre, 1999, Section 7.3), since the reinforcement learning (see Chapter 6), another important topic of this thesis, requires that every state-action pair must be visited infinitely often, which is impossible for a transient MDP.

Finally, note that in the definition of average reward in (5.3), "lim inf", instead "lim", is used, since the limit might not exist for some MDPs. For an example, see Puterman, 1994, Section 8.1.1. On the other hand, the usage of "lim inf" means that this objective function is already risk-averse, for we take a pessimistic or conservative estimation of average rewards, comparing with "lim sup".

5.3 Risk-sensitive MDPs

5.3.1 Setup

We notice that the finite-stage objective function (while other two objectives can be dealt with analogously) can be decomposed as follows,

$$\mathbb{E}^{\boldsymbol{\pi}}_{X_0}[S_T] = r^{\pi_0}(X_0) + \mathbb{E}^{\pi_0}_{X_0}\Bigg[r^{\pi_1}(X_1) + \mathbb{E}^{\pi_1}_{X_1}\Bigg[r^{\pi_2}(X_2) + \ldots + \mathbb{E}^{\pi_{T-1}}_{X_{T-1}}\left[r^{\pi_T}(X_T)\right]\ldots\Bigg]\Bigg] \tag{5.5}$$

where

$$\mathbb{E}_{X_t}^{\pi_t}[v(X_{t+1})] := \int v(X_{t+1}) P^{\pi_t}(dX_{t+1}|X_t)$$

denotes the *conditional expectation* of the function v of the successive state X_{t+1} given current state X_t.

Analogous to the valuation map defined in Section 3.1, we first define the valuation map in a general setting. Given a Borel space X with σ-algebra $\mathcal{B}(\mathsf{X})$. Denote by $\mathscr{P}(\mathsf{X})$ the space of all probability measures on X, and by $\mathscr{L}(\mathsf{X})$ a linear space of real-valued $\mathscr{B}(\mathsf{X})$-measurable functions containing all constant functions.

Definition **5.2.** *A mapping $\mathcal{U}(y,(v,\mu)) : \mathsf{Y} \times \mathscr{L}(\mathsf{X}) \times \mathscr{P}(\mathsf{X}) \to \mathbb{R}$ is said to be a* valuation map on X given Y*, if*

(i) for each $y \in \mathsf{Y}$, $\mathcal{U}(y,(\cdot,\cdot))$ is a valuation function; and

(ii) $\mathcal{U}(\cdot,(v,\mu))$ is $\mathcal{B}(\mathsf{Y})$-measurable for each $(v,\mu) \in \mathscr{L}(\mathsf{X}) \times \mathscr{P}(\mathsf{X})$.

Note that by definition, for each $(x,a) \in \mathsf{K}$, $\mathcal{P}(\cdot|x,a)$ is a probability measure, written as $\mathcal{P}_{x,a}$. Now we define the valuation map in the framework of MDPs.

Definition **5.3.** *A mapping $\mathcal{U}((x,a),v) : \mathsf{K} \times \mathscr{L}(\mathsf{X}) \to \mathbb{R}$ is said to be a* valuation map on an MDP $(\mathsf{X},\mathsf{A},\{\mathsf{A}(x)|x \in \mathsf{X}\},\mathcal{P},r)$*, if there exists a valuation map $\tilde{\mathcal{U}}$ on X given K such that*

$$\mathcal{U}((x,a),v) = \tilde{\mathcal{U}}((x,a),v,\mathcal{P}_{x,a}).$$

Furthermore, $\mathcal{U}((x,a),v)$ is also written as $\mathcal{U}(v|x,a)$ or $\mathcal{U}_{x,a}(v)$ in different contexts and define

$$\mathcal{U}_x^\pi(v) = \mathcal{U}^\pi(v|x) := \int_{\mathsf{A}(x)} \pi(da|x)\mathcal{U}(v|x,a).$$

Remark 5.4. 1) We will explain the choice of $\mathscr{L}(\mathsf{X})$ in the next section.

2) In Definition 5.3, $\mathcal{U}^\pi$ is in fact assumed to be linear to the policy π, which simplifies the optimization problem and is one of the conditions that guarantee the existence of one optimal deterministic policy, "optimal selector" (see the next section).

We replace the conditional expectation in (5.5) with the valuation map defined above and therefore, obtain the following *T-stage risk-sensitive* objective

$$J_T^{\boldsymbol{\pi}} = r^{\pi_0}(X_0) + \mathcal{U}_{X_0}^{\pi_0}\left[r^{\pi_1}(X_1) + \mathcal{U}_{X_1}^{\pi_1}\left[r^{\pi_2}(X_2) + \ldots + \mathcal{U}_{X_{T-1}}^{\pi_{T-1}}[r^{\pi_T}(X_T)]\ldots\right]\right].$$

Its optimization problem can be solved by *dynamic programming* (see e.g. Ruszczyński, 2010), whereas the other two objectives will be defined analogously and discussed in the next section.

In the mathematical finance literature, there exist various ways to extend coherent/convex risk measures to a temporal structure (see Föllmer and Penner,

2006; Cheridito and Kupper, 2011; Ruszczyński, 2010 and references therein). The definition is usually selected based on applications. To compare their subtle differences are out of the scope of this thesis. The evaluation maps defined here are similar to the *risk measure generators* in Cheridito and Kupper (2011) and are implicitly Markovian and time-homogeneous (see also Ruszczyński, 2010), since $\mathcal{U}$ defined above depends merely on the most recent state and action but not the whole history.

The valuation maps used in our risk-sensitive MDP-framework are assumed to be Markovian (cf. the motivation explained in Section 3.1), since in the MDP-framework the underlying stochastic process is Markovian, while the assumption of time homogeneity is due to the fact that since we consider mainly the infinite-horizon criteria (see Section 5.3.2), as in the literature of MDPs, stationary optimal policies are expected. Hence, to comply with the MDP-framework, it is sufficient to construct an operator which replaces the conditional expectation determined by the transition model $\mathcal{P}$ and policy π.

Risk preference

Definition **5.5.** *A valuation map $\mathcal{U}$ is said to be* convex *(respectively* concave, homogeneous*) if $\mathcal{U}(x,a)$ is convex (respectively* concave, homogeneous*), for all* $(x,a) \in \mathsf{K}$.

Applying Proposition 3.9, it is easy to see that concave valuation maps induce risk-averse behavior.

Examples

Note that all examples valuation functions we have presented in Section 2.4.2–2.4.7 can be easily extended to valuation maps correspondingly by replacing μ with some transition kernel $\mathcal{P}$. For instance, the entropic measure $\rho^{\lambda}(v,\mu) = \frac{1}{\lambda}\log\left\{\int_{\Omega} e^{\lambda v} d\mu\right\}$ defined in Section 2.4.2 can be extended to a valuation map, which is said to be an *entropic map*, as follows,

$$\mathcal{U}(v|x,a) := \frac{1}{\lambda}\log\left\{\int_{\mathsf{X}} e^{\lambda v(y)}\mathcal{P}(dy|x,a)\right\}, \lambda \neq 0. \tag{5.6}$$

Other types of valuation maps can be obtained analogously.

5.3.2 Objectives

Let $\mathcal{U}$ be a valuation map and $\gamma \in [0,1]$ be a discount factor. Given $\boldsymbol{\pi} \in \Pi_M$ and $x \in \mathsf{X}$, define

$$J_{\gamma,T}(x,\boldsymbol{\pi}) := r^{\pi_0}(x) + \gamma\mathcal{U}_x^{\pi_0}\left(r^{\pi_1} + \gamma\mathcal{U}^{\pi_1}\left(r^{\pi_2} + \ldots + \gamma\mathcal{U}^{\pi_{T-1}}(r^{\pi_T})\ldots\right)\right). \tag{5.7}$$

We consider the following risk-sensitive objectives:

1) the *T-stage valuation*:

$$J_T(x, \boldsymbol{\pi}) := J_{1,T},$$

2) the *discounted valuation*:

$$J_\gamma(x, \boldsymbol{\pi}) := \lim_{T\to\infty} J_{\gamma,T}(x, \boldsymbol{\pi}), \text{ and} \tag{5.8}$$

3) the *average valuation*:

$$J(x, \boldsymbol{\pi}) := \liminf_{T\to\infty} \frac{1}{T} J_{1,T}(x, \boldsymbol{\pi}), \boldsymbol{\pi} \in \Pi_M, x \in \mathbf{X}. \tag{5.9}$$

Remark 5.6. For the proof of the existence of the limit in (5.8) (under some technical assumptions), see Lemma 5.16 in the next section. For an example where the limit does not exist in (5.9), see (Puterman, 1994, Example 8.1.1). We, therefore, use "lim inf" instead of "lim" in the definition of average valuation.

The optimal control problems for above three risk-sensitive objectives are to maximize the subjective valuation among all Markov policies

$$\begin{aligned} J_T^*(x) &:= \sup_{\boldsymbol{\pi}\in\Pi_M} J_T(x, \boldsymbol{\pi}), \\ J_\gamma^*(x) &:= \sup_{\boldsymbol{\pi}\in\Pi_M} J_\gamma(x, \boldsymbol{\pi}), \text{ and} \\ J^*(x) &:= \sup_{\boldsymbol{\pi}\in\Pi_M} J(x, \boldsymbol{\pi}). \end{aligned}$$

Remarks on time consistency

Let $\gamma = 1$ and consider the T-stage valuation J_T. Following the same line as in Section 3.3 (see also Ruszczyński, 2010), it is easy to verify that given a time-consistent dynamic valuation function, $\{\rho_{t,T}, t = 0, 1, \ldots, T\}$, when applying to the sum S_T defined in (5.3), one can always obtain a backward induction procedure as in (5.7).

Remarks on the definition of discounted valuation

In economics, the time-discount is added to reflect the "time-value" of outcomes: the outcome to be gained in the future is less valuable than the same amount of outcome obtained now. It has similar effects when cost is concerned. Due to its good mathematical properties, exponential discounting scheme, where the cost c_t is multiplied with the time-discount γ^t, is widely applied in economics, finance as well as in MDPs.

A natural extension of classical discounted MDPs, therefore, is

$$D_\gamma(\boldsymbol{\pi}) = r^{\pi_0} + \mathcal{U}^{\pi_0}\left(\gamma r^{\pi_1} + \mathcal{U}^{\pi_1}\left(\gamma^2 r^{\pi_2} + \ldots + \mathcal{U}^{\pi_{T-1}}(\gamma^T r^{\pi_T} + \ldots)\ldots\right)\right).$$

However, since the valuation map $\mathcal{U}$ is not necessarily homogeneous, a stationary policy that optimizes D_γ need not exist. Indeed, it was proved in Chung and Sobel (1987) that for the entropic map, which is not homogeneous, the optimal policy might not be stationary if $D_\gamma(\boldsymbol{\pi})$ is optimized with respect to $\boldsymbol{\pi}$, though D_γ is well-defined for all $\gamma \in [0,1)$. In our definition, discount factor γ is multiplied with $\mathcal{R}$, which has the same "time-discount" effect, where the subjective valuation rather than the immediate reward is discounted. Moreover, it is easy to see that, if $\mathcal{U}$ is homogeneous, D_γ is equivalent to J_γ, the discounted valuation under our definition. Therefore, D_γ defined for any homogeneous valuation map is merely a special case of our definition. Specifically, the classical discounted MDP is indeed a special case of our defined discounted valuation, since it is homogeneous. D_γ was used in Ruszczyński (2010) and the corresponding optimization problem was solved by a value iteration algorithm, since merely the coherent (i.e. concave and homogeneous) valuation maps were considered. Besides, in the proof of the value iteration algorithm, the representation theorem was used, which is valid merely for coherent (i.e. concave and homogeneous) valuation maps. On the contrary, we will see later that the objective J_γ allows a value iteration algorithm for general valuation maps. Therefore, we apply J_γ rather than D_γ.

5.4 Optimization

We derive in this section *value iteration* algorithms to solve the optimization problems proposed in the last section. Among them, the optimal T-stage valuation J_T^* can be easily obtained by *dynamic programming* and is therefore omitted here. The value iteration algorithms for the discount valuation and average valuation are presented in Section 5.4.2 and 5.4.3 respectively. We start first with some preparatory assumptions.

5.4.1 Preparatory assumptions

Optimal selector

Define the following operators

$$\mathcal{F}_\gamma^\pi(v) := r^\pi + \gamma \mathcal{U}^\pi(v),\ \mathcal{F}_\gamma(v) := \sup_{\boldsymbol{\pi}\in\Pi_M} \mathcal{F}_\gamma^\pi(v),$$

where $v \in \mathscr{B}_w$ and $\gamma \in [0,1]$. If $\gamma = 1$, we simply write them as $\mathcal{F}^\pi$ and $\mathcal{F}$ respectively. The operators $(\mathcal{F}_\gamma^\pi)^n$, $n \in \mathbb{N}$, are defined iteratively as $(\mathcal{F}_\gamma^\pi)^0(v) := v$, and $(\mathcal{F}_\gamma^\pi)^n(v) := \mathcal{F}_\gamma^\pi((\mathcal{F}_\gamma^\pi)^{n-1}(v))$, $n = 1, 2, \ldots$, while $\mathcal{F}_\gamma^t$ is defined analogously.

The following assumption is made to ensure the existence of the "selector" in the optimization problem.

Assumption 5.7. For each $x \in \mathsf{X}$,

(i) the reward function $r(x,a)$ is upper semi-continuous on $\mathsf{A}(x)$,

(ii) the action space $\mathsf{A}(x)$ is compact, and

(iii) the function $u'(x,a) := \mathcal{U}_{x,a}(u)$ is continuous on $a \in \mathsf{A}(x)$ for any $u \in \mathscr{B}_w$.

Remark 5.8. This assumption dates back to Schäl (1974) and was later applied to MCPs by Bertsekas and Shreve (1978). Recently, a set of weaker assumptions has been developed by Feinberg et al. (2013).

Proposition **5.9.** *Suppose $\mathcal{U}$ is a valuation map satisfying Assumption 5.7. Then, for all $v \in \mathscr{B}_w$ and $x \in \mathsf{X}$, there exists a deterministic policy $f \in \Delta_D$, such that for any $\gamma \in [0,1]$,*

$$r^f(x) + \gamma \mathcal{U}^f(v|x) = \mathcal{F}_\gamma(v|x) = \sup_{\pi \in \Delta} \{r^\pi(x) + \gamma \mathcal{U}^\pi(v|x)\}. \tag{5.10}$$

Proof. Apparently, for each $x \in \mathsf{X}$,

$$\mathcal{F}_\gamma(v|x) = \sup_{a \in \mathsf{A}(x)} \{r(x,a) + \gamma \mathcal{U}(v|x,a)\}.$$

By Assumption 5.7(i) and (iii), the function

$$u(x,a) := r(x,a) + \gamma \mathcal{U}(v|x,a), \gamma \in [0,1],$$

is upper semi-continuous in $a \in \mathsf{A}(x)$ for each $x \in \mathsf{X}$. Hence, $-u(x,a)$ is lower semi-continuous. Hence, by Assumption 5.7(ii) and Proposition 7.33 of Bertsekas and Shreve (1978) or Lemma 8.3.8(a) of Hernández-Lerma and Lasserre (1999), an optimal selector for $\inf_{a \in \mathsf{A}(x)} -u(x,a)$ exists, which is equivalent to the existence of an optimal selector for $\sup_{a \in \mathsf{A}(x)} u(x,a)$. □

Upper envelope

We introduce below the concept of upper envelope (see also Section 4.3.2) to control the growth of iterations, $\{\mathcal{F}_\gamma^n\}$.

Definition **5.10.** *A convex and homogeneous valuation map $\bar{\mathcal{U}}^{(w,C)}$ is said to be an* upper envelope *of a valuation map $\mathcal{U}$ given a constant $C > 0$, if for all $v, u \in \mathscr{B}_w^{(C)} := \{v \in \mathscr{B}_w | \|v\|_{s,w} \le C\}$,*

$$\mathcal{U}_{x,a}(v) - \mathcal{U}_{x,a}(u) \le \bar{\mathcal{U}}_{x,a}^{(w,C)}(v-u), \forall (x,a) \in \mathsf{X}.$$

Remark 5.11. Apparently, if $\mathcal{U}$ is convex and homogeneous, then $\mathcal{U}$ is an upper envelope of itself for any $C > 0$.

To ensure the existence of the upper bound C, we assume,

Assumption 5.12. There exist a $\mathcal{B}(\mathsf{X})$-measurable function $w_0 : \mathsf{X} \to [0,\infty)$, constants $\alpha_0 \in (0,1)$ and $\tilde{K}_0 > K_0 > 0$ such that

(i) for each $(x,a) \in \mathsf{K}$,

$$[r(x,a) + \mathcal{U}_{x,a}(w_0)] \vee [-r(x,a) - \mathcal{U}_{x,a}(-w_0)] \leq \alpha_0 w_0(x) + K_0;$$

(ii) for all $x, x' \in \mathsf{B}_0 := \{x \in \mathsf{X} | w_0(x) \leq R_0 := \frac{2K_0}{1-\alpha_0}\}$, $a \in \mathsf{A}(x)$, $a' \in \mathsf{A}(x')$, the inequality

$$\mathcal{U}_{x,a}(v) - \mathcal{U}_{x',a'}(v) \leq 2(\tilde{K}_0 - K_0) + \mathcal{U}_{x,a}(w_0) - \mathcal{U}_{x',a'}(-w_0)$$

holds for all v satisfying $|v| \leq w_0 + \tilde{K}_0$.

and obtain the following theorem, which gives us a *bounded forward invariant subset.*

Theorem **5.13.** *Suppose Assumption 5.12 holds. Let* $w := 1 + \tilde{K}_0^{-1} w_0$. *Then for all* $\pi \in \Delta$ *and* $\gamma \in [0,1]$,

$$\|\mathcal{F}_\gamma^\pi(v)\|_{s,w} \leq \tilde{K}_0, \text{ whenever } \|v\|_{s,w} \leq \tilde{K}_0.$$

Proof. The proof is a simple repeat of the proof of Theorem 4.14 and is therefore ignored. □

5.4.2 Discounted valuation

Under Assumption 5.12, we can restrict to the bounded forward invariant subset $\mathcal{B}_w^{(\tilde{K}_0)}$. For the discounted valuation, we need further assumptions

Assumption 5.14. (a) Assumption 5.12 holds with some weight function w_0 and a constant $\tilde{K}$.

(b) Let $w := 1 + \tilde{K}_0^{-1} w_0$. There exists a constant $\bar{w} \in [1, 1/\gamma)$ such that

$$\sup_{a \in \mathsf{A}(x)} \bar{\mathcal{U}}_{x,a}^{(w,\tilde{K}_0)}(w) \leq \bar{w} w(x). \tag{5.11}$$

Proposition **5.15.** *Suppose Assumption 5.14 holds. Then for each* $\pi \in \Delta$ *and* $v, u \in \mathscr{B}_w^{(\tilde{K}_0)}$,

$$\|\mathcal{F}_\gamma^\pi(v) - \mathcal{F}_\gamma^\pi(u)\|_w \leq \bar{w}\gamma \|v - u\|_w.$$

Proof. For each $v, u \in \mathscr{B}_w^{(\tilde{K}_0)}$, we have $\forall (x,a) \in \mathsf{K}$

$$\begin{aligned}
\mathcal{U}_{x,a}(v) - \mathcal{U}_{x,a}(u) \leq & \bar{\mathcal{U}}_{x,a}^{(w,\tilde{K}_0)}(v - u) \\
\leq & \bar{\mathcal{U}}_{x,a}^{(w,\tilde{K}_0)}(\|v - u\|_w w) \\
= & \|v - u\|_w \bar{\mathcal{U}}_{x,a}^{(w,\tilde{K}_0)}(w) \\
\leq & \|v - u\|_w \bar{w} w(x)
\end{aligned}$$

which yields for each $\pi \in \Delta$ and $x \in \mathsf{X}$,

$$\begin{aligned}\mathcal{F}_\gamma^\pi(v|x) - \mathcal{F}_\gamma^\pi(u|x) =& \gamma\left(\mathcal{U}_x^\pi(v) - \mathcal{U}_x^\pi(u)\right)\\ =& \gamma \int_{\mathsf{A}(x)} \pi(da|x)\left(\mathcal{U}_{x,a}(v) - \mathcal{U}_{x,a}(u)\right)\\ \le& \gamma \|v-u\|_w \bar{w} w(x).\end{aligned}$$

Switching v and u, we obtain for each $x \in \mathsf{X}$,

$$|\mathcal{F}_\gamma^\pi(v|x) - \mathcal{F}_\gamma^\pi(u|x)| \le \gamma \|v-u\|_w \bar{w} w(x), \tag{5.12}$$

which implies immediately the required inequality. □

We then show that the limit in the definition of discounted valuation is well defined.

Lemma **5.16.** *Suppose Assumption 5.14 holds. Then for each $\boldsymbol{\pi} \in \Pi_M$,*

(i) $J_\gamma(\boldsymbol{\pi}) = \lim_{T\to\infty} J_{\gamma,T}(\boldsymbol{\pi})$ defined in (5.8) exists in $\mathscr{B}_w^{(\tilde{K}_0)}$, and

(ii) for each $\pi \in \Pi_M$ and $v \in \mathscr{B}_w^{(\tilde{K}_0)}$,

$$J_\gamma(\boldsymbol{\pi}) = \lim_{T\to\infty} \mathcal{F}_\gamma^{\pi_0}(\mathcal{F}_\gamma^{\pi_1} \ldots \mathcal{F}_\gamma^{\pi_T}(v)\ldots).$$

Proof. By iterating Theorem 5.13, we have for each $T > 0$, $J_{\gamma,T}(\boldsymbol{\pi}) \in \mathscr{B}_w^{(\tilde{K}_0)}$. By Assumption 5.14 (a), we have

$$|r(x,a)| \le w_0(x) + \tilde{K}_0 = \tilde{K}_0 w(x), \forall (x,a) \in \mathsf{K},$$

which implies that $\|r^\pi\|_w \le \tilde{K}_0$ holds for all $\pi \in \Delta$. Hence, by iterating Proposition 5.15, we have

$$\begin{aligned}&\|J_{\gamma,T+1}(\boldsymbol{\pi}) - J_{\gamma,T}(\boldsymbol{\pi})\|_w\\ =&\|\mathcal{F}_\gamma^{\pi_0}(\mathcal{F}_\gamma^{\pi_1}\ldots\mathcal{F}_\gamma^{\pi_T}(r^{\pi_{T+1}})\ldots) - \mathcal{F}_\gamma^{\pi_0}(\mathcal{F}_\gamma^{\pi_1}\ldots\mathcal{F}_\gamma^{\pi_{T-1}}(\mathcal{F}_\gamma^{\pi_T}(0))\ldots)\|_w \qquad (5.13)\\ \le&(\overline{w}\gamma)^T \|r^{\pi_{T+1}}\|_w. \qquad (5.14)\end{aligned}$$

Since $\overline{w}\gamma \in [0,1)$, we have

$$\lim_{T\to\infty} \|J_{\gamma,T+1}(\boldsymbol{\pi}) - J_{\gamma,T}(\boldsymbol{\pi})\|_w = 0.$$

This shows the exist exists in $\mathscr{B}_w^{(\tilde{K}_0)}$.

