ABSTRACT

Title of dissertation: STOCHASTIC SYSTEMS WITH CUMULATIVE PROSPECT THEORY

Kun Lin, Doctor of Philosophy, 2013

Dissertation directed by: Professor Steven I. Marcus
Department of Electrical & Computer Engineering

Stochastic control problems arise in many fields. Traditionally, the most widely used class of performance criteria in stochastic control problems is risk-neutral. More recent attempts at introducing risk-sensitivity into stochastic control problems include the application of utility functions. The decision theory community has long debated the merits of using expected utility for modeling human behaviors, as exemplified by the Allais paradox. Substantiated by strong experimental evidence, Cumulative Prospect Theory (CPT) based performance measures have been proposed as alternatives to expected utility based performance measures for evaluating human-centric systems. Our goal is to study stochastic control problems using performance measures derived from the cumulative prospect theory.

The first part of this thesis solves the problem of evaluating Markov decision processes (MDPs) using CPT-based performance measures. A well-known method of solving MDPs is dynamic programming, which has traditionally been applied with an expected utility criterion. When the performance measure is CPT-inspired,

several complications arise. Firstly, when solving a problem via dynamic programming, it is important that the performance criterion has a recursive structure, which is not true for all CPT-based criteria. Secondly, we need to prove the traditional optimality criteria for the updated problems (i.e., MDPs with CPT-based performance criteria). The theorems stated in this part of the thesis answer the question: what are the conditions required on a CPT-inspired criterion such that the corresponding MDP is solvable via dynamic programming?

The second part of this thesis deals with stochastic global optimization problems. Using ideas from the cumulative prospect theory, we are able to introduce a novel model-based randomized optimization algorithm: Cumulative Weighting Optimization (CWO). The key contributions of our research are: 1) proving the convergence of the algorithm to an optimal solution given a mild assumption on the initial condition; 2) showing that the well-known cross-entropy optimization algorithm is a special case of CWO-based algorithms. To the best knowledge of the author, there is no previous convergence proof for the cross-entropy method. In practice, numerical experiments have demonstrated that a CWO-based algorithm can find a better solution than the cross-entropy method.

Finally, in the future, we would like to apply some of the ideas from cumulative prospect theory to games. In this thesis, we present a numerical example where cumulative prospect theory has an unexpected effect on the equilibrium points of the classic prisoner's dilemma game.

STOCHASTIC SYSTEMS WITH CUMULATIVE PROSPECT THEORY

by

Kun Lin

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Steven I. Marcus, Chair/Advisor
Professor Michael C. Fu
Professor P.S. Krishnaprasad
Professor Nuno Martins
Professor Gang Qu

Dedication

To my family.

Acknowledgments

Most importantly, I would like to express my most sincere gratitude to Professor Steven I. Marcus. Without his support and guidance, this dissertation would not have been possible. His integrity, dedication and patience have made a lasting impression on me. A graduate experience, riddled with set-backs and struggles, was made much more pleasant by his willingness to provide sound advice that led me back on the right track.

I would like to thank my committee members for taking the time out of their busy schedules to read this manuscript and attend my defense. From Professor Michael C. Fu, I have gained many insights through our frequent research meetings. Professor P.S Krishnaprasad, Professor Nuno Martins, and Professor Gang Qu have all inspired me through their interesting lectures in their respective fields.

In addition, I would like to express many thanks to my colleagues and officemates, Enlu Zhou, Yongqiang Wang, James Ferlez, and Bhaskar Ramasubramanian, for the many interesting discussions on topics both technical and non-technical.

For the many people not mentioned here, but who have definitely contributed to the completion of this thesis, I would like to express my deepest gratitude.

Contents

List of Tables

List of Figures

List of Algorithms

Nomenclature

ATSP Asymmetric Traveling Salesman Problems

ATSP Asymmetric Traveling Salesman Problems

CDF Cumulative Density Function

CPT Cumulative Propsect Theory

CWO Cumulative Weighting Optimization

MDP Markov Decision Problem

PDE Partial Differential Equation

Chapter 1

Overview

1.1 Motivation

Many relevant real-life problems can be modeled as stochastic systems (e.g., weather, traffic patterns, financial markets, communication systems). A system could be stochastic for many reasons. For one, the randomness could be introduced by inaccurate sensors (i.e., measurement error). Sometimes, we lack sufficient information about the system, and model our ignorance by intentionally incorporating randomness in the model (i.e., model error). Quantum mechanics support the idea that uncertainty is part of the natural order of the universe. For whatever the reason might be, studying stochastic systems is important for solving many real-life problems from various fields. This thesis will try to tackle a few problems in stochastic systems that are inspired by recent advances in decision theory. More specifically, we use some of the latest performance measures suggested by the decision theory community to evaluate the performances of stochastic systems. These novel problems are particularly suited for studying human-centric systems (e.g., war games, consumer behaviors, medical decisions).

1.1.1 Stochastic Optimal Control Problems

> *"The concept of control can be described as the process of influencing the behavior of a dynamical system to achieve a desired goal. If the goal is to optimize some payoff function (or cost function) which depends on the control inputs to the system, then the problem is one of optimal control."*
>
> - Wendell H. Fleming and H. Mete Soner [38]

If a stochastic system has a control input along with a performance criterion, then the resulting problem is a stochastic optimal control problem. Stochastic optimal control problems have many applications in engineering. The evidence of their successful applications can be found in a wide-range of fields (i.e., robotics, route planning, space exploration). In finance, the seminal paper by Black & Scholes in 1973 [14] provides insight into the management of risks, which leads to an equation for valuing options.

There are three approaches for solving stochastic optimal control problems, namely dynamic programming (Hamilton-Jacobi-Bellman equation), the maximum principle, and the martingale and convex duality approach [68]. Dynamic programming, most popular in the analysis of controlled Markov processes, provides sufficient conditions for optimality in its verification theorem, stating that if there exists a policy satisfying the Hamilton-Jacobi-Bellman PDE, then it is an optimal policy. The essence of dynamic programming is Bellman's optimality principle, which roughly says if one knows an optimal policy for an entire period, then starting from any time in that period and at a state along an optimal trajectory, the

same policy is still optimal. This insight leads to the realization that decomposing the optimal value function into two parts (i.e., immediate value and value-to-go) is the key to obtaining an optimal policy. Intuitively, one can obtain the Hamilton-Jacobi-Bellman PDE by first considering the discrete time Bellman equation with time step h, dividing both sides by h, and then taking h to zero. One key feature of dynamic programming is that a solution of the Hamilton-Jacobi-Bellman PDE is a function of the state. In practice, this means the computational complexity of dynamic programming grows exponentially as the number of states increases, leading Bellman to coin the term "curse of dimensionality". On the other hand, a state-feedback policy can be found easily once a solution is obtained, which can be implemented simply as a lookup-table. Fleming & Soner [38] and Bertsekas [8] are excellent textbooks for a comprehensive review of dynamic programming in the controlled Markov process setting. Alternatively, the maximum principle provides necessary conditions for optimality. The stochastic maximum principle is similar in spirit to its deterministic counterpart. For the interested reader, Haussmann [44], Peng [67], Yong and Zhou [89] are excellent sources for a further investigation into this topic. The deterministic maximum principle can be intuitively described as perturbing an optimal control over an interval ϵ. By taking the first order Taylor approximation of the corresponding value function with respect to ϵ and sending ϵ to zero, we obtain a variational inequality. Combining the variational inequality with the co-state equations, the deterministic maximum principle is complete. The stochastic maximum principle differs from its deterministic counterpart in its usage of forward-backward stochastic differential equations to describe the dynamics of

its state and adjoint variables. The martingale approach, which was originated by Pliska [70] and gained popularity in the mathematical finance community, divides the problem into two subproblems: 1) Find the optimizer for the problem at a fixed terminal time T. If the cost function is convex, then the problem of finding the optimizer for the problem can be reduced from an infinite-dimensional problem to a finite-dimensional problem by using convex duality. 2) Use the martingale Representation Theorem to extract the corresponding optimal control. In Pliska's original paper, convexity of the cost function is an important assumption in proving the existence of the solution to the dual static problem. Pham [68] has a recent discussion on this approach.

1.1.2 Dynamic Programming

Of the three approaches we mention above, dynamic programming has proven to be the most popular method for solving dynamic stochastic optimization problems with controlled Markov processes. Numerically, dynamic programming can be applied using either value iteration or policy iteration. In value iteration, the algorithm could start from an infeasible[1] value function and converges to a feasible one. On the other hand, applying policy iteration results in value functions that are feasible. Perhaps the most important reason for dynamic programming's popularity is its production of a feedback policy, which is advantageous for storage, execution (i.e., a table lookup) and robustness. As shown in Bertsekas [8], the breadth of problems that can be solved using dynamic programming includes inventory control, deterministic

[1] A value function is feasible if it is yielded by a feasible policy.

scheduling problems, machine repair, reachability of ellipsoidal tubes, and pursuit-evasion games. Problems presented in his book all have the following flavor: 1) an underlying discrete time dynamic system; 2) a cost function that is additive over time. The natural question to ask for an inquisitive mind is: why do we want to evaluate performances using expectation? If our goal is to predict, in particular, what a human would do in many situations, then expected value contradicts much empirical evidence. Hence, in applying dynamic programming to human decision making processes, we need to have a model that agrees with empirical data. The discussion of the merits of various classes of performance criteria is the focus of the next section. This thesis deviates from the standard dynamic programming approach by updating the dynamic programming framework with a more general class of performance criteria. A host of issues arise from doing this: does a dynamic programming equation even exist?