Statement (ii) is straightforward by replacing $c^{\pi_{T+1}}$ with $v \in \mathscr{B}_w^{(\tilde{K}_0)}$ in (5.13). □

Now, we show that $\mathcal{F}_\gamma$ is a contraction map.

LEMMA **5.17.** *Suppose Assumption 5.7 and 5.14 hold. Then for each* $v, u \in \mathscr{B}_w^{(\tilde{K}_0)}$,

$$\|\mathcal{F}_\gamma(v) - \mathcal{F}_\gamma(u)\|_w \leq \bar{w}\gamma\|v - u\|_w, \bar{w}\gamma \in [0,1).$$

Proof. By Proposition 5.9, the optimal selector f^* always exists for all $v \in \mathscr{B}_w$. Let f_v be the optimal selector for v and f_u be the optimal selector for u. Thus

$$\mathcal{F}_\gamma(v) - \mathcal{F}_\gamma(u) \leq \mathcal{F}_\gamma^{f_v}(v) - \mathcal{F}_\gamma^{f_v}(u) \leq \overline{w}\gamma\|v - u\|_w w,$$

where the last inequality is due to (5.15). By switching v and u, we obtain On the other hand,

$$|\mathcal{F}_\gamma(u) - \mathcal{F}_\gamma(v)| \leq \overline{w}\gamma\|v - u\|_w w.$$

Hence, $\|\mathcal{F}_\gamma(v) - \mathcal{F}_\gamma(u)\|_w \leq \overline{w}\gamma\|v - u\|_w$. □

Hence, by Banach's fixed point theorem, starting from some $v \in \mathscr{B}_w$ satisfying $\|v\|_{s,w} \leq \tilde{K}_0$, $\mathcal{F}_\gamma^n(v)$ converges to a unique fixed point v^* in $\mathscr{B}_w$ satisfying the *Bellman equation*:

$$v^*(x) := \mathcal{F}_\gamma(v^*|x) = \sup_{a \in \mathsf{A}(x)} \{r(x,a) + \gamma \mathcal{U}(v^*|x,a)\}. \tag{5.15}$$

Let f be an optimal selector in the right hand side of the above equation. The following theorem indicates the link between the Bellman equation and the optimization problem of discounted valuation.

THEOREM **5.18.** *Suppose Assumption 5.7 and 5.14 hold. Then* $v^*(x) = J_\gamma^*(x) = J_\gamma(x, f^\infty)$ *for all* $x \in \mathsf{X}$.

Proof. First, we show $v \leq \mathcal{F}_\gamma v$ implies $v \leq J_\gamma^*$. By Proposition 5.9, we assume that the optimal selector for v is f. Hence, we obtain

$$v \leq \mathcal{F}_\gamma^f(v) \leq \mathcal{F}_\gamma^f\mathcal{F}_\gamma^f(v) \leq \mathcal{F}_\gamma^f\mathcal{F}_\gamma^f\mathcal{F}_\gamma^f(v) \leq (\mathcal{F}_\gamma^f)^\infty(v) = J_\gamma(f^\infty) \leq J_\gamma^*$$

where the equality is due to Lemma 5.16(ii).

Second, we show that $v \geq \mathcal{F}_\gamma v$ implies $v \geq J_\gamma^*$. Indeed, let $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots] \in \Pi_M$ be an arbitrary Markov random policy. Then, for all $\pi \in \Delta$, $v \geq \mathcal{F}_\gamma v \geq \mathcal{F}_\gamma^\pi v$. Hence,

$$v \geq \mathcal{F}_\gamma^{\pi_0}(v) \geq \mathcal{F}_\gamma^{\pi_0}\mathcal{F}_\gamma^{\pi_1}(v) \geq \mathcal{F}_\gamma^{\pi_0}\mathcal{F}_\gamma^{\pi_1}\mathcal{F}_\gamma^{\pi_2}(v) \geq \mathcal{F}_\gamma^{\pi_0} \ldots \mathcal{F}_\gamma^{\pi_T}(v) \to J_\gamma(\boldsymbol{\pi}).$$

The limit is due to Lemma 5.16(ii). Since $\boldsymbol{\pi}$ can be arbitrarily chosen, it follows $v \leq \inf_{\boldsymbol{\pi} \in \Pi_M} J_\gamma(\boldsymbol{\pi}) = J_\gamma^*$.

Since $v^* = \mathcal{F}_\gamma v^*$, combining above two steps yields $v^* = J_\gamma^*$. □

By the above theorem, we immediately obtain the following corollary for existence of a stationary deterministic policy.

COROLLARY **5.19.** *Under Assumption 5.14 and 5.7, there exists a stationary deterministic policy* $f^* \in \Delta_D$ *such that* $J_\gamma^* = J_\gamma((f^*)^\infty)$.

Remarks on convex and homogeneous valuation maps

Note that if the valuation map $\mathcal{U}$ is convex and homogeneous, then its upper envelope becomes itself. Assumption 5.14 can therefore be relaxed to

Assumption 5.20. $\mathcal{U}$ is a convex and homogeneous valuation map. Assume

(a) $\bar{r}(x) := \sup_{a\in \mathsf{A}(x)} |r(x,a)| \in \mathscr{B}_w$, and

(b) there exists a nonnegative $\overline{w}$, with $1 \le \overline{w} < 1/\gamma$ such that $\forall x \in \mathsf{X}$,

$$\sup_{a\in \mathsf{A}(x)} \mathcal{U}(w|x,a) \le \overline{w} w(x).$$

and replacing Assumption 5.14 with the above assumption, all the results stated in this subsection hold, where the bounded subspace $\mathscr{B}_w^{\tilde{K}_0}$ can be relaxed to the whole space $\mathscr{B}_w$.

Remarks on finite state-action spaces

In some real-world applications, the state and action spaces are assumed to be finite. Then, Assumption 5.7 automatically holds.

Due to the nonexpansiveness of $\mathcal{U}$ under the sup-norm $\|\cdot\|_\infty$ (see Proposition 4.37(ii)), the statement of Proposition 5.15 holds without Assumption 5.14. It is therefore easy to verify that $\mathcal{F}_\gamma$ is a contraction map under the sup-norm $\|\cdot\|_\infty$ (cf. Lemma 5.17) and Theorem 5.18 holds without Assumptions 5.7 and 5.14.

Value iteration

Finally, according to Lemma 5.17 and Theorem 5.18 we state the following algorithm:

Algorithm 5.1 Value iteration for discounted problems

select one $v_0 \in \mathscr{B}_w^{(\tilde{K}_0)}$, $t = 0$;
repeat
  $v_{t+1} = \mathcal{F}_\gamma(v_t)$ with selector

$$f_{t+1}(x) := \operatorname*{argmax}_{a\in \mathsf{A}(x)} \{r(x,a) + \gamma \mathcal{U}(v_t|x,a)\};$$

  $t = t + 1$;
until $\|v_{t+1} - v_t\|_w < \epsilon$

Theorem 5.18 guarantees $v_n \to J_\gamma^*$ and $f_t \to f^*$, the optimal policy.

5.4.3 Average valuation

We now deal with the average valuation based on the following assumption.

Assumption 5.21 (cf. Assumption 4.17). Let $w_0 : \mathsf{X} \to [0,\infty)$ and $w : \mathsf{X} \to [1,\infty)$ be two real-valued nonnegative $\mathcal{B}(\mathsf{X})$-measurable functions satisfying

(i) $\mathscr{B}_{1+w_0} = \mathscr{B}_w$;

(ii) there exist constants $\gamma \in (0,1)$, $K > 0$ and an upper envelope $\bar{\mathcal{U}}^{(w,\tilde{K}_0)}$ such that
$$\bar{\mathcal{U}}^{(w,\tilde{K}_0)}_{x,a}(w_0) \le \gamma w_0(x) + K, \forall (x,a) \in \mathsf{K};$$

(iii) and furthermore, there exists a probability measure μ such that for all $x, x' \in \mathsf{B} := \{x \in \mathsf{X} | w_0(x) \le R, R > \frac{2K}{1-\gamma}\}$, $a \in \mathsf{A}(x)$, $a' \in \mathsf{A}(x')$, and $v \ge u \in \mathscr{B}_{1+w_0}$,
$$\bar{\mathcal{U}}^{(w,\tilde{K}_0)}_{x,a}(v) - \bar{\mathcal{U}}^{(w,\tilde{K}_0)}_{x,a}(u) \ge \gamma \int (v(x) - u(x))\, \mu(dx).$$

The following lemma shows that $\mathcal{F}^n \to 0$ under the $(1 + \beta w_0)$-seminorm, as $n \to \infty$.

Lemma **5.22.** *Suppose Assumption 5.7 and 5.12 hold. Assume further that Assumption 5.21 holds with the same w_0 as in Assumption 5.12. Then there exists $\bar{\gamma} \in (0,1)$ and $\beta > 0$ such that*
$$\|\mathcal{F}(v) - \mathcal{F}(u)\|_{s,1+\beta w_0} \le \bar{\gamma} \|v - u\|_{s,1+\beta w_0}, \forall v, u \in \mathscr{B}^{(\tilde{K}_0)}_w.$$

Proof. By Proposition 5.9, there exist deterministic policies $f_v, f_u \in \Delta_D$ such that $\mathcal{F}(v) = \mathcal{F}^{f_v}(v)$ and $\mathcal{F}(u) = \mathcal{F}^{f_u}(u)$. Thus
$$\begin{aligned} &\mathcal{F}(v) - \mathcal{F}(u) \le \mathcal{F}^{f_v}(v) - \mathcal{F}^{f_v}(u) = \mathcal{R}^{f_v}(v) - \mathcal{R}^{f_v}(u) \\ \text{and} \quad &\mathcal{F}(u) - \mathcal{F}(v) \le \mathcal{F}^{f_u}(u) - \mathcal{F}^{f_u}(v) = \mathcal{R}^{f_u}(u) - \mathcal{R}^{f_u}(v) \end{aligned}$$
yield for all $x, y \in \mathsf{X}$,
$$\mathcal{F}_x(v) - \mathcal{F}_x(u) + \mathcal{F}_y(u) - \mathcal{F}_y(v) \le \mathcal{R}^{f_u}_x(v) - \mathcal{R}^{f_u}_x(u) + \mathcal{R}^{f_v}_y(u) - \mathcal{R}^{f_v}_y(v).$$
By Assumption 5.21 and repeating the proof in Theorem 4.18, we obtain the required inequality. □

Finally, we show the existence of a solution to the *Poisson equation* and its link to the the optimization problem of average valuation.

Theorem **5.23.** *Under the same assumption as in Lemma 5.22. Then the following* Poisson equation
$$\rho^* + h(x) = \mathcal{F}_x(h) = \sup_{a \in \mathsf{A}(x)} \{r(x,a) + \mathcal{U}(h|x,a)\} \tag{5.16}$$
has a solution $(\rho^*, h) \in \mathbb{R} \times \mathscr{B}_w$, where ρ^* is unique. Furthermore, $\rho^* = J^*(x) = J(x, f^\infty)$ for all $x \in \mathsf{X}$, where f is an optimal selector in the right hand side of (5.16).

Proof. The existence of a unique solution to the Poisson equation is simply due to Lemma 5.22 and Theorem 4.20(i). By Proposition 5.9, the optimal selector f exists. Thus we assume $\mathcal{F}(h) = \mathcal{F}^f(h)$. Iterating (5.16), we obtain

$$(\mathcal{F}^f)^t(h) = (\mathcal{F}^f)^{t-1}(\rho + h) = t\rho + h \quad \Rightarrow \quad \lim_{t\to\infty} \frac{1}{t}\|(\mathcal{F}^f)^t(h) - \rho\|_w = 0.$$

On the other hand, for any $v \in \mathscr{B}_w^{(\tilde{K}_0)}$, by Lemma 4.19,

$$\frac{1}{t}\|(\mathcal{F}^f)^t(v) - (\mathcal{F}^f)^t(h)\|_w \to 0$$

implies that

$$J(x, f^\infty) = \lim_{t\to\infty} \frac{1}{t}(\mathcal{F}^f)_x^t(0) = \rho, \forall x \in \mathsf{X}.$$

Next we prove that $\rho \geq J(x, \boldsymbol{\pi})$ for all $\boldsymbol{\pi} \in \Pi_M$ and $x \in \mathsf{X}$. In fact, let $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots]$ be an arbitrary Markov policy. Then by Lemma 4.19, for all $v \in \mathscr{B}_w^{(\tilde{K}_0)}$,

$$\lim_{t\to\infty} \frac{1}{t}\|(\mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1} \ldots \mathcal{F}^{\pi_t}(v)) - \mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1} \ldots \mathcal{F}^{\pi_t}(0))\|_w = 0. \tag{5.17}$$

By definition $h \geq \mathcal{F}^\pi(h) - \rho, \forall \pi \in \Delta$. Iterating this inequality yields

$$h \geq \mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1}(\ldots \mathcal{F}^{\pi_{t-1}}(h))) - t\rho \tag{5.18}$$

$$\Rightarrow \quad \liminf_{t\to\infty} \frac{1}{t}\mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1}(\ldots \mathcal{F}^{\pi_{t-1}}(h))) \leq \rho. \tag{5.19}$$

Note that by definition

$$J(\boldsymbol{\pi}) = \liminf_{t\to\infty} \frac{1}{t}\mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1}(\ldots \mathcal{F}^{\pi_{t-1}}(0)))$$

and by setting $v = h$ in (5.17), we obtain

$$J(\boldsymbol{\pi}) = \liminf_{t\to\infty} \frac{1}{t}\mathcal{F}^{\pi_0}(\mathcal{F}^{\pi_1}(\ldots \mathcal{F}^{\pi_{t-1}}(h))).$$

Hence, (5.19) implies $\rho \geq J(\boldsymbol{\pi})$. It follows that

$$\rho \geq \sup_{\boldsymbol{\pi}\in\Pi_M} J(\boldsymbol{\pi}) = J^*.$$

Since f^∞ is a valid Markov policy in Π_M, $\rho = J^* = J(f^\infty)$. □

Remarks on convex and homogeneous valuation maps

If $\mathcal{U}$ is convex and homogeneous, then $\mathcal{U}$ itself is an upper envelope $\bar{\mathcal{U}}^{(w,C)}$ for any $C > 0$. In this case, Assumption 5.12 is no longer needed in Lemma 5.22 and Theorem 5.23 to determine *a priori* the size of the bounded forward invariant subset, C. For instance, applying the classical MCP, $\bar{\mathcal{U}}_{x,a}^{(w,\tilde{K}_0)}(v) = \mathbb{E}_{x,a}^{\mathcal{P}}(v)$, and obviously Assumption 5.21 (iii) is equivalent to the classical Doeblin's condition. Hence, Assumption 5.21 becomes the classical condition that has been widely used in the MDP/MCP literature (see Hernández-Lerma et al., 1991; Hernández-Lerma and Lasserre, 1999; Vega-Amaya, 2003 and references therein) for studying the average cost.

Remarks on entropic maps

Similar to the analysis we made in Section 4.4, by Theorem 4.36, we obtain below sufficient conditions for Assumption 5.12 and 5.21.

Proposition **5.24.** *Let $\mathcal{U}$ be the entropic map defined in* (5.6) *with $\lambda = 1$. Suppose the following conditions hold: (i) there exist a function $w_1 : \mathsf{X} \in [1, \infty)$, constants $\gamma_1 \in (0,1)$ and $K_1 > 0$ such that*

$$\mathcal{U}_{x,a}(w_1) \leq \gamma_1 w_1(x) + K_1, \forall (x,a) \in \mathsf{K},$$

(ii) for all $p \in (0,1)$ and all level-sets $\mathsf{C} := \mathscr{B}_{w_1^p}(R)$, $R > 0$, there exist a measure μ_{C} and constants $\lambda_{\mathsf{C}}^+ > \lambda_{\mathsf{C}}^- > 0$ such that $\mu_{\mathsf{C}}(\mathsf{C}) > 0$ and $\forall x \in \mathsf{C}, a \in \mathsf{A}(x)$ and $\mathsf{B} \in \mathcal{B}(\mathsf{X})$,

$$\lambda_{\mathsf{C}}^- \mu_{\mathsf{C}}(\mathsf{B} \cap \mathsf{C}) \leq Q_{x,a}(\mathsf{B} \cap \mathsf{C}) \leq \lambda_{\mathsf{C}}^+ \mu_{\mathsf{C}}(\mathsf{B} \cap \mathsf{C}),$$

and (iii) the cost function c satisfies

$$\bar{c}(x) := \sup_{a \in \mathsf{A}(x)} |c(x,a)| \in \mathscr{B}_{w_1^q} \text{ for some } q \in (0,1).$$

Then Assumption 5.12 holds with $w_0 = w_1^p$ for any $p \in (q,1)$ and some constant $\tilde{K}_0$, and Assumption 5.21 holds with $w = 1 + \tilde{K}_0^{-1} w_0$.

We compare below our results with the most related literature Di Masi and Stettner (2008).

(a) The assumption (A4) in Section 4 of Di Masi and Stettner (2008) requires a positive continuous density, i.e., there exists a positive function q satisfying $Q(dy|x,a) = q(x,a,y)\mu(dy)$ for some reference probability measure μ, which implies the local Doeblin's condition in Assumption 4.32. Hence, our assumption is more general than its counterpart in Di Masi and Stettner (2008).

(b) The assumption (A3) set in Section 3 of Di Masi and Stettner (2008) for the cost function c is implicit and difficult to be verified. On the contrary, the sufficient growth condition for c, (4.46), is explicit in form of the Lyapunov function w_1 w.r.t. the entropic map. Note that, in the example provided by Di Masi and Stettner (2008), the assumption (A3) is also verified with the help of a Lyapunov function.

(c) As an advantage, in comparison with Di Masi and Stettner (2008), the convergence rate of iterations towards the solution to the Poisson equation is explicitly specified by $\bar{\alpha}$ in Lemma 4.18 under the chosen seminorm.

Remarks on finite state-action spaces

For finite state-action spaces, Assumption 5.7 automatically holds. Following the same line as in Section 4.5, Assumptions 5.12 and 5.21 can be reduced to

a set of two conditions similar to Assumption 4.42. In particular, when applying the entropic map, one sufficient condition is to assume that the underlying MDP is ergodict, i.e., there exist positive constants $n_0 \in \mathbb{N}$ and $\epsilon > 0$ such that $(P^{f_0} P^{f_1} \dots P^{f_{n_0-1}})(y|x) \geq \epsilon$ for any $x, y \in \mathsf{X}$ and $f_0 \times f_1 \times \dots \times f_{n_0-1} \in \Delta^{n_0}$. The proof follows the same line as in Example 4.46 and is therefore omitted here.

Value iteration

We state below one value iteration algorithm (see Algorithm 5.2) for the average criterion. Lemma 5.22 and Theorem 5.23 guarantee that $v_{t+1} \to \rho^*$, the optimal

Algorithm 5.2 Value iteration for average problems

select one $v_0 \in \mathscr{B}_w^{(\tilde{K}_0)}$, $t = 0$;
repeat
calculate $v_{t+1} = \mathcal{F}(v_t)$ with selector

$$f_{t+1}(x) := \operatorname*{argmax}_{a \in \mathsf{A}(x)} \{r(x,a) + \mathcal{U}(v_t|x,a)\};$$

$t = t + 1$;
until $\|v_{t+1} - v_t\|_{s,w} < \epsilon$

average valuation, and $f_t \to f^*$, the optimal policy, as $t \to \infty$.

5.5 One example for average valuations with the entropic map

We have already presented in Section 4.4.3 an example of average valuation with the entropic map based on 1-dimensional AR1 process. In the current section, we extend the above example to a d-dimensional process equipped with some control variables.

Let $\mathsf{X} = \mathbb{R}^d$. Consider the following discretized ergodic diffusion $\{x_n \in \mathbb{R}^d\}$ (cf. the example in Di Masi and Stettner, 2008, Section 6):

$$x_{n+1} = Ax_n + b(x_n, a_n) + D(x_n, a_n) w_n,$$

where $\{w_n \in \mathbb{R}^d\}$ is a sequence of i.i.d. standard white noise, $D : \mathsf{K} \to \mathbb{R}^{d \times d}$ is a continuous bounded matrix-valued function which is uniformly elliptic, i.e., there exists a constant $L > 0$ such that

$$L^{-1}\|\xi\|^2 \leq \xi^\top D(x,a) D^\top(x,a) \xi \leq L\|\xi\|^2, \forall (x,a) \in \mathsf{K}, \xi \in \mathbb{R}^d, \tag{5.20}$$

and $b : \mathsf{K} \to \mathbb{R}^d$ is a continuous bounded vector function, and A is a matrix satisfying that there exists a constant $\tilde{\gamma} \in (0,1)$ such that $\xi^\top A^\top A \xi \leq \tilde{\gamma}\|\xi\|^2$,

$\forall \xi \in \mathbb{R}^d$. Then the transition kernel $\mathcal{P}(dy|x,a)$ has the following density with respect to the Lebesgue measure,

$$p(y|x,a) = (2\pi)^{-d/2}|\Sigma|^{1/2} e^{-\frac{1}{2}(y-Ax-b)^\top \Sigma (y-Ax-b)}, \tag{5.21}$$

where $\Sigma = (DD^\top)^{-1}$. Take one $\gamma \in (\tilde{\gamma}, 1)$ and consider the following weight function

$$\hat{w}_1(x) = \frac{\epsilon}{2}\|x\|^2, \text{ with some positive } \epsilon \leq \frac{\gamma - \tilde{\gamma}}{\gamma} L^{-1} < L^{-1}. \tag{5.22}$$

Hence, $\Sigma(x,a) - \epsilon I$ is positive definite for all $(x,a) \in \mathsf{K}$.

Lyapunov function We show that $\hat{w}_1$ is a Lyapunov function with respect to the entropic map satisfying the condition (i) in Proposition 5.24 as follows. By setting $\tilde{x} := Ax + b$, we obtain

$$\begin{aligned}\int \mathcal{P}(dy|x,a) e^{\hat{w}_1(y)} =& (2\pi)^{-d/2}|\Sigma|^{1/2} \int e^{-\frac{1}{2}\left(y^\top(\Sigma-\epsilon I)y - 2y^\top \Sigma \tilde{x} + \tilde{x}^\top \Sigma \tilde{x}\right)} dy \\ =& \frac{|\Sigma|^{1/2}}{|\Sigma - \epsilon I|^{1/2}} e^{\frac{1}{2}\tilde{x}^\top \Sigma \left((\Sigma-\epsilon I)^{-1} - \Sigma^{-1}\right)\Sigma \tilde{x}}\end{aligned}$$

which yields

$$\begin{aligned}\log\left(\mathcal{P}_{x,a}[e^{\hat{w}_1}]\right) =& \log\left(\frac{|\Sigma|^{1/2}}{|\Sigma - \epsilon I|^{1/2}}\right) + \\ & \frac{1}{2}(Ax+b)^\top \Sigma \left((\Sigma - \epsilon I)^{-1} - \Sigma^{-1}\right)\Sigma(Ax+b).\end{aligned}$$

By (5.20) and the choice of ϵ in (5.22), we have

$$\frac{1}{2}x^\top A^\top \Sigma \left((\Sigma - \epsilon I)^{-1} - \Sigma^{-1}\right)\Sigma A x \leq \frac{\gamma\epsilon}{2}\|x\|^2 = \gamma \hat{w}_1(x), \forall (x,a) \in \mathsf{K}.$$

Finally, due to the uniform boundedness of b and $\log\left(\frac{|\Sigma|^{1/2}}{|\Sigma-\epsilon I|^{1/2}}\right)$, we can always select a $\gamma_1 \in (\gamma, 1)$ and $\hat{K}_1 > 0$ such that

$$\log \int \mathcal{P}(dy|x,a) e^{\hat{w}_1(y)} \leq \gamma_1 \hat{w}_1(x) + \hat{K}_1, \forall (x,a) \in \mathsf{K},$$

which confirms that $\hat{w}_1 \geq 0$ is a Lyapunov function with respect to the entropic map. Hence, the condition (i) in Proposition 5.24 holds with $w_1 := \hat{w}_1 + 1$, γ_1 and $K_1 := \hat{K}_1 + 1 - \gamma_1$.

Deoblin's condition Since by (5.21) the transition kernel $\mathcal{P}$ has a positive continuous density function with respect to the Lebesgue measure, the local Doeblin's condition (ii) in Proposition 5.24 is obviously satisfied.

5.6 Discussion

We introduce briefly one alternative way to incorporate risk into the MDP framework. Risk is treated as a constraint, i.e., formally,

$$\sup_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x], \text{ subject to } \varrho^{\boldsymbol{\pi}}(\mathcal{S}|X_0 = x) \leq \theta.$$

Here, $\mathcal{S}$ can be finite-stage S_T, discounted reward S_γ or average reward S_A as in (5.4), while ϱ is a type of risk measure with θ being an acceptable risk level. Applying the Lagrange multiplier (see e.g. Bertsekas, 1999, Chapter 3), the above optimization problem becomes

$$\sup_{\boldsymbol{\pi}, \lambda \geq 0} \mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x] - \lambda \left(\varrho^{\boldsymbol{\pi}}(\mathcal{S}|X_0 = x) - \theta\right).$$

One can furthermore fix λ and obtain the following equivalent problem

$$\sup_{\boldsymbol{\pi}} \mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x] - \lambda \varrho^{\boldsymbol{\pi}}(\mathcal{S}|X_0 = x), \tag{5.23}$$

which can be interpreted as a tradeoff between the mean $\mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x]$ and the risk $\varrho^{\boldsymbol{\pi}}(\mathcal{S}|X_0 = x)$ with λ controlling the degree of risk-preference. If $\lambda > 0$, then the agent should be risk-averse, while $\lambda < 0$ induces risk-seeking behaviors.

Existing literature in this line includes

1) using variance as a measure of risk (see e.g. Sobel, 1982; Filar et al., 1989 and recently rediscovered by Tamar et al. (2012) and Prashanth and Ghavamzadeh (2013) in machine learning);

2) using mean cost as the constraint (see e.g. Altman, 1999; Geibel and Wysotzki, 2005; Dolgov and Durfee, 2003; Feinberg and Shwartz, 1996 and references therein) and

3) a recent work by Borkar and Jain (2010) with conditional value-at-risk (cf. Section 2.4.6) as a measure of risk.

Among them, the mean-variance tradeoff problem cannot be solved by value iterations, since it violates the axiom of monotonicity, as we have explained in Section 2.4.1. In fact, the algorithms stated by Tamar et al. (2012) and Prashanth and Ghavamzadeh (2013) can only ensure a local optimal solution.

All literature mentioned above considers only risk-averse behaviors. One way to induce risk-seeking behaviors or even mixed risk preferences is to apply specially designed risk measures correspondingly. For instance, to induce mixed risk preferences, we can apply the utility based shortfall with S-shaped utility functions (for more details see Section 2.4.5).

Finally, note that under some technical assumptions, we may define a valuation function or map composed of the mean and risk, i.e.

$$\rho(\mathcal{S}|X_0 = x) := \mathbb{E}^{\boldsymbol{\pi}}[\mathcal{S}|X_0 = x] - \lambda \varrho^{\boldsymbol{\pi}}(\mathcal{S}|X_0 = x).$$

Then the objective (5.23) has already been covered by our risk-sensitive framework. In the future, we plan to specify these assumptions and make more detailed comparison of the objective (5.23) with the one considered in this chapter.

6

Risk-sensitive Q-Learning

There are many arts among mankind which are experimental, and have their origin in experience, for experience makes the days of men to proceed according to art, and inexperience according to chance.
— Plato, *Gorgias*

Précis In this chapter, we will restate the framework of risk-sensitive Markov decision processes (MDPs) on finite spaces with, however, a slightly more general setting for noisy rewards. The derived optimization problems are again to be solved by value iteration. After applying a rich family of valuation maps, the utility-based shortfall, to the general framework, we will develop a risk-sensitive Q-learning algorithm, which is necessary for modeling human behavior when transition probabilities are unknown. The derived algorithm applies the utility function to the temporal difference (TD) error, which can be interpreted as applying nonlinear transformations to both rewards and true transition probabilities of the underlying MDP. Finally, we will prove that the proposed algorithm converges to the optimal policy corresponding to the risk-sensitive objective.

Publication related to this chapter This chapter is based on Shen et al., 2014d, Section 3.

6.1 Introduction

We have shown in the last chapter that the objectives of risk-sensitive Markov decision processes (MDPs) can be solved by value iteration algorithms. These algorithms require the entire knowledge of the underlying MDP, namely the reward function and the transition probabilities. In many real-life situations, however, the transition probabilities are unknown, as well as the outcome of an action before its execution. To gather information of rewards and transition probabilities, a decision maker has to explore the whole environment sufficiently with some policies.

Ideally, we may hope that the policy can be gradually improved in course of exploring the environment. *Reinforcement learning* (RL, see e.g. Sutton and Barto, 1998) is exactly the technique that tells decision makers how to improve their policies while exploring the environment. Standard RL algorithms are to maximize cumulative discounted rewards (corresponding to discounted MDPs) or average rewards (corresponding to average MDPs). In this chapter, we accordingly develop RL algorithms for risk-sensitive MDPs developed in the last chapter. In particular, we focus mainly on a special type of RL algorithm, called *Q-learning* (Watkins, 1989), due to the following two reasons:

1) other types of RL algorithms can be derived easily in the same line, and more importantly,

2) Q-learning is a well-developed model also for human decision making and for free choice in nonhumans as well. Similar computational structures, such as dopaminergically mediated reward prediction errors, have been identified across species (Schultz et al., 1997; Schultz, 2002).

In this chapter, the terms *reinforcement learning (RL)* and *Q-learning* will be used in this chapter interchangeably, unless stated otherwise.

Most of the RL literature related to risk focuses on risk-averse control, where the aim is to avoid risk within the framework of MDPs. Coraluppi and Marcus (2000) and Heger (1994) applied the worst case control, which is equivalent to applying the type of valuation function introduced in Section 2.4.3. Another approach is to apply the entropic map, see e.g. Borkar (2002); Liu et al. (2003); Koenig and Simmons (1994). Mihatsch and Neuneier (2002) utilized the valuation function described in (2.15), which is a special case of utility-based shortfall introduced in Section 2.4.5. By selecting proper risk parameters, their approach can also induce risk-seeking behaviors.

All the approaches mentioned above can be viewed as applying special cases of utility-based shortfall. In this chapter, we will show that under some technical assumptions on the utility function, we may derive risk-sensitive Q-learning from utility-based shortfall with arbitrary utility functions. As we have seen in Section 2.4.5, this family of valuation functions can induce all types of risk preferences including mixed ones. For instance, by carefully choosing utility functions, we can also replicate (see Section 2.4.5) key features of human behavior as predicted by prospect theory (Kahneman and Tversky, 1979), e.g., different risk-preferences for gains and losses as well as the shape of subjective probability curves. Hence, the risk-sensitive RL algorithm to be developed in this chapter provides a good framework for quantifying the sequential decision making procedures of humans.

6.2 Risk-sensitive Markov decision processes on finite spaces

6.2.1 Markov decision processes on finite spaces

In the last chapter, we have already introduced the framework of Markov decision processes (MDPs). Here we restate the framework with a slightly more general setting for rewards, that is, the reward can be also "noisy": the reward can be decomposed into two parts: $R = r + n$, where r denotes the reward function as in the framework applied in this chapter, and n denotes some additive noise (real-valued random variable), whose distribution might be dependent on the state-action pair (x, a). In addition, we restrict to finite state-action spaces. The reason of this limitation is two-fold. First, to our best knowledge, mathematically rigorous convergence proofs for the Q-learning algorithm, which is the ultimate goal of this chapter, exist merely for MDPs with finite state-action spaces. Second, the application of the derived risk-sensitive Q-learning, as well as other potential applications in neuroeconomics, has also the same restrictive setting.

Setup

A Markov decision process (see e.g., Puterman, 1994)

$$\mathcal{M} = \{\mathsf{X}, (\mathsf{A}, \mathsf{A}(x), x \in \mathsf{X}), \mathcal{P}, (r, \mathcal{P}_r)\},$$

consists of a finite state space X, admissible finite action spaces $\mathsf{A}(x) \subset \mathsf{A}$ at $x \in \mathsf{X}$, a transition kernel $\mathcal{P}(x'|x, a)$, which denotes the transition probability moving from one state x to another state x' by executing action a, and a reward function r with its distribution $\mathcal{P}_r$. In order to model random rewards, we assume that the reward function has the form

$$r(x, a, \varepsilon) : \mathsf{X} \times \mathsf{A} \times \mathsf{E} \to \mathbb{R}.$$

E (with its Borel σ-algebra $\mathcal{B}(\mathsf{E})$) denotes the noise space with distribution $\mathcal{P}_r$, i.e., given (x, a), $r(x, a, \varepsilon)$ is a random variable with values drawn from $\mathcal{P}_r(\cdot|x, a)$. Let $R(x, a)$ be the *random* reward gained at (x, a), which follows the distribution $\mathcal{P}_r(\cdot|x, a)$. The random state (respectively action) at time t is denoted by X_t (respectively A_t).

Remark 6.1. In standard MDPs, it is sufficient (Puterman, 1994) to consider the *deterministic* reward function

$$\bar{r}(x, a) := \int_{\mathsf{E}} r(x, a, \varepsilon) \mathcal{P}_r(d\varepsilon|x, a),$$

i.e., the mean reward at each (x, a)-pair. In risk-sensitive cases, random rewards cause also risk and uncertainties. Hence, we keep the generality by using random rewards.

Policy and objectives

A *Markov policy* $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots]$ consists of a sequence of single-step Markov policies at times $t = 0, 1, \ldots$, where $\pi_t(A_t = a | X_t = x)$ denotes the probability of choosing action a at state x. Let Π_M be the set of all Markov policies. The optimal policy within a time horizon T is obtained by maximizing the expectation of the discounted cumulative rewards,

$$J_T(\boldsymbol{\pi}, x) := \sup_{\boldsymbol{\pi} \in \Pi_M} \mathbb{E}^{\boldsymbol{\pi}} \left[\sum_{t=0}^{T} \gamma^t R(X_t, A_t) \middle| X_0 = x \right]. \tag{6.1}$$

where $x \in \mathsf{X}$ denotes the initial state and $\gamma \in [0, 1)$ the discount factor. Expanding the sum leads to

$$J_T(\boldsymbol{\pi}, x) = \mathbb{E}^{\pi_0}_{X_0 = x} \Big[R(X_0, A_0) + \gamma \mathbb{E}^{\pi_1}_{X_1} \Big[R(X_1, A_1) + \ldots + \gamma \mathbb{E}^{\pi_T}_{X_T} [R(X_T, A_T)] \ldots \Big] \Big]. \tag{6.2}$$

6.2.2 Discounted risk-sensitive objectives

Valuation maps

Let $\mathscr{L}(\mathsf{X} \times \mathsf{E})$ be the space of all functions $f : \mathsf{X} \times \mathsf{E} \to \mathbb{R}$ such that for each $x \in \mathsf{X}$ $f(x, \cdot)$ is $\mathcal{B}(\mathsf{E})$-measurable. Let $\mathscr{P}(\mathsf{X} \times \mathsf{E})$ be the space of all distributions on $\mathsf{X} \times \mathsf{E}$. Define $\mathsf{K} := \{(x, a) \mid x \in \mathsf{X}, a \in \mathsf{A}(x)\}$. Then the valuation map considered in this chapter is defined as

Definition **6.2.** *A mapping $\mathcal{U}(v, \mu | x, a) : \mathscr{L}(\mathsf{X} \times \mathsf{E}) \times \mathscr{P}(\mathsf{X} \times \mathsf{E}) \times \mathsf{K} \to \mathbb{R}$ is called a* valuation map*, if for each $(x, a) \in \mathsf{K}$, $\mathcal{U}(\cdot | x, a)$ is a valuation function (see Definition 2.6) on $\mathscr{L}(\mathsf{X} \times \mathsf{E}) \times \mathscr{P}(\mathsf{X} \times \mathsf{E})$.*

Let $\mathcal{U}_{x,a}(v, \mu)$ be a short notation of $\mathcal{U}(v, \mu | x, a)$ and let

$$\mathcal{U}^{\pi}_{s}(v, \mu) := \sum_{a \in \mathsf{A}(x)} \pi(a|x) \mathcal{U}(v, \mu | x, a)$$

be the valuation map averaged over all actions. In the context of MDPs defined in this chapter, we consider merely

$$\mu(x', d\varepsilon) = \mathcal{P}(x' | x, a) \mathcal{P}_r(d\varepsilon | x, a)$$

and therefore omit μ in $\mathcal{U}$ in the following.

Risk-sensitive objectives

Replacing the conditional expectation $\mathbb{E}_s^\pi$ with $\mathcal{U}_s^\pi$ in (6.2), the risk-sensitive objective becomes

$$\tilde{J}_{\gamma,T}(\boldsymbol{\pi},x) := \mathcal{U}^{\pi_0}_{X_0=x}\Big[R(X_0,A_0) + \gamma\mathcal{U}^{\pi_1}_{X_1}\Big[R(X_1,A_1) + \ldots + \gamma\mathcal{U}^{\pi_T}_{X_T}\left[R(X_T,A_T)\right]\ldots\Big]\Big]. \tag{6.3}$$

The optimal policy is then given by $\sup_{\boldsymbol{\pi}\in\Pi_M}\tilde{J}_{\gamma,T}(\boldsymbol{\pi},x)$. For infinite-horizon problem, we obtain

$$\sup_{\boldsymbol{\pi}\in\Pi_M}\quad \tilde{J}_\gamma(\boldsymbol{\pi},x) := \lim_{T\to\infty}\tilde{J}_{\gamma,T}(\boldsymbol{\pi},x), \tag{6.4}$$

using the same line of argument.

Finally, to ensure that $\tilde{J}_\gamma$ is finite for all policies, we assume throughout this chapter:

Assumption 6.3. $\bar{R} := \max_{(x,a)\in\mathsf{K}}|\mathcal{U}_{x,a}(R(x,a))| < \infty$.

Proposition **6.4.** *Under Assumption 6.3, for each $x\in\mathsf{X}$ and $\boldsymbol{\pi}\in\Pi_M$,*

(i) $|\tilde{J}_{\gamma,T}(\boldsymbol{\pi},x)| \le \frac{\bar{R}}{1-\gamma}$ *holds for each* $T\in\mathbb{N}$.

(ii) $J_\gamma(\boldsymbol{\pi},x) := \lim_{T\to\infty}\tilde{J}_{\gamma,T}(\boldsymbol{\pi},x)$ *exists and satisfies* $|\tilde{J}_\gamma(\boldsymbol{\pi},x)| \le \frac{\bar{R}}{1-\gamma}$.

Proof. (i) Define $v_T(x) := \mathcal{U}^{\pi_T}_{X_T=x}\left[R(X_T,A_T)\right]$ and

$$v_t(x) := \mathcal{U}^{\pi_t}_{X_t=x}\left[R(X_t,A_t) + \gamma v_{t+1}\right], t = T-1, T-2, \ldots, 0.$$

Then, it is easy to check that $v_0(x) = \tilde{J}_{\gamma,T}(\boldsymbol{\pi},x)$. By Assumption 6.3, $v_T(x) \le \bar{R}, \forall x\in\mathsf{X}$ implies that

$$v_{T-1}(x) \le \mathcal{U}^{\pi_t}_{X_t=x}\left[R(X_t,A_t)\right] + \gamma\max_{x\in\mathsf{X}} v_{t+1}(x) \le \bar{R} + \gamma\bar{R}, \forall x\in\mathsf{X}.$$

By iteration, we obtain, therefore, for each $\forall x\in\mathsf{X}$,

$$v_0(x) \le \left(\sum_{i=0}^{T}\gamma^i\right)\bar{R} \le \frac{\bar{R}}{1-\gamma}.$$

Repeat the same iteration, we obtain analogously that for each $\forall x\in\mathsf{X}$,

$$v_0(x) \ge -\left(\sum_{i=0}^{T}\gamma^i\right)\bar{R} \ge -\frac{\bar{R}}{1-\gamma}.$$

We therefore obtain (i).

(ii) Using the same iterative procedure, it is easy to check for all $x\in\mathsf{X}$ and $\boldsymbol{\pi}\in\Pi_M$ the following inequality holds

$$|J_{\gamma,T+1}(x,\boldsymbol{\pi}) - J_{\gamma,T}(x,\boldsymbol{\pi})| \le \gamma^{T+1}\bar{R},$$

which implies the existence of the limit. Finally, the inequality $|\tilde{J}_\gamma(\boldsymbol{\pi},x)| \le \frac{\bar{R}}{1-\gamma}$ is an immediate result of (i). □

6.2.3 Value iteration

Now we consider the optimization problem of discounted MDPs. Let $|\mathsf{X}|$ be the cardinality of the state space X. Then a function $v : \mathsf{X} \to \mathbb{R}$ can be considered as an $|\mathsf{X}|$-dimensional vector in $\mathbb{R}^{|\mathsf{X}|}$. Define an operator $\mathcal{F}_\gamma : \mathbb{R}^{|\mathsf{X}|} \to \mathbb{R}^{|\mathsf{X}|}$ as

$$\mathcal{F}_\gamma(v)(x) := \max_{a\in\mathsf{A}(x)} \mathcal{U}_{x,a}(R(x,a) + \gamma v), v \in \mathbb{R}^{|\mathsf{X}|}.$$

The norm $\|\cdot\|_\infty$ on $\mathbb{R}^{|\mathsf{X}|}$ is

$$\|v\|_\infty := \max_{x\in\mathsf{X}} |v(x)|, v \in \mathbb{R}^{|\mathsf{X}|}.$$

We first state the following contraction property for $\mathcal{F}_\gamma$, which plays a crucial role in deriving *value iteration* algorithms for discounted risk-sensitive MDPs.

Lemma **6.5.** *Suppose Assumption 6.3 holds. Then for all $v, u \in \mathbb{R}^{|\mathsf{X}|}$,*

$$\|\mathcal{F}_\gamma(v) - \mathcal{F}_\gamma(u)\|_\infty \le \gamma \|v - u\|_\infty.$$

Proof. It is sufficient to show that for each $(x, a) \in \mathsf{K}$,

$$|\mathcal{U}_{x,a}(R(x,a) + \gamma v) - \mathcal{U}_{x,a}(R(x,a) + \gamma u)| \le \gamma \max_{x\in\mathsf{X}} |v(x) - u(x)|.$$

In fact, for each $x \in \mathsf{X}$, we have

$$\underline{d} := \min_{x\in\mathsf{X}} (v(x) - u(x)) \le v(x) - u(x) \le \max_{x\in\mathsf{X}} (v(x) - u(x)) =: \bar{d},$$

which implies that for each $(x, a) \in \mathsf{K}$,

$$\mathcal{U}_{x,a}(R(x,a) + \gamma v) \le \mathcal{U}_{x,a}(R(x,a) + \gamma u + \gamma \bar{d}) = \mathcal{U}_{x,a}(R(x,a) + \gamma u) + \gamma \bar{d}.$$

Hence, for each $(x, a) \in \mathsf{K}$,

$$\mathcal{U}_{x,a}(R(x,a) + \gamma v) - \mathcal{U}_{x,a}(R(x,a) + \gamma u) \le \gamma \bar{d} \le \gamma \max_{x\in\mathsf{X}} |v(x) - u(x)|.$$

Analogously, we obtain for each $(x, a) \in \mathsf{K}$

$$\mathcal{U}_{x,a}(R(x,a) + \gamma v) - \mathcal{U}_{x,a}(R(x,a) + \gamma u) \ge \gamma \underline{d} \ge -\gamma \max_{x\in\mathsf{X}} |v(x) - u(x)|.$$

Combining the above two inequalities yields the required inequality. □

Hence, starting with one vector $v_0 \in \mathbb{R}^{|\mathsf{X}|}$, we consider the iteration:

$$v_{t+1} := \mathcal{F}_\gamma(v_t), t = 0, 1, \ldots.$$

The contraction property proved in the above lemma guarantees that the sequence of $\{v_t\}$ converges to the unique fixed point v^* of $\mathcal{F}_\gamma$ in $\mathbb{R}^{|\mathsf{X}|}$ that satisfies the following *risk-sensitive Bellman equation*:

$$v^*(x) = \max_{a\in\mathsf{A}(x)} \mathcal{U}_{x,a}(R(x,a) + \gamma v^*(x')), x \in \mathsf{X}. \tag{6.5}$$

Furthermore, this unique fixed point is "optimal" in the following sense.