1.1.3 Performance Criteria: From Expected Value to Prospect Theory

Using expected value as a performance criterion has been a long tradition in many engineering and scientific fields. Why is that? Is it only out of its mathematical convenience (i.e., linearity)? In this section, we will trace the development of the expected utility theory and highlight some of its deficiencies. For a more in-depth analysis of the development in the area of prospect theory, the interested reader can refer to [86], which the discussion below draws many facts and examples from.

When given a dynamic control problem, we would like to use an appropriate performance measure for each situation. Hence, understanding its implications is of paramount importance. Using the wrong performance measure might yield an ineffective policy. Each performance criterion encapsulates our preference ordering of the potential outcomes[2] due to our action/decision. In other words, if we know we prefer outcome L_1 to L_2[3] , then we must use a performance criterion that reflects this preference (i.e., $\rho(L_1) > \rho(L_2)$). On the other hand, we would also like the implications of such a preference ordering to be sensible for the problem at hand. For example, in the expected value case, de Finetti [27] (also see a survey by Fishburn [35] for the axiomatization of expected value) shows that the existence of subjective probabilities is equivalent to transitivity, monotonicity and additivity. In addition, the existence of a certainty equivalent (i.e., deterministic) value for each possible outcome guarantees the no-arbitrage condition for the preference system. In other words, expected value performance criteria have properties that we deem rational (i.e., transitivity, monotonicity, and additivity), and lack some undesirable attributes (e.g., arbitrage). However, expected value does have some limitations. One particular limitations that spurred the search for its alternative, expected utility, is best demonstrated in the St. Petersburg paradox[4]. The paradox is an example of a game having an infinite certainty equivalent value (i.e., the price one is willing to pay) under expected value; However, in practice, people often are only willing to pay a finite amount for the game. The paradox was resolved by

[2] An element of the probability space is usually called an outcome.

[3] L_1 is a short-hand notation for lottery 1, not to be confused with the function space.

[4] see appendix on prospect theory

Bernoulli in 1738 [7] by suggesting that people do not evaluate outcomes by their objective values, but rather by their utility. This realization started the new field of expected utility theory.

Expected utility as a performance criterion still remains popular. Perhaps such success can be attributed to its axiomatization by von Neumann & Morgenstern [85]. von Neumann provides the necessary and sufficient conditions for using maximization of expected utility as a function for ordering preferences. Fishburn [34] made updates to von Neumann's work in the 1970s. These conditions include completeness, transitivity, continuity, and substitution. Despite many justifications for using expected utility, it suffers from a well known contradiction with empirical observations demonstrated by the Allais paradox [2], where most people violate the substitution axiom implied by expected utility theory. Lesser known but more recent discussions on the empirical violations of expected utility can be found in [82] and [71].

Cumulative prospect theory resolves all of the paradoxes mentioned above, and has stronger empirical support compared with expected utility theory. There is also a strong behavioral foundation found in its axiomatization [19]. A key feature of cumulative prospect theory is probabilistic sensitivity. This is different from the traditional approach of outcome sensitivity in the expected utility theory. We will first demonstrate, via an example, that risk-aversion can have an equivalent representation outside of expected utility theory.

Example 1. This example is from [86], which demonstrates that risk-sensitivity

	a	b	c	d	e
% outcome is 100	0.10	0.30	0.50	0.70	0.90
% outcome is 0	0.90	0.70	0.50	0.30	0.10
Certainty Equivalent Value	1	9	25	49	81

Table 1.1.1: Example: Data-Equivalence

can be represented either as outcome sensitivity or probability sensitivity. We are given five certainty equivalent (i.e., indifference) values and probability pairs.

From the table above, we see that the value a person places on an outcome (e.g., a,b,c,d or e) might not be a linear evaluation. In other words, in prospect b,

$$0.30 \times 100 + 0.70 \times 0 = 30 \neq 9.$$

We can of course find a utility function U such that

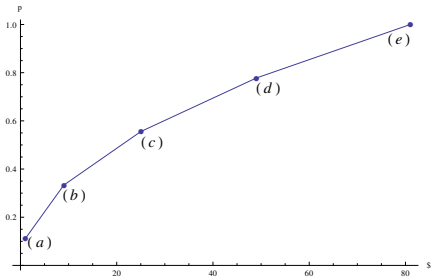
$$\sum U(x)\, p(x)$$

agrees with the values in the table above, where p is a probability mass function and $U(\cdot)$ is a utility function. This operation can be equivalently achieved by using a probability weighting function w,

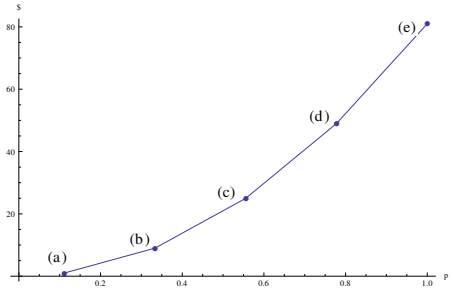
$$\sum w(p)\, x,$$

where instead of transforming the outcomes, we are now transforming the probability weights (see Figure 1.1.1).

Remark 1. In the example above, we are only trying to demonstrate that risk-



(a) Normalized EV vs. Certainty Equivalent



(b) Certainty Equivalent vs. Normalized EV

Figure 1.1.1: Data Equivalence

sensitivity can be expressed equivalently either as outcome sensitivity or probability sensitivity. We want to emphasize the fact that $w(p) : x \rightarrow [0, 1]$, transforms the probabilities based on the entire probability mass function. In other words, $w : P \rightarrow P$, is a mapping from P, the space of probability mass functions, to P.

The example above should convince the reader that every utility function has an equivalent probability weighting function. This, of course, is not the full story. If probability weighting is only equivalent to expected utility, then we would not be so interested in it. After all, if probability weighting functions can predict only as well as expected utility, we will not need to advance risk-sensitive performance measures beyond utility theory. Several sources have demonstrated that in many cases probability weighting gives different predictions than those given by outcome weighting (i.e., expected utility). Onay and Öncüler [66] demonstrate the predictions offered by outcome based risk-aversion are different from that of probability weighting. More importantly, the predictive power of the probability weighting approach is confirmed by their experiments.

1.1.4 Stochastic Optimization with Probability Weighting Functions

Stochastic optimization is another field where a novel approach is inspired by cumulative prospect theory (CPT). In CPT, the probability weighting function has the effect of weighting rare good-news events more than other events. In stochastic optimization, if we can apply this shift in weights iteratively to sampled distributions, we can intuitively understand how we might converge to an optimal value. There are a few desirable properties for stochastic optimization algorithms: 1) we would like the algorithms to increase, in expected value, monotonically; 2) once an optimal solution is obtained, we would like it to be robust again perturbations. As the reader will see in Chapter 3, our method, cumulative weighting optimization (CWO) exhibits both of these properties. We also will develop CWO-based numerical algorithms and present their simulated results in the same chapter. Interestingly, the well-known cross-entropy method is a special case of CWO-based algorithms. In fact, we are able to improve the performance of the cross-entropy method by viewing it as such.

1.1.5 Outline

The thesis is organized by chapters. Chapter 2 will prove the suitability of dynamic programming equations for non-convex performance measures, which include CPT-based criteria. We will present both the finite horizon and infinite horizon cases. In addition, we will also analyze the structure of optimal policies yield by CPT-based criteria and compare them with other risk-sensitive performance criteria. In

Chapter 3, we will provide convergence proofs for cumulative weighting optimization methods. Numerical examples will be provided to demonstrate the performance of our algorithms. In the last chapter, we will discuss our contributions and future work.

Chapter 2

Dynamic Programming with Non-Convex Risk-Sensitive Measures

Dynamic programming with risk-sensitive performance measures has applications in many fields (e.g., operations research, finance, control systems). Historically, risk-sensitive performance measures are represented using expected utility functions. More recently, literature on dynamic performance (i.e., risk or reward) measures has inspired an alternative approach to risk-sensitive performance evaluation. The dynamic performance measure framework is a generalization of the classical work using expected value. One limitation of this approach is that it has only been developed for coherent performance measures, which exclude a large class of important non-convex performance measures (e.g., cumulative prospect theory (CPT) based performance measures). We remedy this limitation by proving the optimality of the dynamic programming equation for non-convex performance measures.

2.1 Introduction

Dynamic programming, introduced by Bellman [5], is a dynamic optimization method. It has been the subject of intense research in the past five decades; see for example [6, 10, 15, 32, 63, 48, 73]. Dynamic optimization problems modeled by controlled Markov processes and solved via dynamic programming are commonly referred to as Markov decision processes (MDPs). Researchers have developed tech-

niques to lift MDPs' curse-of-dimensionality (e.g., approximate dynamic programming [11, 72, 9]), which enable the application of dynamic programming in many fields (e.g., operations research, finance, control systems).

In many applications, risk-sensitive measures are more appropriate than risk-neutral measures [52, 56, 57]. In standard MDPs, the performance measures are frequently expressed as expected utility functions that are risk-sensitive [23, 17, 36, 37, 33, 46, 47, 24, 25]. For example, many problems evaluate their outcomes by using the performance measure $\mathbb{E}[u(X)]$, where u is a risk-sensitive utility function (e.g., exponential), and X is a random variable representing the reward[1]. A notable feature of optimal policies, induced by the risk-sensitive performance measures, is their robustness with respect to modeling errors [30].