THEOREM **6.6.** *Suppose Assumption 6.3 holds. Then*

$$v^*(x) = \max_{\boldsymbol{\pi}\in\Pi_M} \tilde{J}(\boldsymbol{\pi}, x)$$

holds for all $x \in \mathsf{X}$*, whenever* v^* *satisfies the equation* (6.5)*. Furthermore, a stationary deterministic policy* $\boldsymbol{\pi}^* = (\pi^*)^\infty$ *is optimal, if*

$$\pi^*(x) = \operatorname*{argmax}_{a\in\mathsf{A}(x)} \mathcal{U}_{x,a}(R(x,a) + \gamma v^*(x')).$$

Proof. The proof is similar to the proof of Theorem 5.18 and is, therefore, omitted here. □

6.3 Risk-sensitive Q-learning

Define $q^*(x,a) := \mathcal{U}_{x,a}(R(x,a) + \gamma v^*(x,a))$. Then (6.5) becomes

$$q^*(x,a) = \mathcal{U}_{x,a}\left(R(x,a) + \gamma \max_{a\in\mathsf{A}(x')} q^*(x',a)\right), \forall (x,a) \in \mathsf{K}. \tag{6.6}$$

To carry out value iteration algorithms, the MDP $\mathcal{M}$ must be known *a priori*. In many real-life situations, however, the transition probabilities are unknown as well as the outcome of an action before its execution. Therefore, an agent has to explore the environment while gradually improving its policy. We now derive reinforcement learning type algorithms for estimating Q-values of general valuation maps based on the utility-based shortfall defined in Section 2.4.5.

6.3.1 Utility-based shortfall: revisited

Let $u : \mathbb{R} \to \mathbb{R}$ be a continuous, increasing and non-constant utility function satisfying $u(0) = 0$. Recall that the *utility-based shortfall* ρ^u (see also Section 2.4.5) is then defined as

$$\rho^u(v,\mu) := \sup\left\{m \in \mathbb{R} \,\middle|\, \int_\Omega u(v(\omega) - m)\mu(d\omega) \geq 0\right\}. \tag{6.7}$$

For regularity, we assume

Assumption 6.7. There exists a constant $m \in \mathbb{R}$ such that

$$\int_\Omega u(v(\omega) - m)\mu(d\omega) < \infty.$$

This type of valuation function is of particular interest due to the following features.

1. By Proposition 2.15, the optimal m^* in (6.7) is attained when the inequality holds, i.e., we have

$$\mathbb{E}^{\mu}[u(v - m^*)] = 0 \text{ where } m^* := \rho^u(v, \mu).$$

   Hence, one can obtain ρ^u by solving the above stochastic equation.

2. As we have shown in Section 2.4.5, by selecting an appropriate utility function u, the agents' behaviors express key features of human behavior as predicted by prospect theory (Kahneman and Tversky, 1979), for example different risk-preferences for gains and losses as well as the shape of subjective probability curves.

6.3.2 Algorithm for the finite-stage criterion

Setting $\gamma = 1$ in (6.3), we have the following finite-stage risk-sensitive objective

$$\tilde{J}_T(\boldsymbol{\pi}, x) := \mathcal{U}^{\pi_0}_{X_0=x}\left[R(X_0, A_0) + \mathcal{U}^{\pi_1}_{X_1}\left[R(X_1, A_1) + \ldots + \mathcal{U}^{\pi_T}_{X_T}\left[R(X_T, A_T)\right]\ldots\right]\right].$$

Starting with $v_{T+1}(x) := 0, \forall x \in \mathsf{X}$, we consider the following backward induction (called also *dynamic programming*)

$$v_t(x) := \max_{a \in \mathsf{A}(x)} \mathcal{U}_{x,a}(R(x, a) + v_{t+1}(x')), t = T, T-1, \ldots, 0.$$

It is easy to verify that $v_0(x) = \max_{\boldsymbol{\pi} \in \Pi_M} \tilde{J}_T(\boldsymbol{\pi}, x), \forall x \in \mathsf{X}$. Let $q_{T+1}(x, a) := 0$ and $q_t(x, a) := \mathcal{U}_{x,a}(R(x, a) + v_{t+1})$, the above update step becomes

$$q_t(x) := \mathcal{U}_{x,a}(R(x, a) + \max_{a \in \mathsf{A}(x')} q_{t+1}(x')). \tag{6.8}$$

We apply now the utility based shortfall

$$\mathcal{U}_{x,a}(v) = \sup\{m \in \mathbb{R} \mid \mathbb{E}^{\mu_{x,a}}[u(v - m)] \geq 0\}, \tag{6.9}$$

where $\mu_{x,a}(\cdot, \cdot) := \mathcal{P}(\cdot|x, a)\mathcal{P}_r(\cdot|x, a)$ denotes the joint distribution of the successive state x' and the additive noise ε.

At time t, given q_{t+1}, q_t in (6.8) is obtained by solving the following stochastic equation

$$\sum_{x' \in \mathsf{X}} \mathcal{P}(x'|x, a) \int_{\mathsf{E}} \mathcal{P}_r(d\varepsilon|x, a) u\left(r(x, a, \varepsilon) + \max_{a' \in \mathsf{A}(x')} q_{t+1}(x', a') - q_t(x, a)\right) = 0, \forall (x, a) \in \mathsf{K}.$$

Numerically, the above equation can be solved by *stochastic approximation* algorithms (Kushner and Yin, 2003; Borkar, 2008). More specially, at time t, we repeat trying action a at state x a large amount of times and observe N samples of immediate rewards and successive states, $\{R_i, x'_i\}_{i=1,2,\ldots,N}$. Then the q-value at (x, a),

which evaluates the quality of this state-action pair, can be estimated by the following iterative procedure

$$
\begin{aligned}
q_t^{(i+1)}(x,a) =& q_t^{(i)}(x,a) + \frac{1}{i} u(\delta), \\
\text{with } \delta =& R_i + \max_a q_{t+1}(x_i', a) - q_t^{(i)}(x,a).
\end{aligned} \tag{6.10}
$$

Note that q_{t+1} is already known at time t. This iterative procedure is summarized in Algorithm 6.1.

Algorithm 6.1 Risk-sensitive Q learning for the finite-stage criterion

```
initialize q_{T+1}(x,a) = 0 for all x ∈ X, a ∈ A;
for t = T to 0 do
    initialize q_t(s,a) = 0 for all x ∈ X, a ∈ A;
    for each state x ∈ X and a ∈ A do
        for i = 1 to N do
            execute action a at x to obtain sampled
            reward R and successive state x';
            Update q_t(x,a) according to (6.10);
        end for
    end for
end for
```

It is shown in Dunkel and Weber (2010) that if the utility function u satisfies some regularity conditions, then for each time point $t = T, T-1, \ldots, 1$, the policy $\pi_t^{(N)} = \max_{a \in A} q_t^{(N)}(s,a)$ converges to the optimal policy that maximizes the risk-sensitive objective function defined in (6.3), as $N \to \infty$.

6.3.3 Algorithm for the discounted criterion

We again apply the utility-based shortfall defined in (6.9). Suppose $\mathcal{U}_{x,a}(X) = m^*(x,a)$, Proposition 2.15 assures that $m^*(x,a)$ is the unique solution to equation

$$
\mathbb{E}^{\mu_{x,a}}\left[u(X - m^*(x,a))\right] = 0.
$$

Let $v = R + \gamma v^*$. Then $m^*(x,a)$ corresponds to the optimal Q-value $q^*(x,a)$ defined in (6.6), which is equivalent to

$$
\begin{aligned}
\sum_{x' \in \mathsf{X}} \mathcal{P}(x'|x,a) \int_{\mathsf{E}} \mathcal{P}_r(d\varepsilon|x,a) u\left(r(x,a,\varepsilon) + \gamma \max_{a' \in \mathsf{A}(x')} q^*(x',a') - q^*(x,a)\right) \\
= 0, \forall (x,a) \in \mathsf{K}.
\end{aligned} \tag{6.11}
$$

Let $\{X_t, A_t, X_{t+1}, R_t\}$ be the sequence of states, chosen actions, successive states and received rewards, which are all random variables. Analogous to the standard

Q-learning algorithm, we consider the following iterative procedure,

(6.12)
$$q_{t+1}(X_t, A_t) = q_t(X_t, A_t) + \alpha_t(X_t, A_t) u\left(R_t + \gamma \max_a q_t(X_{t+1}, a) - q_t(X_t, A_t)\right),$$

where $\alpha_t \geq 0$ denotes learning rate function that satisfies $\alpha_t(x,a) > 0$ only if (x,a) is updated at time t, i.e., $(x,a) = (X_t, A_t)$. In other words, for all (x,a) that are not visited at time t, $\alpha_t(x,a) = 0$ and their Q-values are not updated. Consider utility functions u with the following properties.

Assumption 6.8. (i) The utility function u is strictly increasing and satisfies $u(0) = 0$.

(ii) There exist positive constants ϵ, L such that $0 < \epsilon \leq \frac{u(x)-u(y)}{x-y} \leq L$, for all $x \neq y \in \mathbb{R}$.

In addition, we assume the random rewards R satisfies

Assumption 6.9. $\mathbb{E}^{\mathcal{P}_r}\left[u^2(R(x,a))\right] < \infty$ holds for each $(x,a) \in \mathsf{K}$.

Remark 6.10. It is easy to check that this assumption is a sufficient condition that implies Assumption 6.7 and Assumption 6.3, and therefore Theorem 6.6 guarantees a unique solution to the Bellman equation (6.5) and a unique solution to the stochastic equation (6.11) as well. Furthermore, note that under Assumption 6.8, a sufficient condition for Assumption 6.9 is

$$\mathbb{E}^{\mathcal{P}_r}\left[R^2(x,a)\right] < \infty, \forall (x,a) \in \mathsf{K},$$

due to the fact that u is Lipschitz.

Then the following theorem holds (for proof see Section 6.3.4).

Theorem **6.11.** *Suppose Assumption 6.8 and 6.9 hold. Consider the generalized Q-learning algorithm stated in* (6.12)*. If the nonnegative learning rates $\alpha_t(x,a)$ satisfy*

$$\sum_{t=0}^{\infty} \alpha_t(x,a) = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \alpha_t^2(x,a) < \infty, \quad \forall (x,a) \in \mathsf{K}, \tag{6.13}$$

then $q_t(x,a)$ converges to $q^*(x,a)$ for all $(x,a) \in \mathsf{K}$ with probability 1.

The assumption in (6.13) requires in fact that all possible state-action pairs must be visited infinitely often. Otherwise, the first sum in (6.13) would be bounded by the setting of the learning rate function $\alpha_t(x,a)$. It means that, similar to the standard Q-learning, the agent has to explore the whole state-action space for gathering sufficient information about the environment. Hence, it can not take a too greedy policy in the learning procedure before the state-action space is well explored. We then introduced the concept of *proper* policies (see also (Bertsekas and Tsitsiklis, 1996, Definition 2.1)) as below.

DEFINITION **6.12.** *A policy is said to be* proper*, if under such policy every state is visited infinitely often.*

A typical policy, which is widely applied in RL literature as well as in models of human reward-based learning, is given by

$$A_t \text{ is drawn according to the distribution } P(a|X_t) := \frac{e^{\beta q(X_t,a)}}{\sum_a e^{\beta q(X_t,a)}}, \tag{6.14}$$

where $\beta \in [0,\infty)$ controls how greedy the policy should be. In Section 6.3.5, we prove that under some technical assumptions upon the transition kernel of the underlying MDP, this policy is always proper. A widely used setting satisfying both conditions in (6.13) is to let $\alpha_t(x,a) := \frac{1}{N_t(x,a)}$, where $N_t(x,a)$ counts the number of times of visiting the state-action pair (x,a) up to time t and is updated trial-by-trial. This leads to the learning procedure shown in Algorithm 6.2 (see also Figure 6.1).

Algorithm 6.2 Risk-sensitive Q-learning

initialize $q(x,a) = 0$ and $N(x,a) = 0$ for all $(x,a) \in \mathsf{X} \times \mathsf{A}$.
for $t = 1$ to T **do**
  at state X_t choose action A_t randomly using a proper policy (e.g. (6.14));
  observe date (X_t, A_t, R_t, X_{t+1});
  $N(X_t,A_t) \Leftarrow N(X_t,A_t) + 1$ and set learning rate: $\alpha_t := 1/N(X_t,A_t)$;
  update q as in (6.12);
end for

The expression

$$TD_t := R_t + \gamma \max_a q_t(X_{t+1},A_t) - q_t(X_t,A_t)$$

inside the utility function of (6.12) corresponds to the standard *temporal difference* (TD) error. Comparing (6.12) with the standard Q-learning algorithm, we find that the nonlinear utility function is applied to the TD error (cf. Figure 6.1). This induces nonlinear transformation not only of the true rewards but also of the true transition probabilities, as has been shown in Section 2.4.5. By applying S-shape utility function, which is partially convex and partially concave, we can therefore replicate key effects of prospect theory without the explicit introduction of a probability-weighting function.

Assumption 6.8 (ii) seems to exclude several important types of utility functions. The exponential function $u(x) = e^x$ and the polynomial function $u(x) = x^p$, $p > 0$, for example, do not satisfy the global Lipschitz condition required in Assumption 6.8 (ii). This problem can be solved by a truncation when x is very large and by an approximation when x is very close to 0. For more details see Section 6.3.5.

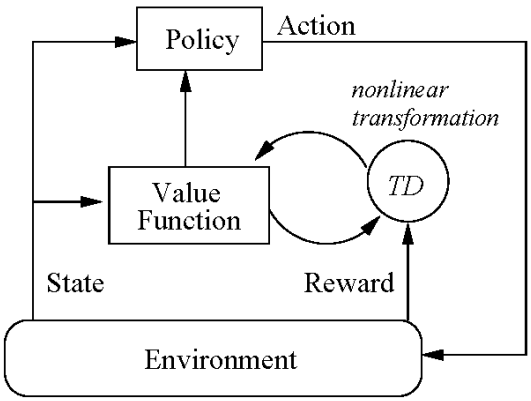


Figure 6.1: Illustration of risk-sensitive Q-learning (cf. Algorithm 6.2). The value function $q(x,a)$ quantifies the current subjective evaluation of each state-action pair (x,a). The next action is then randomly chosen according to a proper policy (e.g. (6.14)) which is based on the current values of q. After interacting with the environment, the agent obtains the reward r and moves to the successor x'. The value function $q(x,a)$ is then updated by the rule given in (6.12). This procedure continues until some stopping criterion is satisfied.

6.3.4 Convergence proof

The Q-learning algorithm (similar to Algorithm 6.2) was first introduced by Watkins (1989) in his PhD thesis, where a sketched proof of its convergence is contained. Due to its wide applications, Tsitsiklis (1994) and Jaakkola et al. (1994) presented new proofs based on the martingale and ordinary differential equation (ODE) methods separately. Q-learning can be in fact viewed as a special case of the asynchronous stochastic approximation approach (see e.g., Borkar, 1998, Borkar, 2008, Chapter 7 and references therein). Our proof below follows the line of a similar proof in Mihatsch and Neuneier (2002), which is based on Tsitsiklis' martingale method (see also Bertsekas and Tsitsiklis, 1996, Chapter 4). In fact, the key result (see Proposition 6.13 below) can be also proved using the ODE method developed by Borkar. For detailed arguments, see Borkar, 2008, Section 7.4 and Section 10.3.

Before proving the risk-sensitive Q-learning, we consider a more general update rule

$$q_{t+1}(i) = (1 - \alpha_t(i))q_t(i) + \alpha_t(i)\left[H(q_t)(i) + w_t(i)\right]. \tag{6.15}$$

where $q_t \in \mathbb{R}^d$, $H : \mathbb{R}^d \to \mathbb{R}^d$ is an operator, w_t denotes some random noise term and α_t is learning rate with the understanding that $\alpha_t(i) = 0$ if $q(i)$ is not updated at time t. Denote by $\mathcal{F}_t$ the history of the algorithm up to time t,

$$\mathcal{F}_t = \{q_0(i), \ldots, q_t(i), w_0(i), \ldots, w_{t-1}(i), \alpha_0(i), \ldots, \alpha_t(i), i = 1, \ldots, t\}.$$

We restate the following proposition.

PROPOSITION **6.13** (Proposition 4.4, Bertsekas and Tsitsiklis (1996)). *Let q_t be the sequence generated by the iteration (6.15). We assume the following*

(a) The learning rates $\alpha_t(i)$ are nonnegative and satisfy

$$\sum_{t=0}^{\infty} \alpha_t(i) = \infty, \quad \sum_{t=0}^{\infty} \alpha_t^2(i) = \infty, \forall i$$

(b) The noise terms $w_t(i)$ satisfy (i) for every i and t, $\mathbb{E}[w_t(i)|\mathcal{F}_t] = 0$; (ii) Given some norm $\|\cdot\|$ on $\mathbb{R}^d$, there exist constants A and B such that $\mathbb{E}[w_t^2(i)|\mathcal{F}_t] \leq A + B\|q_t\|^2$.

(c) The mapping H is a contraction under sup-norm.

Then q_t converges to the unique solution q^* of the equation $H(q^*) = q^*$ with probability 1.

To apply Proposition 6.13, we first reformulate the Q-learning rule (6.12) in a different form

$$q_{t+1}(x,a) = (1 - \frac{\alpha_t(x,a)}{\alpha})q_t(x,a) + \frac{\alpha_t(x,a)}{\alpha}\left[\alpha u(d_t) + q_t(x,a)\right]$$

where α denotes an arbitrary constant such that $\alpha \in (0, \min(L^{-1}, 1)]$. Recall that L is defined in Assumption 6.8. For simplicity, we define

$$D_t(x,a) := R(x,a) + \gamma \max_a q_t(X_{t+1}, a) - q_t(x,a)$$

and set

$$H(q_t)(x,a) := \alpha \mathbb{E}_{x,a} u(R(x,a) + \gamma \max_a q_t(X_{t+1}, a) - q_t(x,a)) + q_t(x,a) \tag{6.16}$$

$$w_t(x,a) := \alpha u(D_t(x,a)) - H(q_t)(x,a) \tag{6.17}$$

More explicitly, $H : \mathbb{R}^d \to \mathbb{R}^d$ is defined as

$$\begin{aligned} H(q)(x,a) =& \alpha \sum_{x' \in \mathsf{X}} \int_{\mathsf{E}} \tilde{\mathcal{P}}(x', d\epsilon|x,a) u\left(r(x,a,\varepsilon) + \gamma \max_{a'} q(x',a') - q(x,a)\right) \\ &+ q(x,a), (x,a) \in \mathsf{K}, \end{aligned}$$

where $\tilde{\mathcal{P}}(x', d\epsilon|x,a) := \mathcal{P}(x'|x,a)\mathcal{P}_r(d\varepsilon|x,a)$. We assume the cardinality of the space K is d.

LEMMA **6.14.** *Suppose that Assumption 6.8 holds and $0 < \alpha \leq \min(L^{-1}, 1)$. Then there exists a real number $\bar{\alpha} \in (0,1)$ such that for all $q, q' \in \mathbb{R}^d$,*

$$\|H(q) - H(q')\|_\infty \leq \bar{\alpha}\|q - q'\|_\infty.$$

Proof. Define $v(x) := \max_a q(x,a)$ and $v'(x) := \max_a q'(x,a)$. Thus,

$$|v(s) - v(s)| \leq \max_{(x,a)\in \mathsf{K}} |q(x,a) - q'(x,a)| = \|q - q'\|_\infty.$$

By Assumption 6.8 (ii) and the monotonicity of u, there exists a $\xi_{(x,y)} \in [\epsilon, L]$ such that $u(x) - u(y) = \xi_{(x,y)}(x - y)$. Analogously, we obtain

$$\begin{aligned}
&(Hq)(x,a) - (Hq')(x,a)\\
=&\sum_{x'} \int_{\mathsf{E}} \tilde{\mathcal{P}}(x',\epsilon|x,a)\{\alpha\xi_{(x,a,\varepsilon,x',q,q')}[\gamma v(x') - \gamma v'(x') - q(x,a) + q'(x,a)]\\
&+ (q(x,a) - q'(x,a))\}\\
=&\alpha\gamma \sum_{x'} \int_{\mathsf{E}} \tilde{\mathcal{P}}(x',\epsilon|x,a)\xi_{(x,a,\varepsilon,x',q,q')}[v(x') - v'(x')]\\
&+ (1 - \alpha \sum_{x'} \int_{\mathsf{E}} \tilde{\mathcal{P}}(x',\epsilon|x,a)\xi_{(x,a,\varepsilon,x',q,q')})[q(x,a) - q'(x,a)]\\
\leq&\left(1 - \alpha(1-\gamma) \sum_{x'} \int_{\mathsf{E}} \tilde{\mathcal{P}}(x',\epsilon|x,a)\xi_{(x,a,\varepsilon,x',q,q')}\right) \|q - q'\|_\infty\\
\leq&(1 - \alpha(1-\gamma)\epsilon)\, \|q - q'\|_\infty
\end{aligned}$$

Hence, $\bar{\alpha} = 1 - \alpha(1-\gamma)\epsilon$ is the required constant. □

Proof of Theorem 6.11. Obviously, Condition (a) in Proposition 6.13 is satisfied and Condition (c) holds also due to Lemma 6.14. It remains to check Condition (b). Let

$$\begin{aligned}
D :=& r(x,a,\epsilon) + \gamma \max_{a'\in \mathsf{A}(x')} q(x',a') - q(x,a)\\
R :=& r(x,a,\epsilon).
\end{aligned}$$

Since u is Lipschitz, there exist coefficients $k_{x,a,x',\varepsilon} \in [\epsilon, L]$ such that

$$u(D) - u(R) = k_{x,a,x',\varepsilon}\,(\gamma v(x') - q(x,a)) = k_{x,a,x',\varepsilon}\,(D - R)$$

Hence, by

$$|\gamma \max_{a'\in \mathsf{A}(x')} q(x',a') - q(x,a)| \leq (1+\gamma)\|q\|_\infty,$$

we have

$$\begin{aligned}
u^2(D) =&(u(R) + k_{x,a,x',\varepsilon}\,(D - R))^2\\
\leq&2\left(u^2(R) + k^2_{x,a,x',\varepsilon}\,(D - R)^2\right)\\
\leq&2u^2(R) + 2k^2_{x,a,x',\varepsilon}(1+\gamma)^2\|q\|^2_\infty. \qquad (6.18)
\end{aligned}$$

We have then

$$\begin{aligned}
\mathbb{E}[w_t^2|\mathscr{F}_t] =& \alpha^2\mathbb{E}_{x,a}\left[u^2\left(r(x,a,\varepsilon)+\gamma\max_{a'\in\mathsf{A}(x')}q_t(x',a')-q_t(x,a)\right)\right] \\
&-\alpha^2\left(\mathbb{E}_{x,a}\left[u\left(r(x,a,\varepsilon)+\gamma\max_{a'\in\mathsf{A}(x')}q_t(x',a')-q_t(x,a)\right)\right]\right)^2 \\
\leq& \alpha^2\mathbb{E}_{x,a}\left[u^2\left(r(x,a,\varepsilon)+\gamma\max_{a'\in\mathsf{A}(x')}q_t(x',a')-q_t(x,a)\right)\right] \\
\text{(by (6.18))}\ \leq& 2\alpha^2\mathbb{E}_{x,a}\left[u^2\left(r(x,a,\epsilon)\right)\right]+(1+\gamma)^2\|q\|_\infty^2 2\alpha^2\mathbb{E}_{x,a}\left[k^2(x,a,x',\epsilon)\right] \\
\leq& 2\alpha^2\mathbb{E}_{x,a}\left[u^2\left(r(x,a,\epsilon)\right)\right]+2(1+\gamma)^2\|q\|_\infty^2,
\end{aligned}$$

where the last inequality is due to the fact that $k(x,a,x',\epsilon) \leq L$ and $\alpha \leq L^{-1}$. Hence, Condition (b) holds. □

6.3.5 Heuristics for utility functions and policies

In Assumption 6.8, the utility function u is required to be globally Lipschitz and furthermore, the slope is lower bounded away from zero. This strict requirement will restrict the application of several important types of utility functions. For instance, $u(x) = x^p, p \in (0,1), x \geq 0$, which is not Lipschitz at the area close to 0. We suggest two types of approximation to avoid this problem.