An important class of risk-sensitive performance measures is coherent risk measures, of which $\mathbb{E}[u(X)]$, where $u(\cdot)$ is a convex function, is a special case [3, 28, 40, 41, 64, 79, 78]. Other well known examples include mean-semideviation and conditional value-at-risk. An important property of coherent risk measures is convexity. Recently, their dynamic counterparts have received great interests in the literature [74, 20, 31, 39, 43, 22, 21, 4, 60]. In many problems, convex performance measures are not the best option for measuring the desirability of outcomes. A well-known example of a non-convex performance measure is suggested by Tversky and Kahneman in the cumulative prospect theory (CPT) [83]. Although CPT had its beginning in the 1990s, its incorporation into dynamic systems is still nascent. Recently, He and Zhou have studied [45] a portfolio choice problem with a non-

[1]Reward is often the sum of per-stage rewards.

convex performance measure. The problem maximizes the terminal wealth of a self-financing portfolio[2] driven by a financial market[3] that is uncontrollable from the perspective of the investor (see [45], Eq. 3). Often, the financial market is assumed to be a Markov semimartingale and has a nonempty set of equivalent martingale measures. Under these assumptions, one can apply the martingale approach (see [68], Chapter 7) to arrive at the desired analytical results. These results become more difficult, if not impossible, to obtain if these assumptions are eliminated. This chapter will study both convex and non-convex performance measures (e.g., CPT-inspired reward measures) when the underlying model is a discrete-time controlled Markov process.

The goal of this chapter is to address, when the underlying system is modeled as a controlled Markov process, the question: *How can we generalize dynamic programming to both convex and non-convex performance measures?* An approach, suggested by Ruszczyński [76], is based on *dynamic risk measures* and *risk transition mapping*s (see [77, Definition 5]). Assuming a sequence of time-consistent[4] risk measures is given (see [76, Theorem 1]), he concludes that if the corresponding one-step dynamic risk measures satisfy the four assumptions of coherent performance measures, namely convexity, monotonicity, translation equivalence, and positive homogeneity, and an equivalent Markov risk transition mapping exists for each one-step dynamic risk measure, then a dynamic programming equation exists for the dynamic optimization problem. Unfortunately, since CPT-inspired measures have

[2]This is just a constraint on the action space of the MDP.

[3]A special case of a controlled semi-martingale Markov process.

[4]Time-consistency is key for rewriting the risk measures into their nested forms, which can be easily optimized via dynamic programming.

nonlinear weighting functions, they do not satisfy some of these assumptions.

We derive the dynamic programming equation for a class of non-convex reward measures (e.g., CPT-inspired reward measures). Our work has many parallels with that of Ruszczyński; however our goal is to generalize his approach to non-convex reward measures. Before we proceed, we will review some background material in controlled Markov processes, CPT, reward transition mappings, and dynamic reward measures.

2.2 Background

In the following sections, we use the following notations:

- $(\cdot)_+ := max(0, \cdot)$; $(\cdot)_- := -min(0, \cdot)$;
- $\mathcal{P}(\cdot)$: the set of probability measures defined on $\cdot$.

2.2.1 Discrete-Time Markov Control Model

We are interested in the case when the underlying system dynamics can be modeled as a discrete-time controlled Markov process. Let us first review the necessary technical background for our discussion. We restate the definition from [48] for the reader's convenience. A Markov control model is a five-tuple, $(\mathbb{X}, \mathbb{A}, \{A(x)|x \in \mathbb{X}\}, Q, r)$, consisting of:

- a Polish space $\mathbb{X}$, called the state space and whose elements are referred to as states;

- a Polish space $\mathbb{A}$, called the control or action set;
- a family $\{A(x) \in \mathbb{A} | x \in \mathbb{X}\}$ of nonempty measurable subsets A(x) of $\mathbb{A}$, where A(x) denotes the set of feasible controls or actions when the system is in state $x \in \mathbb{X}$, and with the property that the set

$$\mathbb{K} := \{(x, a) | x \in \mathbb{X},\ a \in A(x)\} \tag{2.2.1}$$

  of feasible state-action pairs is a measurable subset with respect to the product σ-algebra of $\mathbb{X} \times \mathbb{A}$ (i.e., $\sigma(\mathbb{X} \times \mathbb{A})$);
- a stochastic kernel[5] $Q(\cdot|x, a)$ on $\mathbb{X}$, where $(x, a) \in \mathbb{K}$;
- a measurable function r:$\mathbb{K} \times \mathbb{X} \to \mathbb{R}$ called the *per-stage reward* function.

Remark 2. We can make $A(x)$ and r time-varying, denoted by $A_t(x)$ and r_t, by considering the state space $\mathbb{X}' := \mathbb{X} \cup [0, \ldots, T]$.

Polish spaces include finite-dimensional real spaces, which are important for many real-life applications (e.g. dynamic pricing). The following definition is useful for describing the set of feasible deterministic and randomized policies.

Definition 1. We denote by $\mathbb{F}$ the set of all measurable functions $f : \mathbb{X} \to \mathbb{A}$ such that $f(x) \in A(x)$ for all $x \in \mathbb{X}$. In addition, we let Ψ denote the set of all

[5] A stochastic kernel on X given Y is a function $P(\cdot|\cdot)$ such that

1. $P(\cdot|y)$ is a probability measure on X for each fixed $y \in Y$;
2. $P(B|\cdot)$ is a measurable function on Y for each fixed $B \in \mathcal{B}(X)$.

stochastic kernels ψ in $\mathcal{P}(\mathbb{A}|\mathbb{X})$, the set of probability measures on $\mathbb{A}$ given $\mathbb{X}$, such that $\psi(A(x)|x) = 1$ for every $x \in \mathbb{X}$.

We track a system's history by doing the following: for each $t = 0, 1, \ldots,$ define the space H_t of admissible histories up to time t as $H_0 := \mathbb{X}$, and

$$H_t := \mathbb{K}^t \times \mathbb{X} = \mathbb{K} \times H_{t-1}, \ \forall t = 1, 2, \ldots.$$

The most general policies we investigate are randomized policies, which are defined below.

Definition 2. A *randomized policy* is a sequence $\pi = \{\pi_t, \ t = 0, 1, \ldots\}$ of stochastic kernels $\pi_t \in P(\mathbb{A}|H_t)$ satisfying the constraint

$$\pi_t(A(x_t)|h_t) = 1, \ \forall x_t \in \mathbb{X}, \ h_t \in H_t, \ t = 0, 1, \ldots.$$

The set of all randomized policies is denoted by Π.

A special class of randomized polices is the class of randomized Markov policies.

Definition 3. A randomized policy, $\pi \in \Pi$, is a *randomized Markov policy* if there exists a sequence of stochastic kernels $\psi_t \in \Psi$ such that

$$\pi_t(\cdot|h_t) = \psi_t(\cdot|x_t) = 1$$

$$\forall h_t \in H_t, \, t = 0, 1, \ldots.$$

The policy π is a *randomized stationary policy* if there is a $\psi \in \Psi$ such that

$$\pi_t(\cdot|h_t) = \psi(\cdot|x_t)$$

$$\forall h_t \in H_t,\ t = 0, 1, \ldots.$$

We denote the sets of randomized Markov policies and randomized stationary polices by Π^{RM} and Π^{RS}, respectively.

Furthermore, if there exists a sequence $f_t \in \mathbb{F}$ such that $\psi_t(\cdot|x_t)$ is the Dirac measure concentrated at $f(x_t)$ for all $t = 0, 1, \ldots$, then π is a *deterministic Markov policy*, and $\pi_t := f_t \in \mathbb{F}$. We denote the sets of all deterministic Markov policies and deterministic stationary policies by Π^{DM} and Π^{DS}, respectively.

By fixing a Markov control model, an initial probability distribution v (e.g., a known initial state x_0), and a randomized policy π, we obtain the probability distribution evolution of a discrete-time Markov process. We denote the resulting discrete-time Markov process and action sequence by $\{x_t^\pi\}$ and $\{a_t^\pi\}$ (i.e., a_t^π is a random variable with probability distribution $\pi_t(\cdot|\{x_0, \ldots, x_t\})$) respectively. For ease of notation, we drop the process's dependence on its initial condition, as it is fixed unless stated otherwise. For the rest of the discussion, we are given a fixed Markov control model.

2.2.2 Cumulative Prospect Theory (CPT)

Before introducing cumulative prospect theory, we will first introduce a useful definition.

Definition 4. A *good-news distribution*, $\tilde{F}$, of a random variable is defined as

$$\tilde{F}(x) := 1 - F(x),$$

where $F(x)$ is the cumulative distribution function (CDF). Other names for this distribution are: survival distribution, complementary CDF and reliability distribution.

Remark 3. The above definition should be altered if we are given a minimization problem. In that case, a good-news distribution function should be the cumulative distribution function itself, because smaller values are more favorable.

Another important element of CPT is probability weighting functions, which are defined below.

Definition 5. A *probability weighting function*, w, is a continuous function from $[0, 1]$ to $[0, 1]$.

Prospect theory was suggested in the 1970s by Kahneman and Tversky [59]. They were unsatisfied with the theory and suggested its improved version, cumulative prospect theory (CPT), in the 1990s [83]. CPT asserts that the human decision making process can be modeled by a utility function with the following characteristics:

- The utility function has a reference point against which gains and losses are measured;

- The utility function is concave on gains and convex on losses (i.e., horizontal S-shape);

- A probability weighting function that transforms the probability measure such that a small probability is inflated and a large probability is deflated. For example, a typical weighting function $w : [0,1] \to [0,1]$ is

$$w(y) := \frac{y^\gamma}{(y^\gamma + (1-y)^\gamma)^{\frac{1}{\gamma}}},$$

where $\gamma \in (0,1)$ and y is usually the good-news distribution. This function was originally presented in [83].

Definition 6. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and random variables R and B defined on it. A *CPT performance measure* has the following form:

$$\begin{aligned}\rho(R) := &\int_0^\infty w^+\left(\mathbb{P}\left(u_+\left((R-B)_+\right) > s\right)\right) ds \\ &- \int_0^\infty w^-\left(\mathbb{P}\left(u_-\left((R-B)_-\right) > s\right)\right) ds, \end{aligned} \tag{2.2.2}$$

where $w^+ : [0,1] \to [0,1]$ and $w^- : [0,1] \to [0,1]$ are two continuous non-decreasing functions. $u_+ : \mathbb{R}^+ \to \mathbb{R}^+$ and $u_- : \mathbb{R}^+ \to \mathbb{R}^+$ are two utility functions. The random variable B represents the benchmark we measure the performance against.

The weighting functions used in a CPT performance measure are required to be non-decreasing, which is not necessarily true for a probability weighting function. We apply a CPT-inspired measure to evaluate the expected outcome of a game of

dice in the example below.

Example 2. Cumulative Prospect Theory - Finite State Case with no Control: Consider a game of dice. You roll a die with six possible outcomes $\{1, 2, 3, 4, 5, 6\}$. If the outcome is even, you win an amount equal to the outcome; on the other hand, if the outcome is odd, then you lose an amount equal to the outcome. Thus, the payoffs are {-5, -3, -1, 2, 4, 6}. Furthermore, the payoffs are organized into gains and losses. The probability of gains is derived by assuming the die is fair and written as $\{0 : \frac{1}{2}; 2 : \frac{1}{6}; 4 : \frac{1}{6}; 6 : \frac{1}{6}\}$, which is read as the probability of winning 0 is $\frac{1}{2}$, the probability of winning 2 is $\frac{1}{6}$ and so on. On the down side, a similar calculation leads to $\{0 : \frac{1}{2}; -1 : \frac{1}{6}; -3 : \frac{1}{6}; -5 : \frac{1}{6}\}$. Since the initial state of the die does not matter in this case, the CPT expected value calculation is as follows:

$$\begin{aligned}
\tilde{V} &= u_+\left((2)_+\right)\left(w_+(\frac{1}{2}) - w_+(\frac{1}{3})\right) + u_+\left((4)_+\right)\left(w_+(\frac{1}{3}) - w_+(\frac{1}{6})\right) \\
&+ u_+\left((6)_+\right)\left(w_+(\frac{1}{6}) - w_+(0)\right) - u_-\left((-5)_-\right)\left(w_-(\frac{1}{6}) - w_-(0)\right) \\
&- u_-\left((-3)_-\right)\left(w_-(\frac{1}{3}) - w_-(\frac{1}{6})\right) - u_-\left((-1)_-\right)\left(w_-(\frac{1}{2}) - w_-(\frac{1}{3})\right) \\
&= u_+(2)\left(w_+(\frac{1}{2}) - w_+(\frac{1}{3})\right) + u_+(4)\left(w_+(\frac{1}{3}) - w_+(\frac{1}{6})\right) \\
&+ u_+(6)\left(w_+(\frac{1}{6}) - w_+(0)\right) - u_-(5)\left(w_-(\frac{1}{6}) - w_-(0)\right) \\
&- u_-(3)\left(w_-(\frac{1}{3}) - w_-(\frac{1}{6})\right) - u_-(1)\left(w_-(\frac{1}{2}) - w_-(\frac{1}{3})\right).
\end{aligned}$$

Here, we use different probability weighting functions for gains and losses, namely $w^+ : [0, 1] \to [0, 1]$ and $w^- : [0, 1] \to [0, 1]$. The two functions, $u_+ : \mathbb{R}^+ \to \mathbb{R}^+$ and $u_- : \mathbb{R}^+ \to \mathbb{R}^+$, are two utility functions. In this example, the reference point

is assumed to be zero.

Remark 4. We presented in Eq. 2.2.2 the most general form of a CPT reward measure. In the sequel, we will study various special cases of this reward measure. For example, we are interested in the case when the rewards are strictly positive.

2.2.3 Reward Transition Mappings

In this section, we consider the Markov control model

$$(\mathbb{X}, \mathbb{A}, \{A(x)|x \in \mathbb{X}\}, Q, r).$$

The discrete-time Markov process resulting from applying the policy π and the corresponding action sequence are denoted by $\{x_t^\pi\}$ and $\{a_t^\pi\}$, respectively. A standard finite-horizon dynamic control problem has the following performance measure:

$$\max_{\pi \in \Pi} \mathbb{E}\left[\sum_{t=0}^{T-1} r(x_t^\pi, a_t^\pi, x_{t+1}^\pi) + r_T(x_T^\pi)|x_0\right], \tag{2.2.3}$$

where r_T is a measurable terminal reward function. We would like to solve the optimization problem

$$V_T^*(x) := \max_{\pi \in \Pi} V_T(x, \pi).$$

From [48], the corresponding dynamic programming equation is

$$v_t(x) = \max_{\delta \in \mathcal{P}(A(x))} \int_{\mathbb{X}\times\mathbb{A}} (r(x, a, y) + v_{t+1}(y))\, Q(dy|x, a)\delta(da). \tag{2.2.4}$$

Remark 5. Under some assumptions (i.e., the existence of a measurable deterministic selector), $\delta \in \mathcal{P}(A(x))$ in Eq. 2.2.4 can be replaced by $\delta \in A(x)$ to reflect the fact that a deterministic optimal policy exists.

The right-hand side of Eq. 2.2.4 is a function of the current state x, the reward function parameterized by the current state x (i.e., $g_x(a, y) := r(x, \cdot, \cdot) + v_{t+1}(\cdot) : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$), the transition probability Q, and the randomized control δ. Taking one step further, we can define a function

$$\sigma_t \left(r(x, \cdot, \cdot) + v_{t+1}(\cdot), x, \delta \circ Q(\cdot|x, \cdot) \right) := \int_{\mathbb{X}\times\mathbb{A}} \left(r(x, a, y) + v_{t+1}(y) \right) Q(dy|x, a)\delta(da),$$

and rewrite Eq. 2.2.4 as

$$v_t(x) = \max_{\delta \in \mathcal{P}(A(x))} \sigma_t \left(r(x, \cdot, \cdot) + v_{t+1}(\cdot), x, \delta \circ Q(\cdot|x, \cdot) \right).$$

The sequence, $\{\sigma_t,\ t = 0, \ldots, T-1\}$, is called the reward transition mappings for Eq. 2.2.4. Before we can provide the definition for reward transition mappings, we need to define the term $\delta \circ Q(\cdot|x, \cdot)$ in the equation above.

Definition 7. Given a fixed current state $x \in \mathbb{X}$ and a randomized action $\delta \in \mathcal{P}(\mathbb{A})$, we denote the *one-step state-action measure* (see [18]) with respect to the Markov control model by:

$$[\delta \circ Q_x](B_a \times B_y) := \int_{B_a} Q(B_y|x, a)\delta(da)\ B_y \in \mathcal{B}(\mathbb{X})\ B_a \in \mathcal{B}(\mathbb{A}), \tag{2.2.5}$$

where $Q_x(a) := Q(\cdot|x,a) : \mathbb{A} \to \mathcal{P}(\mathbb{X})$ is the stochastic kernel parameterized by $x \in \mathbb{X}$.

Remark 6. The one-step state-action measure is a measure over $\mathbb{X} \times \mathbb{A}$, which represents the uncertainty over the next state and the current action (i.e., we are interested in randomized policies).

We need to define the space that contains $\delta \circ Q_x$, which was also mentioned in Çavuş & Ruszczyński [18]. Given a probability space $(\mathbb{X} \times \mathbb{A}, \mathcal{B}(\mathbb{X} \times \mathbb{A}), P_0)$, where P_0 is some reference probability measure, the space of p-integrable random variables is denoted by $\mathcal{V} := \mathcal{L}_p(\mathbb{X} \times \mathbb{A}, \mathcal{B}(\mathbb{X} \times \mathbb{A}), P_0),\ p \in [1, \infty)$. Its dual space, $\mathcal{V}'$, is the space of signed measures on $(\mathbb{X} \times \mathbb{A}, \mathcal{B}(\mathbb{X} \times \mathbb{A}))$, which are absolutely continuous with respect to P_0 with their densities in $\mathcal{L}_q(\mathbb{X} \times \mathbb{A}, \mathcal{B}(\mathbb{X} \times \mathbb{A}), P_0)$, where q satisfies the equation $\frac{1}{p} + \frac{1}{q} = 1$. The reference measure, P_0, should be chosen such that all possible measures of the form $\delta \circ Q_x$ are in $\mathcal{V}'$. In the special case of a finite state and control space, P_0 can always be chosen to be uniform. We denote the set of all probability measures in $\mathcal{V}'$ by:

$$\mathcal{M} := \{m \in \mathcal{V}' | m(\mathbb{X} \times \mathbb{A}) = 1,\ m \geq 0\}.$$

Remark 7. The measure defined by Eq. 2.2.5 is an element of $\mathcal{M}$.