1) (re-centralization) Approximate u by $u^\varphi(x) = (x+\varphi)^p - \varphi^p$ with some positive φ.

2) (linearization) approximate u close to 0 by a linear function, i.e.

$$u^\varphi(x) = \begin{cases} u(x) & x \geq \varphi \\ \frac{xu(\varphi)}{\varphi} & x \in [0,\varphi) \end{cases}.$$

In both cases, φ should be set very close to 0.

For $u(x) = x^p, p > 1, x \geq 0$, it violates both bounds. At the area $[0,\varphi]$, where φ is very close to 0, we can again apply above two approximation schemes to overcome the problem by selecting small φ. At the area $[\varphi,\infty)$ with significantly large $\varphi \gg 0$, we consider the following linearization:

$$u^\varphi(x) = u(\varphi) + u'(\varphi)(x-\varphi), x \in [\varphi,\infty).$$

In Section 8.4, for both $p > 1$ and $p \in (0,1)$, we apply the linearization scheme to ensure Assumption 6.8.

Softmax policy

Recall that we call a policy is proper, if under such policy every state is visited infinitely often (see Definition 6.12). In this subsection, we show that under some technical assumptions the softmax policy (6.14) is proper.

Note that a policy $\boldsymbol{\pi} = [\pi_0, \pi_1, \ldots]$ is *deterministic* if for all state x and t, there exists an action $a \in \mathsf{A}(x)$ such that $\pi_t(a|x) = 1$. Under one policy $\boldsymbol{\pi}$, the n-step transition probability $P^{\boldsymbol{\pi}}(X_n = x'|X_0 = x)$ for some $x, x' \in \mathsf{X}$ can be calculated as follows

$$\begin{aligned} & P^{\boldsymbol{\pi}}(X_n = x'|X_0 = x) \\ = & \sum_{X_1, X_2, \ldots, X_{n-1}} P^{\pi_0}(X_1|x) P^{\pi_1}(X_2|X_1) \ldots P^{\pi_{n-1}}(x'|X_{n-1}) \end{aligned}$$

where $P^{\pi}(y|x) := \sum_a \mathcal{P}(y|x,a)\pi(a|x)$ and $\mathcal{P}$ is the transition kernel of the underlying MDP.

Proposition **6.15.** *Assume that the state and action space are finite and the assumptions required by Theorem 6.11 hold. Assume further that for each $x, x' \in \mathsf{X}$, there exist a deterministic policy $\boldsymbol{\pi}_d$, $n \in \mathbb{N}$ and a positive $\epsilon > 0$ such that $P^{\boldsymbol{\pi}_d}(X_n = x'|X_0 = x) > \epsilon$. Then the softmax policy stated in (6.14) is proper.*

Proof. Due to the contraction property of q (see Lemma 6.14), $\{q_t\}$ is uniformly bounded w.r.t. t. Let

$$\boldsymbol{\pi}_s = [\pi_0, \pi_1, \ldots]$$

be a softmax policy associated with $\{q_t\}$. Then, by the definition of softmax policies (see Eq. (6.14)), there exists a positive $\epsilon_0 > 0$ such that $\pi_t(a|x) \geq \epsilon_0$ holds for each $(x,a) \in \mathsf{K}$ and $t \in \mathbb{N}$. It implies that for each $x, x' \in \mathsf{X}$,

$$P^{\boldsymbol{\pi}_s}(X_n = x'|X_0 = x) \geq \epsilon_0^n P^{\boldsymbol{\pi}_d}(X_n = x'|X_0 = x),$$

for any deterministic policy $\boldsymbol{\pi}_d$. Then by the assumption of this proposition, we obtain that for each $x, x' \in \mathsf{X}$, $P^{\boldsymbol{\pi}_s}(X_n = x'|X_0 = x) \geq \epsilon_0^n \epsilon > 0$. It implies that each state will be visited infinitely often. □

The MDP applied in the behavioral experiment in Section 8.4 satisfies above assumptions, since for each $x, x' \in \mathsf{X}$, there exists a deterministic policy $\boldsymbol{\pi}_d$ such that $P^{\boldsymbol{\pi}_d}(S_n = x'|X_0 = x) = 1$, $n \leq 4$, no matter which initial state x we start with.

6.4 Discussion

Some technical extensions are possible within our general risk-sensitive reinforcement learning (RL) framework:

- To derive the risk-sensitive Q-learning, we consider here only the utility-based shortfall. Other types of valuation functions (maps), e.g., the optimized certainty equivalence introduced in Section 2.4.6 and the mean-semideviation trade-off in Section 2.4.7, can also derive Q-learning type algorithms.

- The Q-learning algorithm derived in this paper can be regarded a special type of RL algorithms, TD(0). It can be extended to other types of RL algorithms like SARSA (see e.g. Sutton and Barto, 1998, Chapter 6 for classical MDPs) and TD(λ)(see e.g. Sutton and Barto, 1998, Chapter 7 and Bertsekas and Tsitsiklis, 1996, Chapter 5 for classical MDPs) for $\lambda \neq 0$.
- In Chapter 5, we also provided a framework for the average case. Hence, RL algorithms for the average case can also be derived similar to the discounted case considered in this paper.
- The algorithm in its current form applies to MDPs with finite state spaces only. It can be extended for MDPs with continuous state spaces by applying function approximation technique (see e.g. Böhmer et al., 2013, Bertsekas and Tsitsiklis, 1996, Chapter 6 and Powell, 2007 for classical MDPs).

7

Human Decision under Uncertainty

Truly man is a marvelously vain, diverse, and undulating subject.
It is hard to find any constant and uniform judgment on him.
— Michel de Montaigne, *Essais, Book I*

Précis In this chapter, as a proof of principle for the applicability of the risk-sensitive Q-learning algorithms developed in the last chapter, we apply them to quantify human behavior in a sequential investment task. We find, that the risk-sensitive variant provides a significantly better fit to the behavioral data and that it leads to an interpretation of the subject's responses which is indeed consistent with prospect theory. The analysis of simultaneously measured fMRI signals show a significant correlation of the risk-sensitive TD error with BOLD signal change in the ventral striatum. In addition we find a significant correlation of the risk-sensitive Q-values with neural activity in the striatum, cingulate cortex and insula, which is not present if standard Q-values are used.

Publication related to this chapter The main results have been published in Shen et al., 2014d, Section 4. Among them, the analysis of fMRI results in Section 7.4, along with the supplementary material included at the end of this chapter, was done by Michael J. Tobia.

7.1 Introduction

Reinforcement learning (RL) has been widely applied to quantify sequential human decision making procedures, because similar computational structures, such as dopaminergically mediated reward prediction errors, have been identified in human brains (for a review see e.g. Dayan and Niv, 2008). As we have explained in the last chapter, the objective of the standard RL algorithms is to maximize expected cumulative rewards, which corresponds to risk-neutral behaviors only. However, in our daily life, decisions are usually made in the face of uncertain consequences. Hence, risk derived from these uncertainties has to be taken into

account by human decision makers, consciously or unconsciously. Human agents are not always economically rational. Apparently, some gamblers are risk seeking. Moreover, behavioral studies show that human can even be risk-seeking in one situation while risk-averse in another situation (Kahneman and Tversky, 1979). RL algorithms developed so far cannot effectively model these complicated risk-preferences.

In the literature of cognitive neuroscience, risk sensitive decision making problems have been widely investigated, especially based on prospect theory (see e.g. Berns et al., 2008; Hsu et al., 2005, 2009; Wu et al., 2009; Preuschoff et al., 2008 and references therein). However, most of the studies focuses on one-step decision making problems with full information, i.e., the human subjects were told the outcome and their associated probabilities. Recently, Nagengast et al. (2010), Braun et al. (2011) and Niv et al. (2012) considered sequence decision making problems in learning tasks. Among them, Nagengast et al. (2010); Braun et al. (2011) applied the entropic map, while Niv et al. (2012) applied the reinforcement learning algorithms developed by Mihatsch and Neuneier (2002) which is in fact a special case of utility-based shortfall (for details see Section 2.4.5). All of them can only model uniform risk preferences.

In the last chapter, we have developed a set of risk-sensitive Q-learning algorithms, by applying a family of valuation functions, the utility-based shortfall, to the general framework of risk-sensitive Markov decision processes. We have also shown in Section 2.4.5 that the key features predicted by prospect theory (Kahneman and Tversky, 1979) can be replicated by applying S-shape utility functions. Hence, the new risk-sensitive Q-learning algorithm provides a good candidate model for human risk-sensitive sequential decision-making procedures in learning tasks, where mixed risk-preferences are shown in behavioral studies. In this chapter, we will apply the learning algorithms to quantify human behaviors in a sequential investment task, accompanied with an analysis of simultaneously measured fMRI signals.

7.2 Experiment

Subjects were told that they are influential stock brokers, whose task is to invest into a fictive stock market (see also Tobia et al., 2014). At every trial (see Figure 7.2) subjects had to decide how much ($a = 0, 1, 2,$ or 3 EUR) to invest into a particular stock. After the investment, subjects first saw the change of the stock price and then were informed how much money they earned or lost. The received reward was proportional to the investment. The different trials, however, were not independent from each other (see Figure 7.2). The sequential investment game consisted of 7 states, each one coming with a different set of contingencies, and subjects were transferred from one state to the next dependent of the amount of money they invested. For high investments, transitions followed the path labeled "risk seeking" (RS in Figure 7.2). For low investments, transitions followed the

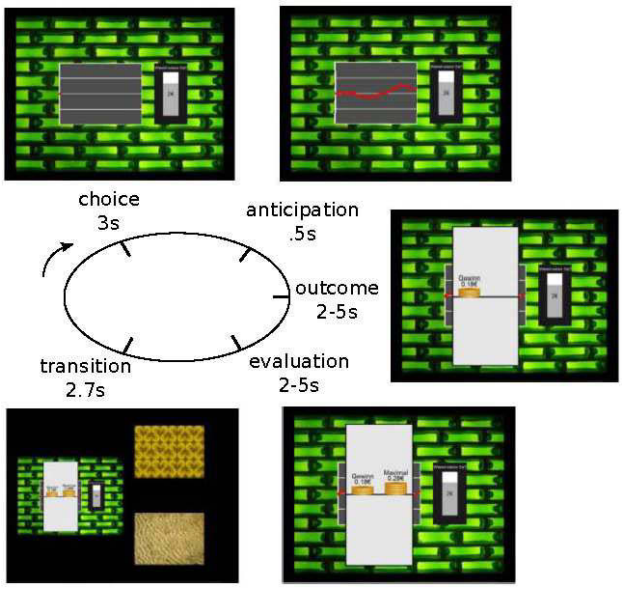


Figure 7.1: Phase transition. Every decision (trial) consists of a choice phase (3s), during which an action (invest 0, 1, 2, or 3 EUR) must be taken by adjusting the scale bar on the screen, an anticipation phase (.5s), an outcome phase (2–5s), where the development of the stock price and the reward (wins and loses) are revealed, an evaluation phase (2–5s), where it reveals the maximal possible reward that could have been obtained for the (in hindsight) best possible action, and a transition phase (2.7s), where subjects are informed about the possible successor states and the specific transition, which will occur. The intervals of the outcome and evaluation phase are jittered for improved fMRI analysis. State information is provided by the colored patterns, the black field provides stock price information during anticipation phase, and the white field provides the reward and the maximal possible reward of this trial. After each round (3 trials), the total reward of this round is shown to subjects.

path labeled "risk averse" (RA in Figure 7.2). After 3 decisions subjects were always transferred back to the initial state, and the reward, which was accumulated during this round, was shown. State information was available to the subjects throughout every trial (see Figure 7.2). Altogether, 30 subjects (young healthy adults) experienced 80 rounds of the 3-decision sequence.

Formally, the sequential investment game can be considered as an MDP with 7 states and 4 actions (see Figure 7.2). Depending on the strategy of the subjects, there are 4 possible paths, each of which is composed of 3 states. The total expected return for each path, averaged over all policies consistent with it, are shown in the right panels of Figure 7.2 ("EV"). On the hand, Path 1 provides the largest expected return per round (EV = 90), while Path 4 leads to an average loss of -9.75. Hence, to follow the on-average highest rewarded path 1, subjects have

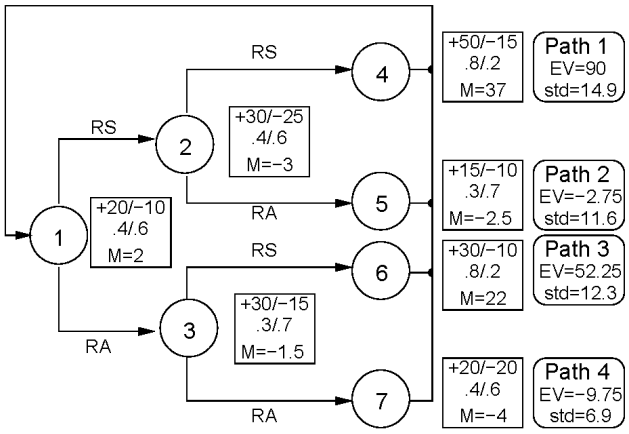


Figure 7.2: Structure of the underlying Markov decision process. The 7 states are indicated by numbered circles; arrows denote the possible transitions. Lables "RS" and "RA" indicate the transitions caused by the two "risk-seeking" (investment of 2 or 3 EUR) and the two "risk-averse" (investment of 0 or 1 EUR) actions. Bi-Gaussian distributions with a standard deviation of 5 are used to generate the random price changes of the stocks. Panels next to the states provide information about the means (top row) and the probabilities (center row) of ever component. M (bottom row) denotes the mean price change. The reward received equals the price change multiplied by the amount of money the subject invests. The rightmost panels provide the total expected rewards (EV) and the standard deviations (std) for all possible state sequences (Path 1 to Path 4) under the assumption that every sequence of actions consistent with a particular sequence of states is chosen with equal probability.

to take "risky" actions (investing 2 or 3 EUR at each state). Always taking conservative actions (investing 0 or 1 EUR) results in Path 4 and a high on-average loss. On the other hand, since the standard deviation of the return R of each state equals $\mathrm{std}(R) = a \times \mathrm{C}$, where a denotes the action (investment) the subject takes and C denotes the price change, the higher the investment, the higher the risk. Path 1 has, therefore, the highest standard deviation (std = 14.9) of the total average reward, whereas the standard deviation of Path 4 is smallest (std = 6.9). Path 3 provides a trade-off option: it has slightly lower expected value (EV = 52.25) than Path 1 but comes with a lower risk (std = 12.3). Hence, the paradigm is suitable for observing and quantifying the risk-sensitive behavior of subjects.

7.3 Risk-sensitive model of human behavior

Figure 7.3 summarizes the strategies which were chosen by the 30 subjects. 17 subjects mainly chose Path 1, which provided them high rewards. 6 subjects chose Path 4, which gave very low rewards. The remaining 7 subjects show no signifi-

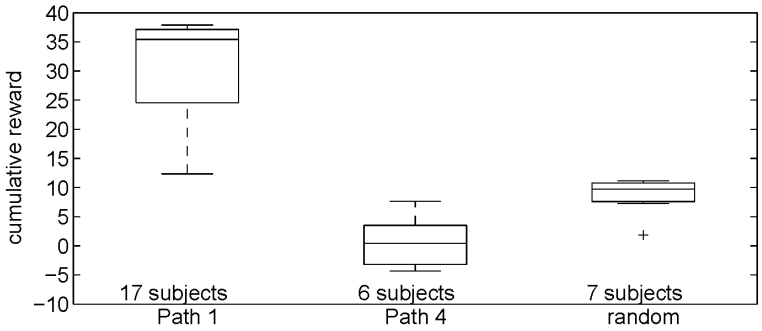


Figure 7.3: Distribution of "strategies" chosen by the subjects in the sequential investment game and the corresponding cumulative rewards. Subjects are grouped according to the sequence of states (Path 1 to Path 4, cf. Figure 7.2) they chose during the last 60 trials of the game. If a path i is chosen in more than 60% of the trials, the subject is assigned the group "Path i". Otherwise, subjects are assigned the group labeled "random". The vertical axis denotes the cumulative reward obtained during the last 60 trials.

cant preference among all 4 paths and the rewards they received are on average between the rewards received by the other 2 groups. The optimal policy for maximizing expected reward is the policy that follows Path 1. The results shown in Figure 7.3, however, indicate that the standard model fails to explain the behavior of more than 40% of the subjects.

We now quantify subjects' behavior by applying three classes of Q-learning algorithm:

(1) standard Q-learning,

(2) the risk-sensitive Q-learning (RSQL) method described by Algorithm 6.2 in Section 6.3.3, and

(3) an expected utility (EU) algorithm with the following update rule

$$Q(s_t, a_t) \Leftarrow Q(s_t, a_t) + \alpha \left(u(r_t) + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right), \tag{7.1}$$

where the nonlinear transformation is applied to the reward r_t directly. The latter one is a straightforward extension of expected utility theory. Risk sensitivity is implemented via the nonlinear transformation of the true reward r_t.

For both risk-sensitive Q-learning methods (RSQL and EU), we consider the family of polynomial mixed utility functions

$$u(x) = \begin{cases} k_+ x^{l_+} & x \geq 0 \\ -k_- (-x)^{l_-} & x < 0 \end{cases} . \tag{7.2}$$

branch $x \geq 0$	shape	risk pref.
$0 < l_+ < 1$	concave	risk-averse
$l_+ = 1$	linear	risk-neutral
$l_+ > 1$	convex	risk-seeking

branch $x < 0$	shape	risk pref.
$0 < l_- < 1$	convex	risk-seeking
$l_- = 1$	linear	risk-neutral
$l_- > 1$	concave	risk-averse

Table 7.1: Parameters for the two branches $x \geq 0$ (left) and $x < 0$ (right) of the polynomial utility function $u(x)$ (7.2), its shape and the induced risk preference.

The parameters $k_\pm > 0$ and $l_\pm > 0$ quantify the risk-preferences separately for wins and losses (see Table 7.1). Hence, there are 4 parameters for u which have to be determined from the data. For all three classes, actions are generated according to the "softmax" policy (6.14), which is a proper policy for the paradigm (for a proof see Section 6.3.5), and the learning rate α is set constant across trials.

For RSQL, the learning rate is absorbed by the coefficients $k_\pm$. Hence, there are 6 parameters

$$\{\beta, \gamma, k_\pm, l_\pm\} =: \theta$$

which have to be determined. Standard Q-learning corresponds to the choice $l_\pm = 1$ and $k_\pm = \alpha$. The risk-sensitive model applied by Niv et al. (2012) is also a special case of the RSQL-framework and corresponds $l_\pm = 1$. For the EU algorithm, there are 7 parameters, $\{\alpha, \beta, \gamma, k_\pm, l_\pm\}$, which have to be fitted to the data. $l_\pm = 1$ and $k_\pm = 1$ again corresponds to the standard Q-learning method.

Parameters were determined subject-wise by maximizing the log-likelihood of the subjects' action sequences,

$$\max_\theta L(\theta) := \sum_{t=1}^{T} \log p(a_t|s_t, \theta) = \sum_{t=1}^{T} \log \frac{e^{\beta Q(s_t, a_t|\theta)}}{\sum_a e^{\beta Q(s_t, a|\theta)}} \tag{7.3}$$

where $Q(s, a|\theta)$ indicates the dependence of the Q-values on the model parameters θ. Since RSQL/EU and the standard Q-learning are nested model classes, we apply the Bayesian information criterion (BIC, see e.g., Ghosh et al., 2006)

$$B := -2L + k \log(n)$$

for model selection. L denotes the log-likelihood (7.3). k and n are the number of parameters and trials respectively.

To compare results, we report relative BIC scores, $\Delta B := B - B_Q$, where B is the BIC score of the candidate model and B_Q is the BIC score of the standard Q-learning model. We obtain

$$\begin{aligned} \Delta B = &\quad -500.14 \quad \text{for RSQL, and} \\ \Delta B = &\quad -23.10 \quad \text{for EU.} \end{aligned}$$

The more negative the relative BIC score is, the better the model fits data. Hence, the RSQL algorithm provides a significantly better explanation for the behavioral

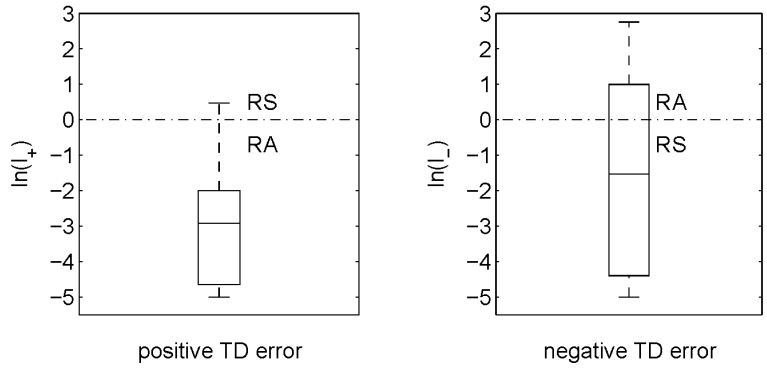


Figure 7.4: Distribution of values for the shape parameters l_+ (left) and l_- (right) for the RSQL model.

data than the EU algorithm and standard Q-learning. In the following, we only discuss the results obtained with the RSQL model.