The space $\mathcal{V}'$ (and thus $\mathcal{M}$) is endowed with the Prokhorov topology (weak convergence). For $p \in [1, \infty)$ we will endow $\mathcal{V}$ with the strong (i.e., norm) topology.

However, if $p = \infty$, we will endow $\mathcal{V}$ with the topology induced by the form:

$$\langle \psi, m \rangle = \int_{\mathbb{X}\times\mathbb{A}} \psi(x,a) m(dx,da),\ \psi \in \mathcal{V},\ m \in \mathcal{V}'.$$

Definition 8. A mapping $\sigma : \mathcal{V} \times \mathbb{X} \times \mathcal{M} \to \mathbb{R}$ is a *reward transition mapping* if for every $x \in \mathbb{X}$ and every $m \in \mathcal{M}$ fixed (denote $\sigma(\cdot) := \sigma(\cdot, x, m)$), the following conditions are true:

1) if $\phi \leq \psi$ then $\sigma(\phi) \leq \sigma(\psi),\ \forall \phi, \psi \in \mathcal{V}$;

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

This definition is more general than Definition 3.1 in [18]. Since CPT inspired performance measures are distorted by nonlinear probability weighting functions, they generally do not satisfy the convexity and translation invariant requirements satisfied by convex risk measures. By removing these two requirements, we are able to work with non-convex CPT inspired performance measures.

2.2.4 Generalized Markov Dynamic Reward Measures

The definition of dynamic reward measures varies based on the objective of the analysis. The dynamic risk measure community, for example [76], defines dynamic performance measures based on coherent risk measures. Since our goal is to define a class of dynamic performance measures that contains non-convex performance measures (e.g., CPT inspired reward measures), we need to modify the definitions used by the dynamic risk measure community. However, because we are maximizing rewards rather than minimizing risks, our definitions are defined by switching the

direction of the analogous inequalities.

Given a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ with $\mathcal{F}_0 = \{\Omega, \emptyset\}$, we define the spaces $\mathcal{L}_t = \mathcal{L}_p(\Omega, \mathcal{F}_t, P),\ p \in [1, \infty],\ t = 0, 1, \ldots, T,\ \textit{and}\ \mathcal{L}_{t,T} = \mathcal{L}_t \times \cdots \times \mathcal{L}_T$.

Remark 8. Given a Markov control model, Ω can be thought of as H_T and $\mathcal{F}_T$ as $\sigma(H_T)$.

Definition 9. A mapping $\rho_{t,T} : \mathcal{L}_{t,T} \to \mathcal{L}_t$, where $1 \leq t \leq T$, is called a *conditional reward measure*, if it has the following monotonicity property:

$$Z \geq W \text{ implies } \rho_{t,T}(Z) \geq \rho_{t,T}(W),\ \forall Z, W \in \mathcal{L}_{t,T}.$$

Definition 9 was first presented by Ruszczyński in [76]. The inequality above is meant to be component-wise almost surely. Intuitively, the definition above says a conditional reward measure preserves the order of the rewards. Furthermore, taking $R \in \mathcal{L}_{t,T}$ to be a sequence of future rewards, $\rho_{t,T}(R)$ gives the price, at time t, that one is willing to pay to obtain the payoff sequence R.

Definition 10. A *dynamic reward measure* is a sequence of conditional reward measures $\{\rho_{t,T},\ t = 1, \ldots, T\}$.

In other words, a dynamic reward measure is a time-varying mapping that reflects the present value of a sequence of future rewards. It can be utilized as a performance measure in many real-life scenarios. One important concept in dynamic reward measures is time-consistency, which is defined below.

Definition 11. A dynamic reward measure $\{\rho_{t,T}\}_{t=1}^{T}$ is called *time-consistent* if for all $1 \leq \tau < \theta \leq T$ and all sequences $Z, W \in \mathcal{L}_{\tau,T}$ the conditions

$$Z_k = W_k,\ k = \tau, \ldots, \theta - 1 \text{ and } \rho_{\theta,T}(Z_\theta, \ldots, Z_T) \geq \rho_{\theta,T}(W_\theta, \ldots, W_T)$$

imply that

$$\rho_{\tau,T}(Z_\tau, \ldots, Z_T) \geq \rho_{\tau,T}(W_\tau, \ldots, W_T).$$

In applications, a time-consistent dynamic reward measure can be more conveniently represented by its corresponding sequence of one-step conditional reward measures, whose definition is given below.

Definition 12. A mapping $\rho_t : \mathcal{L}_{t+1} \to \mathcal{L}_t$ is called a *one-step conditional reward measure* if

$$\rho_t(Z) = \rho_{t,t+1}(0, Z),\ Z \in \mathcal{L}_{t+1}.$$

For this thesis, we are only interested in one-step conditional reward measures that satisfy the assumption below.

Assumption 1. *A one-step conditional reward measure satisfies the following conditions:*

1. If $Z \leq W$ then $\rho_t(Z) \leq \rho_t(W)$, $\forall Z, W \in \mathcal{L}_{t+1}$;
2. $\rho_t(\beta Z) = \beta \rho_t(Z),\ \forall Z \in \mathcal{L}_{t+1}, \beta \geq 0$.

Below are several one-step conditional rewards that satisfy Assumption 1.

Example 3. The following reward measures are both convex reward measures.

Mean-semideviation model:

$$\rho_t(Z_{t+1}) = \mathbb{E}\left[Z_{t+1}|\mathcal{F}_t\right] + \kappa\mathbb{E}\left[\left(\left(Z_{t+1} - \mathbb{E}\left[Z_{t+1}|\mathcal{F}_t\right]\right)_+\right)^r |\mathcal{F}_t\right]^{\frac{1}{r}}.$$

Here, $r \in [1, p]$ and $\kappa \in [0, 1]$ may be any $\mathcal{F}_t$-measurable random variables.

Another interesting example is the Average Conditional Value at Risk:

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

where the infimum is point-wise, and α is any $\mathcal{F}_t$-measurable function with values in an interval $[\alpha_{min}, \alpha_{max}] \in (0, 1)$.

The next example is an example of a non-convex reward measure in Z_{t+1}.

Example 4. Cumulative Prospect Theory:

$$\rho_t(Z_{t+1}) = \int_0^\infty w_t^+ \left(P\left(u_t^+\left(\left(Z_{t+1}\right)_+\right) > s\right)\right) ds - \int_0^\infty w_t^- \left(P\left(u_t^-\left(\left(Z_{t+1}\right)_-\right) > s\right)\right) ds, \tag{2.2.6}$$

where $w_t^+,\ w_t^+,\ u_t^-,$ and u_t^- are $\mathcal{F}_t$-measurable functions with values in the function spaces $[0, 1] \to [0, 1],\ [0, 1] \to [0, 1],\ \mathbb{R}^+ \to \mathbb{R}^+$ and $\mathbb{R} \to \mathbb{R}^+$, respectively (see Eq. 2.2.2), and P is an appropriate probability measure. Here, the benchmark random variable B in Eq. 2.2.2 is zero.

Remark 9. The performance measures in Example 3 satisfy the convexity and trans-

lation invariance assumptions in Ruszczyński's work, whereas the performance measure (i.e., Eq. 2.2.6) in Example 4 does not. Eq. 2.2.6 is the main motivation for us to generalize Ruszczyński's approach.

Applying a one-step conditional reward measure to a controlled Markov process, we ideally would like to obtain an optimal Markov policy. However, we cannot expect this to be true in general, because the one-step reward measure could depend on the past history of the underlying Markov process (i.e., h_t). In order to overcome this difficulty, we follow Ruszczyński's ([76]) definition of the one-step Markov conditional reward measure.

2.2.4.1 Markov Conditional Reward Measures

As we mentioned in the previous section, one-step conditional reward measures might not be Markov. However, if a one-step conditional reward has a corresponding reward transition mapping, then it only depends on the current state of the system, hence it is Markov. The following condition is important for the integrability of Markov conditional reward measures.

Definition 13. A function g is said to be b-bounded if $\exists C > 0$ and $b : \mathbb{X} \to [1, \infty),\ b \in \mathcal{V}$ and

$$|g(x, a, y)| \leq C\left(b(x) + b(y)\right),\ \forall x \in \mathbb{X},\ a \in A(x),\ y \in \mathbb{X}.$$

We denote the function $g(x, a, y) : \mathbb{X} \times \mathbb{A} \times \mathbb{X}$ with the x argument parameterized by $g_x : \mathbb{A} \times \mathbb{X} \to \mathbb{R}$ (i.e., $g_x(a, y) := g(x, a, y)$). In addition, the notation $\pi_{t,x}$

denotes the measure $\pi_t(\cdot|x) \in \mathcal{P}(\mathbb{A})$. We remind the reader that $Q_{t,x}$, the transition probability at time t, is a mapping $a \to Q_t(\cdot|x,a)$.

We consider the filtered probability space $(H_T, \sigma(H_T), \mathcal{F}_t, \mathbb{P}^\pi)$, where $\mathcal{F}_t$ is the σ-field generated by the state-action trajectory (i.e., $\{x_0^\pi, a_0^\pi, \ldots, x_t^\pi\}$) of the controlled Markov process $\{x_t^\pi\}$. The space $\mathcal{L}_t$ in the definition below is defined with respect to the filtered probability space $(H_T, \sigma(H_T), \mathcal{F}_t, \mathbb{P}^\pi)$. More specifically, elements of $\mathcal{L}_t$ are functions of $\{x_0^\pi, a_0^\pi, \ldots, x_t^\pi\}$.