Figure 7.4 shows the distribution of best-fitting values for the two parameters $l_\pm$ which quantify the risk-preferences of the individual subjects. We conclude (cf. Table 7.1) that most of the subjects are risk-averse for positive and risk-seeking for negative TD errors. The result is consistent with previous studies from the economics literature (see Tversky and Kahneman, 1992 and references therein).

After determining the parameters $\{k_\pm, l_\pm\}$ for the utility functions, we perform an analysis similar to the analysis discussed in Section 2.4.5. Given an observed reward sequence $\{r_i\}_{i=1}^N$, the empirical subjective mean m_{sub} is obtained by solving the following equation

$$\frac{1}{N}\sum_{i=1}^{N} u(r_i - m_{sub}) = 0.$$

If subjects are risk-neutral, then $u(x) = x$, and $m_{sub} = m_{emp} = \frac{1}{N}\sum_{i=1}^N r_i$ is simply the empirical mean. Following the idea of prospect theory, we define a normalized subjective probability Δp,

$$\Delta p := \frac{m_{sub} - \min_i r_i}{\max_i r_i - \min_i r_i} - \frac{m_{emp} - \min_i r_i}{\max_i r_i - \min_i r_i} = \frac{m_{sub} - m_{emp}}{\max_i r_i - \min_i r_i}. \tag{7.4}$$

If Δp is positive, the probability of rewards is overestimated and the induced policy is, therefore, risk-seeking. If Δp is negative, the probability of rewards is underestimated and the policy is risk-averse. Figure 7.5 summarizes the distribution of normalized subjective probabilities for subjects from the "Path 1", "Path 4" and "random" groups of Figure 7.3. For subjects within group "Path 1", $|\Delta p|$ is small and their behaviors are similar to those of risk-neutral agents. This is consistent with their policy, because both risk-seeking and risk-neutral agents should prefer Path 1. For subjects within groups "Path 4" and "random", the normalized subjective probabilities are on average 10% lower than those of risk-neutral agents. This

explains why subjects in these groups adopt the conservative policies and only infrequently choose Path 1.

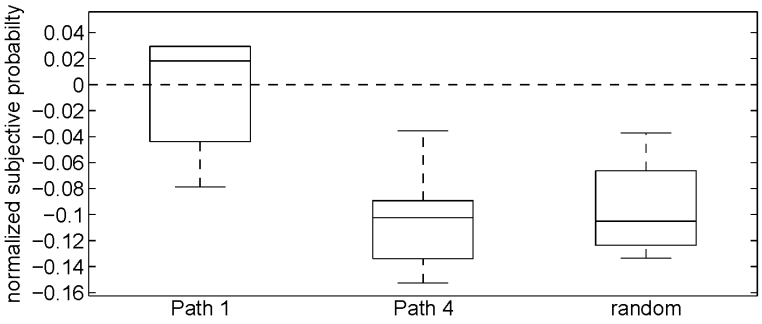


Figure 7.5: Distribution of normalized subjective probabilities, Δp, defined in (7.4), for the different subject groups defined in Figure 7.3.

7.4 fMRI results

Functional magnetic resonance imaging (fMRI) data were simultaneously recorded while subjects played the sequential investment game. The analysis of fMRI data was conducted in SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Details of the magnetic resonance protocol and data processing are presented at the end of this chapter. The sequence of Q-values for the action chosen at each state were used as parametric modulators during the choice phase, and temporal difference (TD) errors were used at the outcome phase (see Figure 7.2).

Figure 7.6 (left) shows that the sequence of TD errors for the RSQL model (with best fitting parameters) positively modulated the BOLD signal in the subcallosal gyrus extending into the ventral striatum (-14 8 -16) (marked by the cross in Figure 7.6 (left)), the anterior cingulate cortex (8 48 6), and the visual cortex (-8 -92 16; $z = 7.9$). The modulation of the BOLD signal in the ventral striatum is consistent with previous experimental findings (cf. Schultz, 2002; O'Doherty, 2004), and supports the primary assertion of computational models that reward-based learning occurs when expectations (here, expectations of "subjective" quantities) are violated (Sutton and Barto, 1998).

Figure 7.6 (right) shows the results for the sequence of Q-values for the RSQL model (with best fitting parameters), which correspond to the subjective (risk sensitive) expected value of the reward for each discrete choice. Several large clusters of voxels in cortical and subcortical structures were significantly modulated by the Q-values at the moment of choice. The sign of this modulation was negative. The peak of this negative modulation occurred in the left anterior insula (-36 12 -2, $z = 4.6$), with strong modulation also in the bilateral ventral striatum (14 8 -4, marked by the cross in Figure 7.6(right); -16 4 0) and the cingulate cortex (4 16 28). The reward prediction error processed by the ventral striatum (and other regions

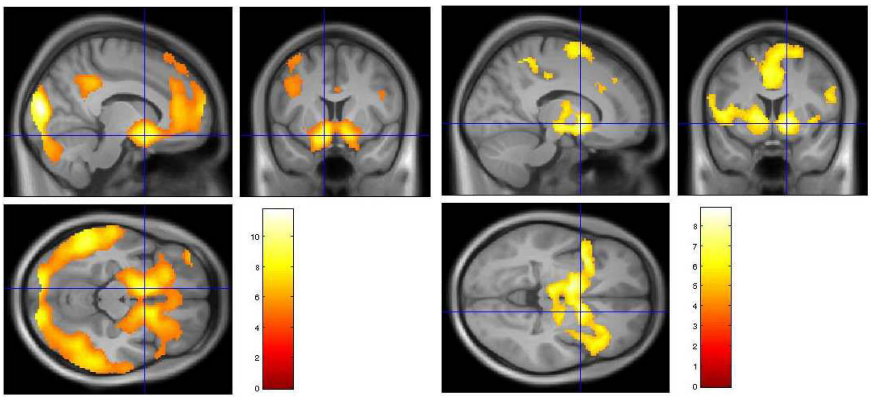


Figure 7.6: Modulation of the fMRI BOLD signal by (left) TD errors and by (right) Q-values generated by the RSQL model with best fitting parameters. The data is shown whole-brain corrected to $p < .05$ (voxel-wise $p < .001$ and minimum 125 voxels). The color bar indicates the t-value ranging from 0 to the maximal value. The cross indicates location of strongest modulation for TD errors (in the left ventral striatum (-14 8 -16)) and for Q-values (in the right ventral striatum (14 8 -4)). However, it is remarkable that for both TD errors and Q-values, modulations in the left and right ventral striatum are almost equally strong with a slight difference.

noted above) would not be computable in the absence of an expectation, and as such, this activation is important for substantiating the plausibility for the RSQL model of learning and choice. Sequences of Q-values obtained with standard Q-learning (with best fitting parameters), on the other hand, failed to predict any changes in brain activity even at a liberal statistical threshold of $p < .01$ (uncorrected). This lack of neural activity for the standard Q model, in combination with the significant activation for our RSQL, supports the hypothesis that some assessment of risk is induced and influences valuation. Whereas the areas modulated by Q-values differ from what has been reported in other studies (i.e., the ventromedial prefrontal cortex as in Gläscher et al., 2009), it does overlap with the representation of TD errors. Furthermore, the opposing signs of the correlated neural activity suggests that a neural mismatch of signals in the ventral striatum between Q-value and TD errors may underlie the mechanism by which values are learned.

7.5 Discussion

We applied the risk-sensitive Q-learning (RSQL) method to quantify human behavior in a sequential investment game and investigated the correlation of the predicted TD- and Q-values with the neural signals measured by fMRI.

We first showed that the standard Q-learning algorithm cannot explain the

behavior of a large number of subjects in the task. Applying RSQL generated a significantly better fit and also outperformed the expected utility algorithm. The risk sensitivity revealed by the best fitting parameters is consistent with the studies in behavioral economics, that is, subjects are risk-averse for positive while risk-seeking for negative TD errors. Finally, the relative subjective probabilities provide a good explanation why some subjects take conservative policies: they underestimate the true probabilities of gaining rewards.

The fMRI results showed that TD sequence generated by our model has a significant correlation with the activity in the subcallosal gyrus extending into the ventral striatum. The sequence of Q-values has a significant correlation with the activities in the left anterior insula. Previous studies (see e.g., Glimcher et al., 2008, Chapter 23 and Symmonds et al., 2011) suggest that higher order statistics of outcomes, e.g., variance (the 2nd order) and skewness (the 3rd order), are encoded in human brains separately and the individual integration of these risk metrics induces the corresponding risk sensitivity. Our results indicate, however, that the risk sensitivity can be simply induced (and therefore encoded) by a nonlinear transformation of TD errors and no additional neural representation of higher order statistics is needed.

Supplementary material: magnetic resonance protocol and data processing

Magnetic resonance (MR) images were acquired with a 3T whole-body MR system (Magnetom TIM Trio, Siemens Healthcare) using a 32-channel receive-only head coil. Structural MRI were acquired with a T1 weighted magnetization-prepared rapid gradient-echo (MPRAGE) sequence with a voxel resolution of $1 \times 1 \times 1$ mm^3, coronal orientation, phase-encoding in left-right direction, FoV = 192 × 256 mm, 240 slices, 1100 ms inversion time, TE = 2.98 ms, TR = 2300 ms, and 90 flip angle. Functional MRI time series were recorded using a T2* GRAPPA EPI sequence with TR = 2380 ms, TE = 25 ms, anterior-posterior phase encode, 40 slices acquired in descending (non- interleaved) axial plane with $2 \times 2 \times 2$ mm^3 voxels (204 × 204 mm FoV; skip factor = .5), with an acquisition time of approximately 8 minutes per scanning run.

Structural and functional magnetic resonance image analyzes were conducted in SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Anatomical images were segmented and transformed to Montreal Neurological Institute (MNI) standard space, and a group average T1 custom anatomical template image was generated using DARTEL. Functional images were corrected for slice-timing acquisition offsets, realigned and corrected for the interaction of motion and distortion using unwarp toolbox, co-registered to anatomical images and transformed to MNI space using DARTEL, and finally smoothed with an 8 mm FWHM isotropic Gaussian kernel.

Functional images were analyzed using the general linear model (GLM) imple-

mented in SPM8. First level analyzes included onset regressors for each stimulus event excluding the anticipation phase (see Figure 7.2), and a set of parametric modulators corresponding to trial-specific task outcome variables and computational model parameters. Trial-specific task outcome variables (and their corresponding stimulus event) include the choice value of the investment (choice phase) and the total value of rewards (gains/losses) over each round (corresponding to multi-trial feedback event). Model derived parametric modulators included the time series of Q values for the selected action (choice phase), TD (outcome phase). Reward value was not modeled as a parametric modulator because the TD error time series and trial-by-trial reward values were strongly correlated (all rs $>$ .7; ps $<$ .001). The configuration of the first-level GLM regressors for the standard Q-learning model was identical to that employed in the risk-sensitive Q-learning model. All regressors were convolved with a canonical hemodynamic response function. Prior to model estimation, coincident parametric modulators were serially orthogonalized as implemented in SPM (i.e., the Q-value regressor was orthogonalized with respect to the choice value regressor). In addition, we included a set of regressors for each participant to censor EPI images with large, head movement related spikes in the global mean. These first level beta values were averaged across participants and tested against zero with a t-test. Monte Carlo simulations determined that a cluster of more than 125 contiguous voxels with a single-voxel threshold of $p < .001$ achieved a corrected p-value of .05.

8

Risk-averse Algorithmic Trading

I of dice possess the science
and in numbers thus am skilled.
— Mahābhārata

Précis In this chapter, we apply the risk-averse reinforcement learning algorithms developed in Chapter 6 to algorithmic trading. Our approach is tested in an experiment based on 1.5 years of millisecond time-scale limit order data from NASDAQ, which contain the data around the 2010 flash crash. The results show that our algorithm outperforms the risk-neutral reinforcement learning algorithm by 1) keeping the trading cost at a substantially low level at the spot when the flash crash happened, and 2) significantly reducing the risk over the whole test period.

Publication related to this chapter This chapter is based on Shen et al. (2014a), presenting work done in collaboration with Ruihong Huang.

8.1 Introduction

Algorithmic trading contributes a major part of trading volumes to modern electronic equity markets (Securities and Exchange Commission, 2010). Though details of trading algorithms in practice remain unrevealed, solid evidence (see, e.g., Menkveld and Yueshen, 2013, Kirilenko et al., 2011) found around the *2010 Flash Crash*

> *"On May 6, 2010, the prices of many U.S.-based equity products experienced an extraordinarily rapid decline and recovery. That afternoon, major equity indices in both the futures and securities markets, each already down over 4% from their prior-day close, suddenly plummeted a further 5-6% in a matter of minutes before rebounding almost as quickly."* – *U.S. Commodity Futures Trading Commission and Securities & Exchange Commission (2010)*

indicates that most of them failed to adapt their trading strategies to the extreme market event. This failure harms not only performance of trading algorithms, but also market's stability. Hence, there is an increasing interest to incorporate a robust risk control into trading strategies, especially into the algorithms used for algorithmic trading.

In this chapter, we focus on a common high-frequency trade problem faced by an institutional trader: the *optimal trade execution* (see e.g. Bertsimas and Lo, 1998; Almgren and Chriss, 2001; Nevmyvaka et al., 2006 and references therein), i.e., to liquid a huge inventory over a short time horizon. We model this problem by a Markov decision processes with finite state and action spaces. In particular, in order to control the risk in trade, we apply a *risk-averse* valuation map within the framework of risk-sensitive Markov decision processes developed in Section 6.2. The corresponding optimization problem is then solved by a data-driven technique, the risk-averse reinforcement learning (RL), or more specifically, risk-averse Q-learning, proposed in Section 6.3.

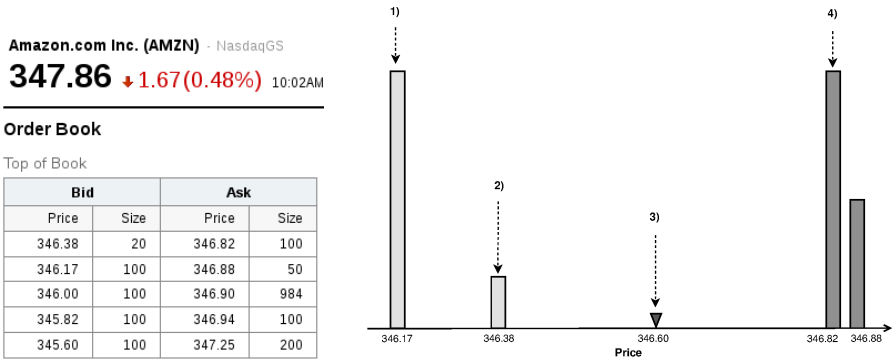
By employing a huge data set containing the high-frequency order books (one and half years in millisecond time scale) of Amazon.com Inc. (AMZN) traded on NASDAQ, we show that the risk-averse RL outperforms the standard RL (which is risk-neutral) by a reduction of 10 to 15 basis points on the risk, at a price of 2 to 3 basis points on the average avenue, depending on the experimental cases. More importantly, unlike the risk-neutral RL, which results in a huge cost on the flash-crash spot, the risk-averse RL is able to keep its corresponding cost at a substantially low level.

Our contribution to the literature on algorithmic trading are twofold. First, our risk-averse RL generalizes the risk-neutral RL (see e.g., Bertsimas and Lo, 1998; Nevmyvaka et al., 2006) by allowing traders to explicitly control the risk according to their risk sensitivities. Second, we introduce the RS-MDP model into algorithmic trading. Its basic idea on risk control can be valuable for traders to improve the robustness of their trading algorithms under extreme market conditions.

Our study is also related to the works by Almgren and Chriss (2001); Alfonsi et al. (2010) and Obizhaeva and Wang (2012), where explicit structural models are proposed to describe market dynamics. Based on these models, the risk in trade execution is controlled by minimizing quadratic utility or value at risk. Our approach is distinguished from them by applying an RL-type algorithm to solve the risk-averse optimization problem. Since RL is a data-driven technique, it requires no explicit structural model on the market dynamics. Hence, our approach is more practical in high-frequency markets where modeling the market dynamics is a huge challenge.

8.2 Trading on limit order markets

Institutional traders generally acquire their long-term (e.g. a day or longer) target positions by assigning small tasks in a short time-frame, typically in minutes. In



Amazon.com Inc. (AMZN) · NasdaqGS

347.86 ↓1.67(0.48%) 10:02AM

Order Book

Top of Book

Bid		Ask	
Price	Size	Price	Size
346.38	20	346.82	100
346.17	100	346.88	50
346.00	100	346.90	984
345.82	100	346.94	100
345.60	100	347.25	200

Figure 8.1: Left: a snapshot of AMZN order book in NASDAQ (source: Yahoo! Finance). Right: a graphical representation of the order book with two price levels. The position and height of each bar represent the price and size of each order respectively. Bids (buy orders) are drawn in green, while asks (sell orders) are in red. The mid-quote price, i.e., the average of the highest bid and lowest ask, is marked by a triangle. 1) – 4) label 4 admissible strategies for a seller (for details see Section 8.2).

most electronic limit order markets, e.g., NASDAQ, traders compete for trading by posting limit orders. A limit order is specified by its trade direction (buy or sell), limit price and quantity. During the continuous trading phase, a centralized computer system collects all incoming limit orders and executes them according to the following price-time procedure rule: the buy (respectively sell) order with a higher (respectively lower) price is executed earlier and, among all orders with the same price, the first coming order is executed first. In the case where a buy (respectively sell) order with a price higher (respectively lower) than the best price of all existing sell (respectively buy) orders, it becomes a so-called market(able) order and is executed immediately. All unexecuted limit orders are aggregated as a limit order book which is publicly visible to all market participants.

To illustrate how trade is executed, suppose a trader plans to sell 100 shares of AMZN (Amazon.com Inc.). Figure 8.1 shows a snapshot of AMZN's limit order book in NASDAQ. Here are some feasible choices for her (for the sake of simplicity, we do not consider hidden liquidity in this example. Nevertheless, the data of hidden liquidity is considered and used in our trade simulation in Section 8.4.):

1. submit an order at price \$346.17 and get an immediate full execution with the average price \$346.212 (80 shares at \$346.17 and 20 at \$346.38).

2. submit an order at price \$346.38 and get a partial execution for 20 shares.

The remaining 80 shares will wait in the order book until incoming buyers take them.

3. submit an order at price \$346.60. Since it is the first order at this price, it will be executed (possible partially) when a buyer bids a price higher than or equal to \$346.60.

4. submit an order at price \$346.82. It will be executed when an incoming buyer's bid is higher than or equal to this price, but only after the execution of the existing order with 100 shares at the same price.

We see that a strategy pursuing a better selling price is always associated with a longer (expected) time-to-execution. In practice, the trader typically has to liquid a position inside a strict time-horizon. The long time-to-execution could be very costly for her, since she probably has to trade at a substantially bad price at the end of the time-horizon, especially when the market price moves downward.

To solve this problem, the trader can divide the whole time-horizon into several time points and individual decisions are made at each time point according to updates (or *states*, which will be specified in the next section) of the order book, in order to achieve the long-term goal. In the previous example, she can place her order (with the remaining open quantity) at a lower (respectively higher) price when observing a downward (respectively upward) movement of the price.

Since the market is usually unstable, the optimal strategy for the long-term goal in one time period, e.g., several days, might not work in another time period. A usual way to solve this problem is to define the long-term goal in the sense of average, i.e., we seek a strategy that maximizes the long-term goal averaged over all time periods. But even so, the derived optimal strategy might still cause huge losses in some extreme market conditions, e.g., the 2010 flash crash, since the extreme market conditions might be far away from the average.

In the high-frequency world, the averagely optimal strategy is usually obtained by *reinforcement learning* (RL, see e.g., Nevmyvaka et al., 2006). Since it does not require an explicit structural model for the underlying market dynamics, RL has been proven to be a very flexible and powerful tool for maximizing average rewards. In this chapter, we are going to avoid high losses at extreme market events by adding risk control into RL algorithms.

8.3 Risk-averse Q-learning

We restate the main results of risk-sensitive Markov decision processes and Q-learning developed in Chapter 6 in the context of optimal trade execution.

A *Markov decision process* (MDP, see e.g. Puterman, 1994)

$$\mathscr{M} := \{\mathsf{X}, \mathsf{A}, \mathcal{P}, (r, \mathcal{P}_r)\}. \tag{8.1}$$

consists of

- a state space X, where each state $s \in \mathsf{X}$ consists of a set of discretized variables representing the state of limit order book and the trade execution,
- an action space A composed of all admissible trading options (denoted by a),
- a transition kernel $\mathcal{P}(x'|x,a)$ specifying the transition probability moving from state $x \in \mathsf{X}$ to $x' \in \mathsf{X}$ after taking action a,
- a random reward R with the form

$$R(x,a) := r(x,a) + \epsilon(x,a),$$

where the determinant function $r : \mathsf{X} \times \mathsf{A} \to \mathbb{R}$ denotes the expected reward at each $(s,a) \in \mathsf{X} \times \mathsf{A}$. The stochastic function ϵ represents a reward noise due to the uncertain future dynamics of the market. Let $\mathcal{P}_r(\epsilon|s,a)$ denote the distribution of the noise with the event space E. We assume that the zero-mean condition

$$\int_{\mathsf{E}} \epsilon \mathcal{P}_r(\epsilon|x,a)d\epsilon = 0 \tag{8.2}$$

holds for all $(x,a) \in \mathsf{X} \times \mathsf{A}$.

There are two sources of uncertainty in our model: (i) the successive state x' (due to the stochastic dynamics of limit order book) and (ii) the immediate reward (due to noise ϵ). They can be quantified by the following joint distribution

$$\mu_{x,a}(x',\epsilon) := \mathcal{P}(x'|x,a)\mathcal{P}_r(\epsilon|x,a). \tag{8.3}$$

Let $t = 1,\ldots,T$ denote the tth time point in the discretized trading time-horizon and $\pi_t : \mathsf{X} \to \mathsf{A}$ be the corresponding decision rule. The optimal *risk-neutral* trading strategy is then the multistage policy $\boldsymbol{\pi} = [\pi_1, \pi_2, \ldots, \pi_T]$ that maximizes the following objective function (see e.g. Bertsimas and Lo, 1998):

$$\begin{aligned} J(\boldsymbol{\pi},x) :=& \mathbb{E}[\sum_{t=1}^{T} R(X_t,A_t)|X_1 = x, \boldsymbol{\pi}] \\ =& \mathbb{E}^{\pi_1}_{X_1=x}[R(X_1,A_1) + \mathbb{E}^{\pi_2}_{X_2}[R(X_2,A_2) + \ldots + \mathbb{E}^{\pi_T}_{X_T}[R(X_T,A_T)]\ldots]]. \end{aligned}$$

Note that comparing with the setup in Section 6.2, we considered here merely deterministic policies, due to the fact that one has to select a specific action at each trading time point.