Definition 14. A one-step conditional reward measure $\rho_t : \mathcal{L}_{t+1} \to \mathcal{L}_t$ is a *Markov reward measure with respect to a controlled Markov process* $\{x_t^\pi\}$ *and its controls* $\{a_t^\pi\}$, if there exists a reward transition mapping $\sigma_t : \mathcal{V} \times \mathbb{X} \times \mathcal{M} \to \mathbb{R}$, such that for any b-bounded measurable functions $g : \mathbb{X} \times \mathbb{A} \times \mathbb{X} \to \mathbb{R}$, there is a feasible control $\pi_t : \mathbb{X} \to \mathcal{P}(\mathbb{A}(x))$ such that the following equation holds

$$\rho_t \left(g(x_t^\pi, a_t^\pi, x_{t+1}^\pi) \right) = \sigma_t(g_{x_t^\pi}, x_t^\pi, \pi_{t,x_t^\pi} \circ Q_{t,x_t}), \ a.s. \tag{2.2.7}$$

Remark 10. In the sequel, we use the term one-step Markov reward measure for both ρ_t and its corresponding σ_t. Furthermore, the right-hand side of Eq. 2.2.7 can be thought of as a function parameterized by the current state x_t^π.

Definition 15. A one-step conditional reward measure ρ_t is *Markov*, if ρ_t is a Markov reward measure with respect to all feasible controlled Markov processes and controls $\{\{x_t^\pi\}, \{a_t^\pi\} | \pi \in \Pi\}$ and σ_t is the same for all $\pi \in \Pi$. Furthermore, a dynamic reward measure $\{\rho_t\}$ is Markov, if each of the one-step conditional reward measure ρ_t is Markov.

In other words, if a conditional reward measure ρ_t is Markov, then we can replace it with its Markov counterpart σ_t when calculating the reward at time t.

2.3 Dynamic Programming

2.3.1 Finite-Horizon

Given a time-consistent dynamic reward measure $\{\rho_{t,T}\}_{t=0}^{T-1}$ and its corresponding one-step dynamic reward measure $\{\rho_t\}_{t=0}^{T-1}$, we can write the corresponding value function starting at x_0 with a control policy $\pi \in \Pi$ and the resulting state-action trajectory $\{x_0^\pi, a_0^\pi, \ldots, x_T^\pi\}$ as:

$$\begin{aligned} V_T(x_0, \pi) = \rho_0 \left(r(x_0^\pi, a_0^\pi, x_1^\pi) + \rho_1 \left(r(x_1^\pi, a^\pi{}_1, x_2^\pi) \right. \right. \\ + \rho_2 \left(r(x_2^\pi, a^\pi{}_2, x_3^\pi) + \cdots + \right. \\ \left. \left. \left. \rho_{T-1} \left(r(x_{T-1}^\pi, a_{T-1}^\pi, x_T^\pi) + r_T(x_T^\pi) \right) \cdots \right) \right) \right). \end{aligned}$$

The equation above is obtained by applying Definition 3 and Theorem 1 in [76] [6]. We are interested in optimization problems of the form:

$$V_T^*(x_0) := \max_{\pi \in \Pi} V_T(x_0, \pi). \tag{2.3.1}$$

In the rest of this section, we prove the optimality of the dynamic programming equation that solves this optimization problem. The state space $\mathbb{X}$ is extended with time variable (i.e., $\mathbb{X} \cup [0, \ldots T]$) to model the time-varying nature of the reward

[6] The policies considered in [76] are deterministic, but we are considering randomized policies.

functions r_t, and action space constraint $A_t(x)$. The extended state space is denoted by $\mathbb{X}'$.

Theorem 1. *Assume the following conditions hold:*

1) $\forall x \in \mathbb{X}$, *the stochastic kernels* $Q_{t,x} : a \to Q_t(\cdot|x,a)$ *are continuous;*

2) The one-step dynamic reward measure $\{\rho_t\}_{t=0}^{T-1}$ *is Markov (see Definition 15), and there exists a sequence of corresponding reward transition mappings* $\sigma_t : m \to \sigma(\psi, x, m),\ t = 0, \ldots, T-1$ *that are upper semi-continuous;*

3) The functions $\{r_t(\cdot,\cdot,\cdot)\}_{t=0}^{T-1}$ *are b-bounded, measurable, and* $a \to r_t(\cdot, a, \cdot)$ *is upper semi-continuous;*

4) For every $x \in \mathbb{X}$ *and* $t \in [0, \ldots, T-1]$ *the set* $A_t(x)$ *is compact;*

5) The function $r_T(\cdot)$ *is b-bounded and measurable;*

Then a maximizer for the dynamic programing equation:

$$v_t(x) = \max_{\delta \in \mathcal{P}(A(x))} \sigma_t \left(r_t(x,\cdot,\cdot) + v_{t+1}(\cdot), x, \delta \circ Q_{t,x} \right)$$

$$v_T(x) = r_T(x) \quad x \in \mathbb{X},\ t = 1, \ldots, T-1, \tag{2.3.2}$$

exists. Furthermore, an optimal policy, $\pi^* := \left\{\pi_0^*, \pi_1^*, \cdots \pi_{T-1}^*\right\}$ exists and each $\pi_t^*(x)$ is a maximizer for the right-hand side of Eq. *2.3.2 at time t for all* $x \in \mathbb{X}$*; In addition, every measurable solution of Eq. 2.3.2 at time 0,* v_0*, is an optimal solution for Eq. 2.3.1.*

Proof. Let π denote an arbitrary randomized policy and $\{x_0^\pi, a_0^\pi, \ldots, x_T^\pi\}$, the resulting state-action trajectory of the controlled Markov process. We denote the

reward-to-go function by :

$$\begin{aligned} R_t(x,\pi) :&= \rho_t\left(r(x_t^\pi, a_t^\pi, x_{t+1}^\pi)+\right.\\ &\quad \rho_{t+1}\left(r(x_{t+1}^\pi, a_{t+1}^\pi, x_{t+2}^\pi)+\cdots+\right.\\ &\quad \left.\left.\rho_{T-1}\left(r(x_{T-1}^\pi, a_{T-1}^\pi, x_T^\pi)+r_T(x_T^\pi)\right)\cdots\right)\right)\\ R_T(x,\pi) :&= r_T(x). \end{aligned}$$

This is the total reward from time t onwards when the policy π is applied at the initial state x. In particular, we know

$$V_T(x,\pi) = R_0(x,\pi).$$

We first prove that a solution to Eq. 2.3.2 exists. By assumption 1, $Q_{t,x}$ is a continuous stochastic kernel, which implies $\delta \circ Q_{t,x} : \mathcal{P}(\mathbb{A}) \to \mathcal{M}$ is continuous in δ. Here, $Q_{t,x} : \mathbb{A} \to \mathcal{P}(\mathbb{X})$ is the stochastic kernel parameterized by x at time t (see Definition 7). By assumption 2, we know that $\delta \to \sigma(\psi, x, \delta \circ Q_{t,x})$ is upper semi-continuous in δ. Assumptions 3, 4, 5 imply that the set $\mathcal{P}(A(x))$ is weakly-compact, hence a maximizer exists for Eq. 2.3.2.

We denote an optimal policy by

$$\pi^* = \left\{\pi_0^*, \ldots, \pi_{T-1}^*\right\},$$

where $\pi_t^*(x)$ is a maximizer for Eq. 2.3.2 for all $x \in \mathbb{X}$. We need to show that for

$t = 0, \ldots T$,

$$R_t(x, \pi) \leq v_t(x), \tag{2.3.3}$$

and with equality if $\pi = \pi^*$, i.e.,

$$R_t(x, \pi^*) = v_t(x). \tag{2.3.4}$$

In particular, if Eq. 2.3.3 is true, we have $V_T(x, \pi) = R_0(x, \pi) \leq v_0(x)$ and $V_T(x, \pi^*) = R_0(x, \pi^*) = v_0(x)$, which prove the statement regarding $v_0(x)$ being the solution for the optimization problem stated in Eq. 2.3.1.

We show Eq. 2.3.3 to be true by backward induction. We first note the fact that

$$R_T(x, \pi) = v_T(x) = r_T(x).$$

Assuming the induction hypothesis that for some $t = T - 1, \ldots, 0$,

$$R_{t+1}(x, \pi) \leq v_{t+1}(x), \ x \in \mathbb{X},$$

the reward-to-go equation at time t satisfies the following inequalities:

$$\begin{aligned} R_t(x, \pi) &= \rho_t \left(r(x, a_t^\pi, x_{t+1}^\pi) + R_{t+1}(x_{t+1}^\pi, \pi) \right) \\ &\leq \rho_t \left(r(x, a_t^\pi, x_{t+1}^\pi) + v_{t+1}(x_{t+1}^\pi) \right) \\ &\leq \max_{\delta \in \mathcal{P}(A(x))} \sigma_t \left(r_t(x, \cdot, \cdot) + v_{t+1}(\cdot), x, \delta \circ Q_x \right) \\ &:= v_t(x). \end{aligned}$$

The second line in the equation above is due to part 1 of Assumption 1 (i.e., monotonicity) and the induction hypothesis. The third line in the equation above is true by the virtue of ρ_t being Markov (the second assumption of this theorem). This proves Eq. 2.3.3. If we assume $R_{t+1}(x, \pi^*) \geq v_{t+1}(x),\ x \in \mathbb{X}$, then we conclude $R_t(x, \pi^*) \geq v_t(x)$ using a similar induction argument as above, which proves Eq. 2.3.4. It should be easy to see that $R_{T-1}(x, \pi^*) = v_{T-1}(x)$, since π^* is in Π by definition. Repeating the same steps as above for $T-2, T-3, \ldots, 0$, we obtain the desired result

$$V_T^*(x) = V_T(x, \pi^*) = R_0(x, \pi^*) = v_0(x).$$

□

2.3.1.1 Application: Cumulative Prospect Theory Measures

We assume that we are given a one-step dynamic reward measure of the form in Eq. 2.2.6, where $u^+(x) = x$ and $u^-(x) = x$. We would like to evaluate the performance of the random variable ψ_x at each time t, assuming the dynamic reward measure is Markov, and the following reward transition mapping satisfies Eq. 2.2.7:

$$\begin{aligned}\sigma_t(\psi_x, x, m) = &\int_0^\infty w^+\left(m\left((\psi_x)_+ > s\right)\right) ds \\ &- \int_0^\infty w^-\left(m\left((\psi_x)_- > s\right)\right) ds, \end{aligned} \tag{2.3.5}$$

where ψ is a $\mathcal{B}(\mathbb{K} \times \mathbb{X})$-measurable random variable (e.g., $\psi = r + v$), and $m \in \mathcal{M}$. We denote the function $\psi(x, \cdot, \cdot)$ by $\psi_x \in \mathcal{L}_{t+1}$, which is a $\mathcal{B}(\mathbb{X} \times \mathbb{A})$-measurable

random variable. $w^+ : [0,1] \to [0,1]$ and $w^- : [0,1] \to [0,1]$ are two continuous monotonically non-decreasing functions.

We would like to apply Theorem 1 to Eq. 2.2.6, given the assumption that it is Markov, by proving that Eq. 2.3.5 is a reward transition mapping.

Theorem 2. *σ_t defined by equation 2.3.5 is a reward transition mapping. Furthermore, it is continuous in m.*

Proof. First we need to show that Eq. 2.3.5 satisfies the two properties in Definition 14.

1) prove: if $\phi_x \leq \psi_x$ then $\sigma(\phi_x) \leq \sigma(\psi_x),\ \forall \phi_x, \psi_x \in \mathcal{V}$;

We need to break σ_t into two parts, $\int_0^\infty w^+\left(m\left((\psi_x)_+ > s\right)\right) ds$ and

$\int_0^\infty w^-\left(m\left((\psi_x)_- > s\right)\right) ds$.

We first look at $\int_0^\infty w^+\left(m\left((\psi_x)_+ > s\right)\right) ds$. Since $\phi_x \leq \psi_x$, we have

$m\left((\phi_x)_+ > s\right) \leq m\left((\psi_x)_+ > s\right)$. Using the fact that w^+ is a monotonically non-decreasing function,

we have $w^+\left(m\left((\phi_x)_+ > s\right)\right) \leq w^+\left(m\left((\psi_x)_+ > s\right)\right)$, which implies

$$\int_0^\infty w^+\left(m\left((\phi_x)_+ > s\right)\right) ds \leq \int_0^\infty w^+\left(m\left((\psi_x)_+ > s\right)\right) ds.$$

Similarly, we have $m\left((\phi_x)_- > s\right) \geq m\left((\psi_x)_- > s\right)$, using the fact the w^- is a monotonically non-decreasing function, we have

$$w^-\left(m\left((\phi_x)_- > s\right)\right) \geq w^-\left(m\left((\psi_x)_- > s\right)\right),$$

which implies

$$\int_0^\infty w^- \left(m\left((\phi_x)_- > s\right)\right) ds \geq \int_0^\infty w^- \left(m\left((\psi_x)_- > s\right)\right) ds.$$

Conclusion 1 follows from the previous inequality.

2) prove: $\sigma(\beta\phi_x) = \beta\sigma(\phi_x),\ \forall\phi_x \in \mathcal{V},\ \beta \geq 0$;

Since

$$\int_0^\infty w^+ \left(m\left((\beta\psi_x)_+ > s\right)\right) ds - \int_0^\infty w^- \left(m\left((\beta\psi_x)_- > s\right)\right) ds$$

$$= \int_0^\infty w^+ \left(m\left((\psi_x)_+ > \frac{s}{\beta}\right)\right) ds - \int_0^\infty w^- \left(m\left((\psi_x)_- > \frac{s}{\beta}\right)\right) ds,$$

we do a change of variable with $z = \frac{s}{\beta}$. Rewriting the equation in terms of z we have

$$\beta\left(\int_0^\infty w^+ \left(m\left((\psi_x)_+ > z\right)\right) dz - \int_0^\infty w^- \left(m\left((\psi_x)_- > z\right)\right) dz\right).$$

To prove the continuity of σ_t, we explicitly prove that

$$\int_0^\infty w^+ \left(m\left((\beta\psi_x)_+ > s\right)\right) ds$$

is continuous in m, because the proof for the second part of the equation will follow similarly. We prove the continuity of σ_t by appealing to the fact that the sum of two continuous functions is continuous.

We denote the Prokhorov metric (see [16, Section 2.1]) by:

$$d(\mu,\nu) := \inf\left\{\epsilon | \mu(A) \leq \nu(A^{\epsilon}) + \epsilon,\ \nu(A) \leq \mu(A^{\epsilon}) + \epsilon \right.$$
$$\left. \forall A \in \mathcal{B}(\mathbb{X} \times \mathbb{A})\right\}.$$

For the purpose of readability, we define the function

$$f_{\psi_x}^{\mu,\nu}(s) = \left| w^{+}\left(\nu\left(\left(\psi_x\right)_{+} > s\right)\right) - w^{+}\left(\mu\left(\left(\psi_x\right)_{+} > s\right)\right)\right|,$$

and its associated sets

$$B_{\delta_1} = \left\{s : s \in [0, M],\ f_{\psi_x}^{\mu,\nu}(s) \leq \delta_1\right\}, \text{ and}$$
$$\bar{B}_{\delta_1} = \left\{s : s \in [0, M],\ f_{\psi_x}^{\mu,\nu}(s) > \delta_1\right\}.$$

Since the total reward is the sum of a finite number (i.e., finite-horizon) of per-stage rewards, it is bounded by $M \in \mathbb{R}$.

Given an arbitrary $\epsilon > 0$, we need to find a δ_1 such that it satisfies the following

equations:

$$\left|\int_0^M w^+\left(\nu\left(\left(\psi_x\right)_+ > s\right)\right) ds - \int_0^M w^+\left(\mu\left(\left(\psi_x\right)_+ > s\right)\right) ds\right|$$
$$\leq \int_0^M \left|w^+\left(\nu\left(\left(\psi_x\right)_+ > s\right)\right) - w^+\left(\mu\left(\left(\psi_x\right)_+ > s\right)\right)\right| ds \leq$$
$$\int_0^M 1_{B_{\delta_1}} \delta_1 ds + \int_0^M 1_{\bar{B}_{\delta_1}} ds =$$
$$\delta_1 \times M \times \int_0^M \frac{1_{B_{\delta_1}}}{M} ds + M \times \int_0^M \frac{1_{\bar{B}_{\delta_1}}}{M} ds =$$
$$\delta_1 \times M \times \int_0^M \frac{1_{B_{\delta_1}}}{M} ds + M \times \left(1 - \int_0^M \frac{1_{B_{\delta_1}}}{M} ds\right) \leq$$
$$2 \times \delta_1 \times M = \epsilon.$$

By letting $1 > \delta_1 = \frac{\epsilon}{2M} > 0$, the last line in the equation above holds.

Our goal is to prove that given an arbitrary δ_1, there always exists a $\delta \leq \delta_1$ such that all ν in the δ-neighborhood of μ satisfy the following

$$\int_0^M \frac{1_{B_{\delta_1}}}{M} ds \geq 1 - \delta_1,$$

which implies

$$\delta_1 \times M \times \int_0^M \frac{1_{B_{\delta_1}}}{M} ds + M \times \left(1 - \int_0^M \frac{1_{B_{\delta_1}}}{M} ds\right) \leq \epsilon.$$

Since w^+ is continuous, for any $\delta_1 > 0$ there exists a δ_2 such that

$$\left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \leq \delta_2 \implies$$
$$\left|w^+\left(\nu\left((\psi_x)_+ > s\right)\right) - w^+\left(\mu\left((\psi_x)_+ > s\right)\right)\right| \leq \delta_1.$$

Hence, for any $\delta_1 > 0$ we can always find a δ_2 such that

$$\int_0^M \frac{1_{\left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \leq \delta_2}}{M} ds \geq 1 - \delta_1 \implies$$
$$\int_0^M \frac{1_{B_{\delta_1}}}{M} ds \geq 1 - \delta_1. \tag{2.3.6}$$

From the Markov inequality, we have

$$\int_0^M 1_{\left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \leq \delta_2} \frac{1}{M} ds \geq$$
$$1 - \frac{\int_0^M \left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \frac{1}{M} ds}{\delta_2}. \tag{2.3.7}$$

Next, we need to find a δ such that the following equations hold:

$$\int_0^M \left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \frac{1}{M} ds \leq$$
$$\int_0^M d(v, \mu)^2 \frac{1}{M} ds \leq \delta^2 = \delta_1 \times \delta_2.$$

Finally, letting $\delta := +\sqrt{\delta_1 \times \delta_2}$, it is true that for any v in the δ-neighborhood of μ,

the following equations hold:

$$\int_0^M \left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \frac{1}{M} ds \leq \delta_1 \times \delta_2$$

$$\implies 1 - \frac{\int_0^M \left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \frac{1}{M} ds}{\delta_2} \geq$$

$$1 - \frac{\delta_1 \times \delta_2}{\delta_2} = 1 - \delta_1$$

$$\implies \int_0^M 1_{\left|\nu\left((\psi_x)_+ > s\right) - \mu\left((\psi_x)_+ > s\right)\right| \leq \delta_2} \frac{1}{M} ds \geq 1 - \delta_1$$

$$\implies \int_0^M \frac{1_{B_{\delta_1}}}{M} ds \geq 1 - \delta_1 \text{ (Eq. 2.3.6).}$$

The second implication is due to Eq. 2.3.7 and the third implication is due to Eq. 2.3.6. The second assertion of the theorem is proved. □

Below is an example where we use a Markov CPT dynamic reward measure.