To incorporate risk into the multistage decision-making procedure, we replace the risk-neutral expectation $\mathbb{E}$ by the utility-based shortfall, $\mathcal{U}$,

$$\mathcal{U}_{x,a}(X) = \sup\{m \in \mathbb{R} \mid \mathbb{E}_{x,a}[u(X-m)] \geq 0\}. \tag{8.4}$$

Here, u denotes a concave, continuous and strictly increasing utility function satisfying $u(0) = 0$. It can be chosen freely according to how much risk the trader is willing to take. We can, therefore, control the risk-sensitivity of our trading strategies by properly adjusting its form. The risk-averse optimal trade strategy is then obtained by maximizing the following *risk-averse objective*:

(8.5)
$$\tilde{J}(\boldsymbol{\pi}, x) := \mathcal{U}^{\pi_1}_{X_1=x}[R(X_1, A_1) + \mathcal{U}^{\pi_2}_{X_2}[R(X_2, A_2) + \ldots + \mathcal{U}^{\pi_T}_{X_T}[R(X_T, A_T)]\ldots]],$$

where $\mathcal{U}^{\pi}_{x}$ is defined by $\mathcal{U}^{\pi}_{x}(\cdot) := \mathcal{U}_{x,\pi(x)}(\cdot)$. In this chapter, we assume that u takes the following form,

$$u(x) = \begin{cases} \frac{1}{\lambda}[(x+1)^{\lambda} - 1] & x \geq 0 \\ x & x < 0 \end{cases}, \tag{8.6}$$

where $\lambda \in (0, 1]$. For $\lambda < 1$, u is concave and the profits (i.e., $x > 0$) are always less valued, while the losses ($x < 0$) keep equal. Since u is a concave function, by Proposition 2.16 and 3.9, its corresponding optimal trading strategy is therefore risk-averse. Furthermore, the smaller λ is, the more punishment are applied to large profits, which leads to a more risk-averse trading strategy. If setting $\lambda = 1$, it corresponds to the linear function $u(x) = x$ implying that the evaluation function coincides with the standard expectation, i.e., $\mathcal{U}_{s,a}(\cdot) = \mathbb{E}_{x,a}(\cdot)$. In this case, our model reduces to the risk-neutral model.

As we have shown in Section 6.3.2, the optimization problem can be solved by a reinforcement learning (RL) algorithm (see Algorithm 6.1). Suppose at tth time point, we repeat trying action a at state x a large amount of times and observe N samples of immediate rewards and successive states, $\{R_i, x'_i\}_{i=1,2,\ldots,N}$. Then the q-value at (x, a), which evaluates the quality of this state-action pair, can be estimated by the following iterative procedure

$$q_t^{(i+1)}(x, a) = q_t^{(i)}(x, a) + \frac{1}{i} u\left(R_i + \max_a q_{t+1}(x'_i, a) - q_t^{(i)}(x, a)\right) \tag{8.7}$$

provided that we have already known the q-value at $(t+1)$th time point at each state-action. Since the utility function u of the form (8.6) is sufficiently regular (in fact, it is easy to verify that u is Lipschitz for all $\lambda \in (0, 1]$), by the standard result in stochastic approximation (see e.g. Borkar, 2008, Theorem 2), for each $t = T, T-1, \ldots, 1$, the policy $\pi_t^{(N)}(x) = \max_{a \in A} q_t^{(N)}(x, a)$ converges to the optimal policy that maximizes the risk-averse objective function defined in (8.5), as $N \to \infty$.

8.4 Experiments

Data

We perform our experiment on selling AMZN stocks in NASDAQ. The order book data are provided by LOBSTER (Huang and Polak, 2011)[1] with two price levels,

[1] For more information see the website http://lobsterdata.com.

i.e., only the best two asks and bids (see Figure 8.1 (bottom) for an example) are available at each time stamp. LOBSTER uses

1. historical NASDAQ TotalView streaming messages, which are provided by NASDAQ as a standard data feed for real-time trading, and

2. an algorithm, which continuously updates order books by all order events, including order submissions, cancellations and executions, as well as hidden order executions, recorded with millisecond time stamps.

Therefore, our experimental environment effectively represents the real algorithmic trading environment in NASDAQ.

In order to analyze our model's performance during and after the flash crash on May 6, 2010, we use the one year data previous to the flash crash, i.e., from May 1, 2009 to April 30, 2010, to train the model and obtain the best strategy. This strategy is then tested in the succeeding half-year period from May 1, 2010 to October 31, 2010, which contains the flash crash.

MDP formulation

In our experiment, we analyze the performance of the risk-averse RL by comparing to the risk-neutral RL proposed by Nevmyvaka et al. (2006) who show that it substantially outperforms the trading strategy, *submit and leave*, which is commonly used in practice. For comparison, we specify the structure of the underlying MDP defined in (8.1) by applying a setup similar to the one used in Nevmyvaka et al. (2006).

Time resolution We discretize the total time horizon equally with different scales. Given the total time horizon H and a time resolution T, we assume that limit orders will be submitted to the market at each time point $t = n \cdot H/T$, $n = 0, 1, \ldots, T-1$, according to the trading policy determined by our algorithm. In this study, for covering the "flash crash" period, we set $H = 10$ minutes for all experimental cases.

States We consider two variables: the seller's inventory and the bid-ask spread of the order book.Let V denote the target volume and I be the number of inventory units, between which one strategy can distinguish. Then, if v shares are remained, the state of inventory is $i := \lfloor v \cdot I/V \rfloor$. The spread is discretized to three states: small, middle and big, according to the 33.3% to 66.7% empirical quantiles of the spread in training data set.

Actions We use relative prices as actions. Specifically, action a corresponds to submitting a sell order at price $ask - a$ (unit: US cent) with all of the remaining

shares[2]. Hence, $a \leq 0$ represents a strategy of submitting a limit order at or behind the market, e.g., strategy 4) in Sec. 8.2, while $a > 0$ corresponding to strategies undercutting ask. In the case that a is greater than the spread, it is the strategy of submitting market(able) order, e.g., strategy 1) in Section 8.2. The range of actions are determined based on gaps between second best asks and second best bids in historical data.

Rewards The immediate reward of an action a is the cash inflow resulted from a (partial) execution of the limit order placed at $ask - a$ in the next H/T time interval. Moreover, we assume that selling V shares is mandatory. Any inventory remaining at time H must be cleaned up by using a market order, walking through the lower prices on the bid side of the order book until all remaining volumes are sold. Note that this execution is not a part of policy, but serves as a penalty for the failure of fulfilling the trade target.

Performance evaluation

We evaluate performance of algorithms by calculating *trading costs* and their associated *risk* in the test period using the optimal strategies learned by algorithms.

In particular, the trading cost of each time horizon H is defined as the average execution price achieved by the strategy relative to the mid-quote (average of bid and ask) at the start time point of H, i.e.,

$$cost = \frac{\text{mid-quote at time 0} - \text{average execution price}}{\text{mid-quote at time 0}} \times 10000.$$

Risk is evaluated by two criteria:

1) the standard deviation and

2) 95% quantile of costs over the whole test period.

The latter one is in spirit of the concept of *value at risk* (see e.g. Jorion, 2007), which is widely used in industry to measure risk. By the 95% quantile cost, we measure in fact the upside risk of extremely large losses (costs).

The unit of both costs and their associated risk, including the standard deviation and 95% quantile, is simply the basis point (= 1/100 of a percent).

8.5 Results

Tuning λ

The free parameter λ in (8.6) determines how conservative (risk-averse) the derived optimal strategy is with regard to the uncertainty of rewards. How to choose

[2]Our model can be extended to include the order size as another dimension of the action space, given an estimate for the market impact of limit order (see e.g., Hautsch and Huang, 2012), which is left as future work.

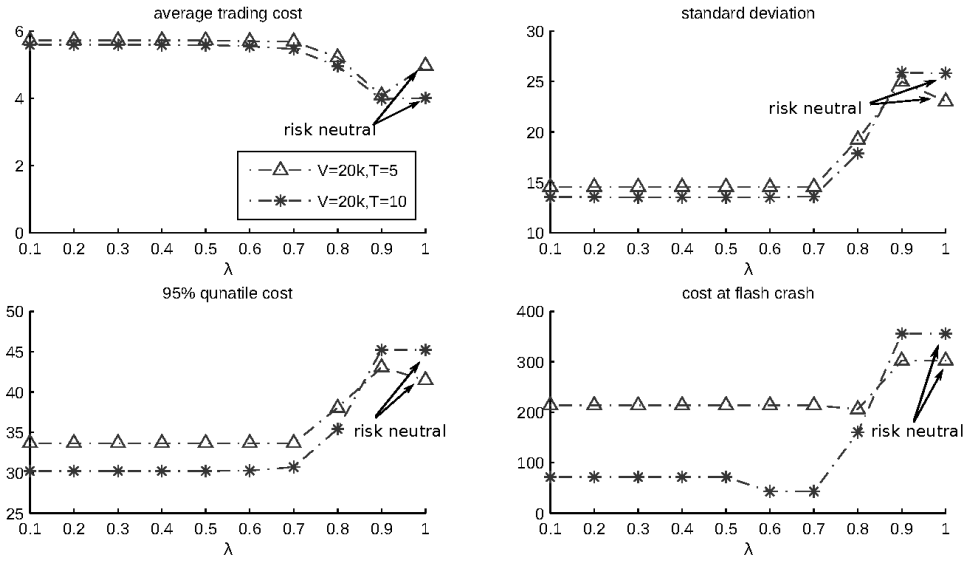


Figure 8.2: Performance of the risk-averse RL algorithm with the utility function defined in (8.6) against the choice of risk parameter λ. The choice of $\lambda = 1$ corresponds to the *risk-neutral* RL algorithm. The curve with "+" is obtained with total inventory $V = 20000$, time resolution $T = 5$ and inventory resolution $I = 5$, while the curve with "*" is with the setting $V = 20000$, $T = 10$ and $I = 10$.

its value is, therefore, of critical importance for the performance of the risk-averse RL algorithm. To this end, we first conduct a series of experiments with 10 different modes of increasing λ from 0.1 to 1. Figure 8.2 depicts the corresponding costs and risk for the two cases

1) $V = 20000$, $T = 5$ and $I = 5$, and

2) $V = 20000$, $T = 10$ and $I = 10$.

We observe that the average *cost* decreases while the standard deviation, 95% quantile cost and the cost at flash crash increase as λ increases. The underlying reason is that the greater the value of λ is, the closer is the shape of function in (8.6) to the linear function, and thus the less risk-averse is the algorithm. Indeed, with $\lambda = 1$, the algorithm reduces to the risk-neutral one which results in a slightly lower average of trading cost (approximately 2 basis points) during the whole test period, but a significantly higher risk (approximately 15 basis points, in terms of both standard deviation and 95%-quantile cost) than the risk-averse RL with $\lambda = 0.6$. Furthermore, during the flash crash period, the risk-averse algorithm outperform the risk-neutral one by 100 points for the case $T = 5$ or by 200 points for $T = 10$ (see the next paragraph for more comparisons of performances during the flash crash). In Figure 8.2, we can observe a switch point

around $\lambda \in [0.6, 0.7]$ indicating an optimal choice. We also identify the same pattern over other experimental cases (the graph is available upon request), and thus set $\lambda = 0.6$ in the following analysis.

Performance during the flash crash

The flash crash on May 6, 2010 provides a natural test case to examine the robustness of RL trading algorithms under extreme market conditions. Within ten-minute period from 14:40 to 14:50, the price of AMZN first dropped sharply down by approximately 4.23% followed by a quick recovery, ca. 2.94%, resulting a total decline, ca. 1.45%, over the period.

Figure 8.3 shows the dynamics of mid-quote prices and the corresponding trading costs for the case V=10000, T=10 and I, resulting from the risk-neutral and risk-averse RL algorithms from 12:00 to 16:00. We see that the risk-averse RL performs more stably than the risk-neutral RL over the period. Especially, during 10-minute period around the flash crash, the risk-averse RL managed to keep its trade cost at 41 basis points, while the risk-neutral RL led to a huge cost higher than 300 basis points. It indicates that the risk-neutral RL does not adapt to the extreme market event.

Besides the large jump of price, the flash crash is also characterized by a dramatically large width of bid-ask spread. In order to see whether including the spread as a state variable would remedy the risk-neutral RL's deflection, we compare the trading costs for all experimental cases under these two setups in Figure 8.4. It shows that the risk-neutral RL's performance during the flash crash does not significantly change, despite of adding substantial information, due to its non-sensitivity to spread state. We shall discuss the underlying reason in detail in the following paragraphs.

Overview of the performance in the whole test period

Figure 8.5 shows the cost and risk of the risk-neutral and risk-averse RL algorithms for all experimental cases under two setups, i.e. without and with the spread as an additional state variable. Overall, we find that risk-averse RL significantly reduces the risk by decreasing 10 to 15 basis points on the standard deviation and 95% quantile of the resulting trading cost, at the price of a slight increase of the average, ca. 2 to 3 basis points. It indicates that the proposed algorithm can not only avoid black swan events like the flash crash, but also reduce the overall risk in the whole test period at a price of a slight increase of average trading cost.

Furthermore, the results show that after introducing spreads as state variables, the trading risk resulting from the optimal policy of the risk-averse RL decreases remarkably, but hardly changes the risk of applying the risk-neutral RL. It is due to the fact that the optimal policy of the risk-neutral RL is based on the maximization of the expected reward (i.e., the expectation of shares executed in the next

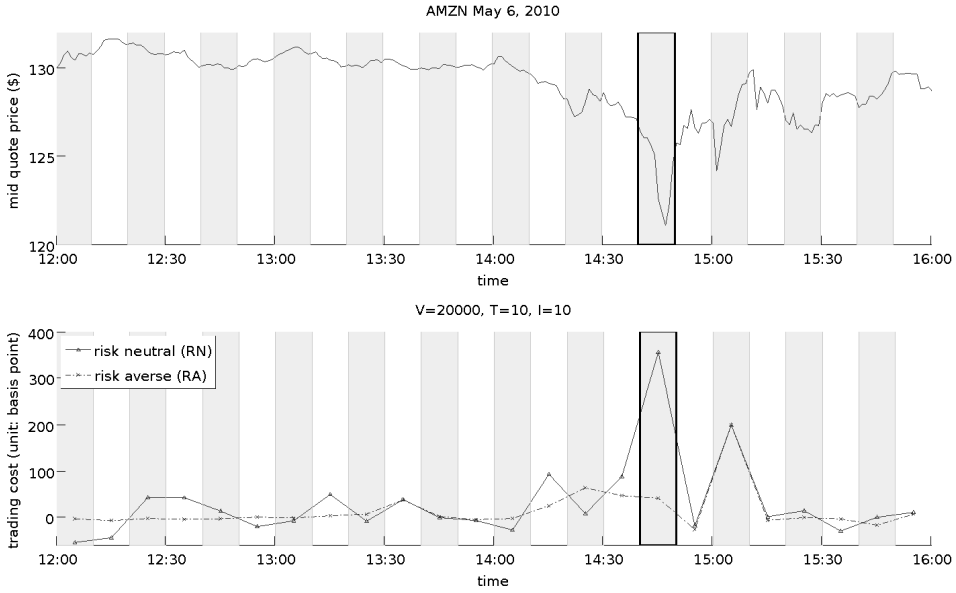


Figure 8.3: Top: mid quote price curve of AMZN on May 6, 2010. Each colored time interval is the trade time-horizon $H = 10$ minutes. The time interval when flash crash happened is highlighted by a black rectangular. Bottom: trading costs of the risk-neutral (labeled by "RN") RL and risk-averse (labeled by "RA") RL with $\lambda = .6$.

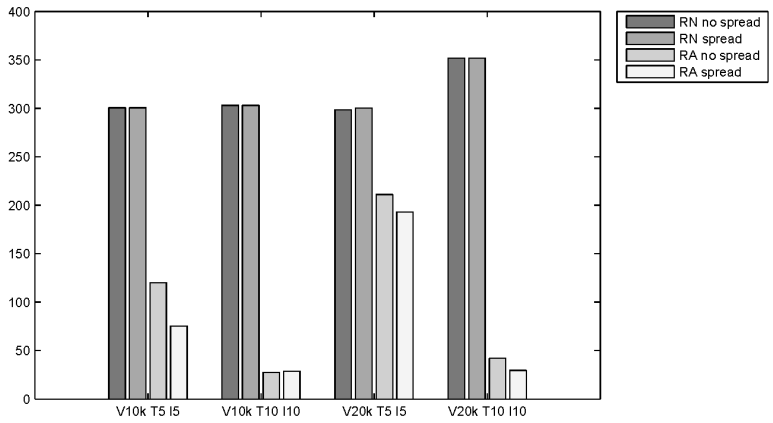


Figure 8.4: Trading costs on the flash crash spot (14:40–14:50 on May 6, 2010) with different combinations of algorithms (risk-neutral "RN" and risk-averse "RA"), volumes ($V = 10000, 20000$), time resolutions ($T = 5, 10$) and volume resolutions ($I = 5, 10$) and whether spreads are included in the state space ("no spread", "spread").

H/T time interval) among all admissible actions (i.e., decisions on the limit price of the order). Since the change of spread affects the expected rewards with different prices comparably, the risk-neutral RL will not adjust its strategy significantly according to the spread. This non-sensitivity of risk-neutral RL to spread explains why including the spread into simulation does not improve its performance during the flash crash.

Different from the risk-neutral RL, the risk-averse RL optimizes its policy by considering both reward and its uncertainty (i.e., the variance of the executed shares in the next time interval). The latter could change distinguishably among all admissible actions, when the spread changes. Therefore, the risk-averse RL tends to adjust its strategies substantially according to the spread state.

8.6 Conclusion

We have proposed a risk-averse reinforcement learning (RL) algorithm for optimal trade execution based on the risk-senstive MPD model for sequential trading decision. Our method has been tested by using 1.5 year high-frequency limit order book data in NASDAQ, which covers 2010 flash crash. Comparing with the risk-neutral RL, the risk-averse RL 1) significantly reduces the trading cost on the spot of flash-crash, and 2) is associated with a remarkable lower risk in the whole test period at the price of a slight increase of average trading cost.

We believe that incorporating risk control into trading algorithms with a nonlinear utility function would be a valuable guideline for the practitioners to im-

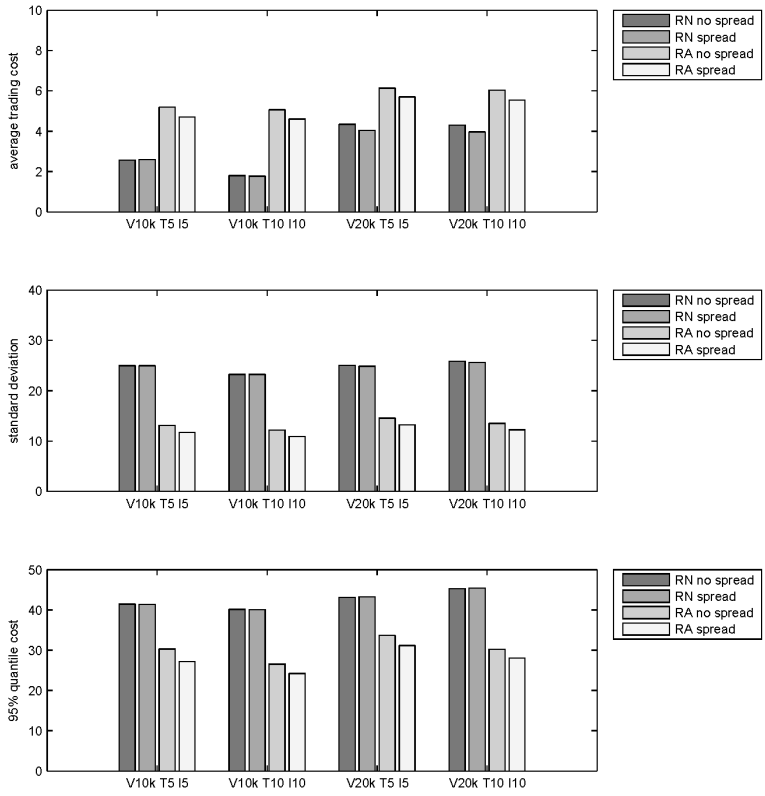


Figure 8.5: Performance of the risk-neutral (RN) and risk-averse (RA) RL with different combinations of volumes ($V = 10000, 20000$), time resolutions ($T = 5, 10$) and volume resolutions ($I = 5, 10$) and whether spreads are included in the state space ("no spread", "spread").

prove the robustness of their existing algorithms at extreme market events.

Bibliography

Alfonsi, A., Fruth, A., and Schied, A. (2010). Optimal execution strategies in limit order books with general shape functions. *Quantitative Finance*, 10(2):143–157. (Cited on p. 126)

Almgren, R. and Chriss, N. (2001). Optimal execution of portfolio transactions. *Journal of Risk*, 3:5–40. (Cited on p. 126)

Altman, E. (1999). *Constrained Markov Decision Processes*. CRC Press. (Cited on pp. 77, 92)

Arapostathis, A., Borkar, V., Fernández-Gaucherand, E., Ghosh, M., and Marcus, S. (1993). Discrete-time controlled Markov processes with average cost criterion: a survey. *SIAM Journal on Control and Optimization*, 31(2):282–344. (Cited on p. 41)

Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3):203–228. (Cited on pp. 2, 7, 12, 14, 15, 74)

Artzner, P., Delbaen, F., Eber, J., Heath, D., and Ku, H. (2007). Coherent multiperiod risk adjusted values and Bellman's principle. *Annals of Operations Research*, 152(1):5–22. (Cited on p. 33)

Avila-Godoy, G. and Fernández-Gaucherand, E. (1998). Controlled Markov chains with exponential risk-sensitive criteria: modularity, structured policies and applications. In *Proceedings of the 37th IEEE Conference on Decision and Control*, pages 778–783. (Cited on pp. 55, 74)

Ben-Tal, A. and Teboulle, M. (1987). Penalty functions and duality in stochastic programming via φ-divergence functionals. *Mathematics of Operations Research*, 12(2):224–240. (Cited on p. 24)

Ben-Tal, A. and Teboulle, M. (2007). An old-new concept of convex risk measures: the optimized certainty equivalent. *Mathematical Finance*, 17(3):449–476. (Cited on p. 24)

Bentham, J. (1789). *An Introduction to the Principles of Morals and Legislation*. T. Payne and Son. (Cited on p. 73)

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36. (Cited on p. 11)

Berns, G. S., Capra, C. M., Chappelow, J., Moore, S., and Noussair, C. (2008). Nonlinear neurobiological probability weighting functions for aversive outcomes. *NeuroImage*, 39(4):2047–2057. (Cited on p. 114)

Bernstein, P. L. (1997). *Against the Gods*. Simon & Schuster. (Cited on p. 7)

Bertsekas, D. and Shreve, S. (1978). *Stochastic Optimal Control: The Discrete Time Case*. Academic Press. (Cited on p. 82)

Bertsekas, D. and Tsitsiklis, J. (1996). *Neuro-Dynamic Programming*. Athena Scientific. (Cited on pp. 4, 76, 77, 104, 106, 107, 111)

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific, 2 edition. (Cited on p. 92)

Bertsimas, D. and Lo, A. W. (1998). Optimal control of execution costs. *Journal of Financial Markets*, 1(1):1–50. (Cited on pp. 126, 129)

Böhmer, W., Grünewälder, S., Shen, Y., Musial, M., and Obermayer, K. (2013). Construction of approximation spaces for reinforcement learning. *Journal of Machine Learning Research*, 14:2067–2118. (Cited on p. 111)

Borkar, V. S. (1998). Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851. (Cited on p. 106)

Borkar, V. S. (2002). Q-learning for risk-sensitive control. *Mathematics of Operations Research*, pages 294–311. (Cited on pp. 2, 96)

Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems View Point*. Cambridge University Press. (Cited on pp. 102, 106, 130)

Borkar, V. S. and Jain, R. (2010). Risk-constrained Markov decision processes. In *Proceedings of 49th IEEE Conference on Decision and Control (CDC)*, pages 2664–2669. IEEE. (Cited on p. 92)

Borkar, V. S. and Meyn, S. (2002). Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, pages 192–209. (Cited on pp. 55, 74)

Braun, D., Nagengast, A., and Wolpert, D. (2011). Risk-sensitivity in sensorimotor control. *Frontiers in Human Neuroscience*, 5. (Cited on pp. 2, 114)

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer. (Cited on p. 70)

Cavazos-Cadena, R. (2010). Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains. *Mathematical Methods of Operations Research*, 71(1):47–84. (Cited on pp. 55, 74)

Cavazos-Cadena, R. and Hernández-Hernández, D. (2009). Necessary and sufficient conditions for a solution to the risk-sensitive poisson equation on a finite state space. *Systems & Control Letters*, 58(4):254–258. (Cited on p. 70)

Chateauneuf, A. and Cohen, M. (2008). Cardinal extensions of the EU model based on the Choquet integral. *Decision-making Process*, pages 401–433. (Cited on p. 74)

Cheridito, P., Delbaen, F., and Kupper, M. (2006). Dynamic monetary risk measures for bounded discrete-time processes. *Electronic Journal of Probability*, 11(3):57–106. (Cited on pp. 4, 29, 74)

Cheridito, P. and Kupper, M. (2011). Composition of time-consistent dynamic monetary risk measures in discrete time. *International Journal of Theoretical and Applied Finance*, 14(1):137–162. (Cited on pp. 74, 79)

Cheridito, P. and Li, T. (2009). Risk measures on Orlicz hearts. *Mathematical Finance*, 19(2):189–214. (Cited on pp. 16, 18, 56)

Choquet, G. (1953). Theory of capacities. In *Annales de l'institut Fourier*, volume 5. (Cited on pp. 12, 16, 19)

Chung, K. and Sobel, M. (1987). Discounted MDPs: distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25:49. (Cited on pp. 55, 74, 81)

Coraluppi, S. and Marcus, S. (2000). Mixed risk-neutral/minimax control of discrete-time, finite-state Markov decision processes. *IEEE Transactions on Automatic Control*, 45(3):528–532. (Cited on pp. 18, 19, 55, 74, 96)

Csiszar, I. (1967). On topological properties of f-divergences. *Studia Sci. Math. Hungar.*, 2:329–339. (Cited on p. 24)

Dayan, P. and Niv, Y. (2008). Reinforcement learning: the good, the bad and the ugly. *Current Opinion in Neurobiology*, 18(2):185–196. (Cited on p. 113)

Del Moral, P., Ledoux, M., and Miclo, L. (2003). On contraction properties of Markov kernels. *Probability Theory and Related Fields*, 126(3):395–420. (Cited on p. 37)

Delbaen, F. (2000). Coherent risk measures on general probability spaces. *Advances in Finance and Stochastics Essays in Honour of Dieter Sondermann*, pages 1–37. (Cited on pp. 14, 15, 16, 19)

Denneberg, D. (1994). *Non-additive Measure and Integral*. Kluwer Academic Publishers. (Cited on pp. 12, 19)

Detlefsen, K. and Scandolo, G. (2005). Conditional and dynamic convex risk measures. *Finance and Stochastics*, 9(4):539–561. (Cited on pp. 32, 33)

Di Masi, G. and Stettner, L. (2008). Infinite horizon risk sensitive control of discrete time Markov processes under minorization property. *SIAM Journal on Control and Optimization*, 46(1):231. (Cited on pp. 55, 74, 89, 90)

Dolgov, D. A. and Durfee, E. H. (2003). Approximating optimal policies for agents with limited execution resources. In *IJCAI*, pages 1107–1112. (Cited on p. 92)

Doob, J. L. (1953). *Stochastic Processes*. New York Wiley. (Cited on p. 39)

Douc, R., Fort, G., Moulines, E., and Priouret, P. (2009). Forgetting the initial distribution for hidden Markov models. *Stochastic Processes and Their Applications*, 119(4):1235–1256. (Cited on pp. 60, 70)

Douc, R., Moulines, E., and Rosenthal, J. S. (2004). Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Annals of Applied Probability*, pages 1643–1665. (Cited on p. 37)

Dunkel, J. and Weber, S. (2010). Stochastic root finding and efficient estimation of convex risk measures. *Operations Research*, 58(5):1505–1521. (Cited on pp. 20, 103)

Feinberg, E. A., Kasyanov, P. O., and Zadoianchuk, N. V. (2013). BergeâĂŹs theorem for noncompact image sets. *Journal of Mathematical Analysis and Applications*, 397(1):255–259. (Cited on p. 82)

Feinberg, E. A. and Shwartz, A. (1996). Constrained discounted dynamic programming. *Mathematics of Operations Research*, 21(4):922–945. (Cited on p. 92)

Filar, J., Kallenberg, L., and Lee, H. (1989). Variance-penalized Markov decision processes. *Mathematics of Operations Research*, pages 147–161. (Cited on p. 92)

Fleming, W. and Hernández-Hernández, D. (1997). Risk-sensitive control of finite state machines on an infinite horizon I. *SIAM Journal on Control and Optimization*, 35(5):1790–1810. (Cited on pp. 55, 74)

Föllmer, H. and Penner, I. (2006). Convex risk measures and the dynamics of their penalty functions. *Statistics & Decisions*, 24(1):61–96. (Cited on pp. 74, 78)

Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4):429–447. (Cited on pp. 2, 7, 12, 14, 15, 17, 18, 56, 74)

Föllmer, H. and Schied, A. (2004). *Stochastic Finance*. Walter de Gruyter & Co., Berlin. Extended edition. (Cited on pp. 20, 21, 56)

Gaubert, S. and Gunawardena, J. (2004). The Perron-Frobenius theorem for homogeneous, monotone functions. *Transactions American Mathematical Society*, 356(12):4931–4950. (Cited on pp. 38, 70)

Geibel, P. and Wysotzki, F. (2005). Risk-sensitive reinforcement learning applied to control under constraints. *J. Artif. Intell. Res.*, 24:81–108. (Cited on p. 92)

Ghosh, J., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis*. Springer, New York. (Cited on p. 118)

Gilboa, I. (2009). *Theory of Decision under Uncertainty*. Cambridge University Press. (Cited on pp. 1, 9)

Gläscher, J., Hampton, A., and O'Doherty, J. (2009). Determining a role for ventromedial prefrontal cortex in encoding action-based value signals during reward-related decision making. *Cerebral Cortex*, 19(2):483–495. (Cited on p. 121)

Glimcher, P., Camerer, C., Fehr, E., and Poldrack, R. (2008). *Neuroeconomics: Decision Making and the Brain*. Academic Press. (Cited on pp. 3, 122)

Glynn, P. W. and Meyn, S. P. (1996). A Liapounov bound for solutions of the Poisson equation. *The Annals of Probability*, pages 916–931. (Cited on p. 37)

Gollier, C. (2004). *The Economics of Risk and Time*. The MIT Press. (Cited on pp. 1, 2, 74)

Good, I. (1986). A minor comment concerning Hempel's paradox of confirmation. *Journal of Statistics, Computation and Simulation*, 24(3-4):320–321. (Cited on p. 9)

Gunawardena, J. and Keane, M. (1995). On the existence of cycle times for some nonexpansive maps. Technical Report HPL-BRIMS-95-003, Hewlett-Packard Labs. (Cited on p. 64)

Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications*. Cambridge University Press. (Cited on p. 65)

Hairer, M. and Mattingly, J. (2011). Yet another look at Harris' ergodic theorem for Markov chains. In *Seminar on Stochastic Analysis, Random Fields and Applications VI*, pages 109–117. Springer. (Cited on pp. 4, 37, 38, 39, 63, 70)

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press. (Cited on p. 40)

Hamm, A.-M., Salfeld, T., and Weber, S. (2013). Stochastic root finding for optimized certainty equivalents. In *Winter Simulation Conference*, pages 922–932. (Cited on p. 25)

Harris, T. (1956). The existence of stationary measures for certain Markov processes. In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, volume 2, pages 113–124. (Cited on p. 37)

Hautsch, N. and Huang, R. (2012). The market impact of a limit order. *Journal of Economic Dynamics & Control*, 36:501–522. (Cited on p. 132)

Heger, M. (1994). Consideration of risk in reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 105–111. (Cited on pp. 2, 96)

Hempel, C. G. (1945). Studies in the Logic of Confirmation I. *Mind*, pages 1–26. (Cited on p. 9)

Hernández-Hernández, D. and Marcus, S. (1996). Risk sensitive control of Markov processes in countable state space. *Systems & Control Letters*, 29(3):147–155. (Cited on pp. 2, 55, 74)

Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes*. Springer. (Cited on p. 38)

Hernández-Lerma, O. and Lasserre, J. (1996). *Discrete-time Markov Control Processes: Basic Optimality Criteria*. Springer. (Cited on pp. 1, 73, 75)

Hernández-Lerma, O. and Lasserre, J. (1999). *Further Topics on Discrete-Time Markov Control Processes*. Springer Verlag. (Cited on pp. 1, 73, 75, 77, 82, 88)

Hernández-Lerma, O. and Lasserre, J. (2003). *Markov Chains and Invariant Probabilities*. Birkhäuser Verlag. (Cited on pp. 30, 37)

Hernández-Lerma, O., Montes-De-Oca, R., and Cavazos-Cadena, R. (1991). Recurrence conditions for Markov decision processes with Borel state space: a survey. *Annals of Operations Research*, 28(1):29–46. (Cited on p. 88)

Howard, R. and Matheson, J. (1972). Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369. (Cited on pp. 2, 17, 55)

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310(5754):1680–1683. (Cited on p. 114)

Hsu, M., Krajbich, I., Zhao, C., and Camerer, C. F. (2009). Neural response to reward anticipation under risk is nonlinear in probabilities. *The Journal of Neuroscience*, 29(7):2231–2237. (Cited on p. 114)

Huang, R. and Polak, T. (2011). LOBSTER: The limit order book reconstructor. Technical report, School of Business and Economics, Humboldt Universität zu Berlin. http://lobster.wiwi.hu-berlin.de/Lobster/LobsterReport.pdf. (Cited on p. 130)

Iyengar, G. (2005). Robust dynamic programming. *Mathematics of Operations Research*, pages 257–280. (Cited on p. 18)

Jaakkola, T., Jordan, M., and Singh, S. (1994). On the Convergence of Stochastic Iterative Dynamic Programming Algorithms. *Neural Computation*, 6(6):1185–1201. (Cited on p. 106)

Jorion, P. (2007). *Value at Risk: the New Benchmark for Managing Financial Risk*. McGraw-Hill New York. (Cited on p. 132)

Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, 47(2):263–292. (Cited on pp. 2, 3, 7, 11, 12, 22, 96, 102, 114)

Kirilenko, A., Kyle, A., Samadi, M., and Tuzun, T. (2011). The flash crash: The impact of high frequency trading on an electronic market. *SSRN:1686004*. (Cited on p. 125)

Koenig, S. and Simmons, R. (1994). Risk-sensitive planning with probabilistic decision graphs. In *Proceedings of Principles of Knowledge Representation and Reasoning*, pages 363–373. Morgan Kaufmann. (Cited on p. 96)

Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer. (Cited on p. 8)

Kontoyiannis, I. and Meyn, S. (2003). Spectral theory and limit theorems for geometrically ergodic Markov processes. *Annals of Applied Probability*, pages 304–362. (Cited on p. 63)

Kontoyiannis, I. and Meyn, S. (2005). Large deviations asymptotics and the spectral theory of multiplicatively regular Markov processes. *Electron. J. Probab.*, 10(3):61–123. (Cited on pp. 37, 63)

Koopmans, T. C. (1960). Stationary ordinal utility and impatience. *Econometrica: Journal of the Econometric Society*, pages 287–309. (Cited on p. 33)

Kreps, D. M. and Porteus, E. L. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica: Journal of the Econometric Society*, pages 185–200. (Cited on p. 33)

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86. (Cited on p. 17)

Kushner, H. and Yin, G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Verlag. (Cited on pp. 20, 76, 102)

Ledoux, M. (2001). *The Concentration of Measure Phenomenon*. American Mathematical Society. (Cited on p. 56)

Lemmens, B. and Nussbaum, R. (2012). *Nonlinear Perron-Frobenius Theory*. Number 189. Cambridge University Press. (Cited on p. 70)

Liu, Y., Goodwin, R., and Koenig, S. (2003). Risk-averse auction agents. In *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 353–360. ACM. (Cited on p. 96)

Makowski, A. M. and Shwartz, A. (2002). The Poisson equation for countable Markov chains: probabilistic methods and interpretations. In *Handbook of Markov Decision Processes*, pages 269–303. Springer. (Cited on p. 37)

Marcus, S., Fernández-Gaucherand, E., Hernández-Hernandez, D., Coraluppi, S., and Fard, P. (1997). Risk sensitive Markov decision processes. *Progress in Systems and Control Theory*, 22:263–280. (Cited on pp. 55, 74)

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, 7(1):77–91. (Cited on p. 17)

Menkveld, A. J. and Yueshen, B. (2013). Anatomy of the flash crash. *SSRN:2243520*. (Cited on p. 125)

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag. (Cited on pp. 37, 47, 71, 75)

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49(2):267–290. (Cited on pp. 2, 22, 96, 106, 114)

Morgenstern, O. and Neumann, J. V. (1944). *Theory of Games and Economic Behavior*. Princeton University Press. (Cited on pp. 10, 11)

Nagengast, A., Braun, D., and Wolpert, D. (2010). Risk-sensitive optimal feedback control accounts for sensorimotor behavior under uncertainty. *PLoS Computational Biology*, 6(7):e1000857. (Cited on pp. 2, 114)

Nevmyvaka, Y., Feng, Y., and Kearns, M. (2006). Reinforcement learning for optimized trade execution. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 673–680. ACM. (Cited on pp. 126, 128, 131)

Niv, Y., Edlund, J., Dayan, P., and O'Doherty, J. (2012). Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *The Journal of Neuroscience*, 32(2):551–562. (Cited on pp. 2, 3, 114, 118)

Obizhaeva, A. A. and Wang, J. (2012). Optimal trading strategy and supply/demand dynamics. *Journal of Financial Markets*. (Cited on p. 126)

O'Doherty, J. (2004). Reward representations and reward-related learning in the human brain: insights from neuroimaging. *Current Opinion in Neurobiology*, 14(6):769–776. (Cited on p. 120)

Ogryczak, W. and Ruszczyński, A. (1999). From stochastic dominance to mean-risk models: Semideviations as risk measures. *European Journal of Operational Research*, 116(1):33–50. (Cited on pp. 17, 25, 26)

Powell, W. B. (2007). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley & Sons. (Cited on p. 111)

Prashanth, L. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. In *Advances in Neural Information Processing Systems*, pages 252–260. (Cited on p. 92)

Preuschoff, K., Quartz, S., and Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *The Journal of Neuroscience*, 28(11):2745–2752. (Cited on pp. 3, 114)

Puterman, M. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc. (Cited on pp. 1, 70, 73, 75, 76, 77, 80, 97, 128)

Riedel, F. (2004). Dynamic coherent risk measures. *Stochastic processes and their applications*, 112(2):185–200. (Cited on p. 32)

Rockafellar, R. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42. (Cited on p. 24)

Roorda, B., Schumacher, J., and Engwerda, J. (2005). Coherent acceptability measures in multiperiod models. *Mathematical Finance*, 15(4):589–612. (Cited on p. 74)

Ruszczyński, A. (2010). Risk-averse dynamic programming for Markov decision processes. *Mathematical Programming*, pages 1–27. (Cited on pp. 2, 4, 29, 32, 33, 74, 78, 79, 80, 81)

Ruszczyński, A. and Shapiro, A. (2006). Conditional risk mappings. *Mathematics of Operations Research*, 31(3):544–561. (Cited on p. 74)

Ruszczynski, A. and Shapiro, A. (2006). Optimization of convex risk functions. *Mathematics of operations research*, 31(3):433–452. (Cited on pp. 16, 25)

Savage, L. (1972). *The Foundations of Statistics*. Dover Publications. (Cited on pp. 2, 9, 74)

Savage, S. L. (2009). *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. John Wiley & Sons. (Cited on p. 1)

Schäl, M. (1974). A selection theorem for optimization problems. *Archiv der Mathematik*, 25(1):219–224. (Cited on p. 82)

Schied, A. (2007). Optimal investments for risk-and ambiguity-averse preferences: a duality approach. *Finance and Stochastics*, 11(1):107–129. (Cited on p. 24)

Schied, A., Föllmer, H., and Weber, S. (2009). Robust preferences and robust portfolio choice. *Handbook of Numerical Analysis*, 15:29–87. (Cited on pp. 16, 20, 24)

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, 36(2):241–263. (Cited on pp. 1, 96, 120)

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599. (Cited on pp. 1, 96)

Securities and Exchange Commission (2010). Concept release on equity market structure. *Federal Register*, 75(13):3594–3614. (Cited on p. 125)

Shen, Y., Huang, R., Yan, C., and Obermayer, K. (2014a). Risk-averse reinforcement learning for algorithmic trading. In Serguieva, A., Maringer, D., Palade, V., and Almeida, R., editors, *Proceedings of 2014 IEEE Computational Intelligence for Financial Engineering and Economics*, pages 391–398. (Cited on p. 125)

Shen, Y., Stannat, W., and Obermayer, K. (2013). Risk-sensitive Markov Control Processes. *SIAM Journal on Control and Optimization*, 51(5):3652–3672. (Cited on pp. 7, 35, 73)

Shen, Y., Stannat, W., and Obermayer, K. (2014b). Risk-sensitive Markov control processes with general convex risk maps. *Arxiv preprint arXiv:1403.3321*. (Cited on pp. 35, 73)

Shen, Y., Stannat, W., and Obermayer, K. (2014c). A unified framework for risk-sensitive Markov control processes. *To Appear in Proceedings of 53rd IEEE Conference on Decision and Control*. (Cited on p. 73)

Shen, Y., Tobia, M., Sommer, T., and Obermayer, K. (2014d). Risk-sensitive reinforcement learning. *Neural Computation*, 26(7):1298–1328. (Cited on pp. 7, 95, 113)

Simon, M. and Barry, R. (1980). *Methods of Modern Mathematical Physics*, volume I. New York, Academic Press. (Cited on p. 10)

Sobel, M. (1982). The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802. (Cited on p. 92)

Sutton, R. and Barto, A. (1998). *Reinforcement Learning*. MIT Press. (Cited on pp. 1, 76, 96, 111, 120)

Svindland, G. (2009a). *Convex Risk Measures Beyond Bounded Risks*. PhD thesis, Ludwig-Maximilians-Universität München. (Cited on p. 19)

Svindland, G. (2009b). Subgradients of law-invariant convex risk measures on L^1. *Statistics & Decisions*, 27(2):169–199. (Cited on pp. 16, 43, 46)

Symmonds, M., Wright, N., Bach, D., and Dolan, R. (2011). Deconstructing risk: Separable encoding of variance and skewness in the brain. *NeuroImage*, 58(4):1139–1149. (Cited on pp. 3, 122)

Tamar, A., Di Castro, D., and Mannor, S. (2012). Policy gradients with variance related risk criteria. In *International Conference on Machine Learning*. (Cited on p. 92)

Tobia, M., Guo, R., Schwarze, U., Böhmer, W., Gläscher, J., Finckh, B., Marschner, A., Büchel, C., Obermayer, K., and Sommer, T. (2014). Neural systems for choice and valuation with counterfactual learning signals. *NeuroImage*, 89:57–69. (Cited on p. 114)

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16(3):185–202. (Cited on p. 106)

Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323. (Cited on pp. 12, 20, 22, 23, 74, 119)

U.S. Commodity Futures Trading Commission and Securities & Exchange Commission (2010). Findings regarding the market events of May 6, 2010. *Report of the Staffs of the CFTC and SEC to the Joint Advisory Committee on Emerging Regulatory Issues*. (Cited on p. 125)

Vega-Amaya, O. (2003). The average cost optimality equation: a fixed point approach. *Bol. Soc. Mat. Mexicana*, 9(1):185–195. (Cited on p. 88)

Wakker, P. P. (2010). *Prospect Theory: for Risk and Ambiguity*. Cambridge University Press. (Cited on p. 12)

Walters, P. (2000). *An Introduction to Ergodic Theory*. Springer. (Cited on p. 37)

Watkins, C. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College. (Cited on pp. 96, 106)

White, D. J. (1993). *Markov Decision Processes*. John Wiley & Sons New York, NY. (Cited on pp. 1, 73, 75, 77)

Wu, S.-W., Delgado, M., and Maloney, L. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences*, 106(15):6088–6093. (Cited on pp. 3, 114)