Example 5. The following example attempts to explain why people become entrepreneurs. We assume a person could be in several states {poor, middle, upper-middle, super-rich}. If one decides to become an entrepreneur, one has the following transition probability matrix.

	poor	middle	upper-middle	super-rich
poor	.999	0	0	.001
middle	.999	0	0	.001
upper-middle	.999	0	0	.001
super-rich	.001	0	0	.999

Table 2.3.1: Transition probability matrix for becoming an entrepreneur

One could also choose to pursuit a normal job with the following transition

probabilities.

	poor	middle	upper-middle	super-rich
poor	0	1	0	0
middle	0	0	1	0
upper-middle	0	0	1	0
supper-rich	0	0	0	1

Table 2.3.2: Transition probability matrix for taking a normal job

The action space is {entrepreneur (E), normal (N)}. We define the random variable x_t to represent the current state of the controlled Markov process:

$$x_t(\omega) = \begin{cases} 1 & \omega = \text{poor} \\ 2 & \omega = \text{middle} \\ 3 & \omega = \text{upper-middle} \\ 4 & \omega = \text{super-rich} \end{cases}.$$

In this example, the per-stage reward function is given as:

$$r(x,a) := \begin{cases} x - 1/x & x \leq 3 \text{ and } a = E \\ x & x \leq 3 \text{ and } a = N \\ 100 - 1/x & x > 3 \text{ and } a = E \\ 100 & x > 3 \text{ and } a = N \end{cases},$$

and the terminal reward function is:

$$r_2(x) := \begin{cases} x & x \leq 3 \\ & \\ 100 & x > 3 \end{cases}.$$

We want to solve the optimization problem stated in Eq. 2.3.1 given a dynamic reward measure of the form in Eq. 2.2.6[7]:

$$\rho_t(Z_{t+1}) = \int_0^\infty w^+(P\left((Z_{t+1}|x_t) > s\right))ds,$$

with the nonlinear weighting function $w^+(F)$:=$e^{-\delta(-ln(F))^\gamma}$, where $0 < \gamma < 1$ and $\delta > 0$. For the purpose of this example, we take γ to be 0.9 and δ to be 0.5. Since we are only dealing with positive rewards in this example, w^- and u_- need not be given. Furthermore, we note that the dynamic reward measure is Markov and has a sequence of transition mappings σ_t of the form:

$$\sigma_t(r + v_{t+1}, x_t, \lambda \circ Q_{x_t}) =$$
$$\int_0^\infty w^+\left(\lambda \circ Q_{x_t}\left(r + v_{t+1}(x_{t+1}) > s\right)\right) ds.$$

The table below shows the value function at times 0 and 1 by applying the

[7]For simplicity, we dropped u_+ and u_-.

dynamic programming equation

$$v_t(x) = \max_{p_E \in [0,1]} \left\{ \int_0^\infty w^+ \left(Q_{x,E} \left(r(x,E) + v_{t+1}(x_{t+1}) > s \right) p_E \right. \right.$$

$$\left. \left. + Q_{x,N} \left(r(x,N) + v_{t+1}(x_{t+1}) > s \right) (1 - p_E) \right) ds \right\},$$

where $v_2 = r_2$ and the variable p_E is the probability of becoming an entrepreneur.

Time=1

x_1	p_E	$v_1(x_1)$
poor	0.850471	7.1229
middle	0.787238	8.81843
upper-middle	0.808817	9.91617
super-rich	0	200

Time = 0

x_0	p_E	$v_0(x_0)$
poor	0.919031	18.6786
middle	0.886583	20.3504
upper-middle	0.896001	21.4663
super-rich	0	300

Table 2.3.3: An optimal solution for Ex. 5 (a value function and an optimal policy) at time 0 and 1.

Since Table 2.3.3 above shows the likelihood of becoming an entrepreneur is higher if one is younger, it agrees with our intuition that one should pursue entrepreneurship while still young. For example, an individual, starting out poor, should be entrepreneurial almost 92% of the time; On the other hand, if the same individual is a year older, he or she should only be entrepreneurial 85% of the time. This result also agrees with our tendency to become more risk-averse as we grow older.

Our approach yields an optimal randomized policy, which is different from

the standard approach (see Eq. 2.2.3), where an optimal solution is deterministic. Non-convex reward measures are useful for modeling many real-life problems. More specifically, CPT-inspired reward measures are derived from experimental data and have been proven to model several key characteristics of human behavior well. In this section, we proved the optimality of dynamic programming equations for the optimization problem described by Eq. 2.3.1. In addition, we provided a numerical example demonstrating the intuitiveness of the optimal policies obtained. In the next section, we will apply dynamic programming to infinite-horizon MDPs with non-convex reward measures.

2.3.2 Discounted Infinite-Horizon

As in the finite-horizon case, we assume that we are given a time-consistent dynamic reward measure $\{\rho_{t,\infty}\}_{t=0}^{\infty}$ and its corresponding time-invariant one-step dynamic reward measure $\{\rho\}$. Here, ρ does not depend on t anymore. From Definition 3 and Theorem 1 in [76], we write the corresponding value function starting at x_0 with a control policy π and the resulting state-action trajectory $\{x_0^\pi, a_0^\pi, \ldots, x_T^\pi\}$ as:

$$\begin{aligned} V(x_0,\pi) = \rho\left(r(x_0^\pi, a_0^\pi, x_1^\pi) + \beta\rho\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right. \\ +\beta\rho\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right. \\ \left.\left.\beta\rho\left(r(x_\infty^\pi, a_\infty^\pi, x_\infty^\pi)\right)\cdots\right)\right)\right). \end{aligned}$$

We would like to consider the following optimization problem with $\beta \in (0,1)$:

$$V^*(x_0) := \max_{\pi \in \Pi} V(x_0, \pi). \tag{2.3.8}$$

In this section, we are interested in the case when $r : \mathbb{K} \times \mathbb{X} \to \mathbb{R}$ is bounded, i.e., $\exists \bar{r} \in \mathbb{R}^+$ such that $|r| \leq \bar{r}$. We assume r to be a non-positive valued function. The non-negative case can be argued by symmetry. We denote the t-stage-reward function resulting from applying a policy π by:

$$\begin{aligned} J_t(x_0, \pi) = \rho\left(r(x_0^\pi, a_0^\pi, x_1^\pi) + \beta\rho\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right. \\ +\beta\rho\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right. \\ \left.\left.\beta\rho\left(r(x_{t-1}^\pi, a_{t-1}^\pi, x_t^\pi)\right)\cdots\right)\right). \end{aligned} \tag{2.3.9}$$

Since $r \leq 0$ and $V_t \leq V_{t-1} \leq 0$, by the Monotone Convergence Theorem, we may write

$$V(x_0, \pi) = \lim_{t \to \infty} J_t(x_0, \pi), \ \forall \pi \in \Pi.$$

From Section 2.3.1, we know the solution to Eq. 2.3.9 can be obtained by iterating the following equation:

$$v_t(x) := \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r(x, \cdot, \cdot) + \beta v_{t-1}(\cdot), x, \delta \circ Q_x\right).$$

In other words,

$$v_t(x) = \max_{\Pi} \ J_t(\pi, x), \ \forall x \in \mathbb{X}.$$

The equation above is the backward form of the finite-horizon dynamic programming equation; this is different from the forward equation we used in the previous section. In the finite-horizon case, we are able to prove the optimality of the dynamic programming equation (i.e., Eq. 2.3.2) by backward induction; this approach will not suffice in the infinite-horizon case. In the infinite-horizon case, we need to appeal to the Banach fixed-point theorem to prove the existence of a measurable function v^* such that $v^* = Tv^*$. Lastly, we need to prove that the solution v^* is indeed equal to

$$V^*(x) := \max_{\Pi} V(x, \pi), \ \forall x \in \mathbb{X}.$$

We used [48, 49] as the main technical references for the proofs below.

Definition 16. Let $\mathbb{M}(\mathbb{X})^-$ denote the cone of non-positive measurable functions on $\mathbb{X}$. For every $v \in \mathbb{M}(\mathbb{X})^-$, Tv is defined as a mapping from $\mathbb{X}$, i.e.,

$$Tv(x) := \max_{\delta \in \mathcal{P}(A(x))} \sigma(r_x + \beta v, x, \delta \circ Q_x), \ \forall x \in \mathbb{X},$$

where r_x is the reward function with x held fixed, i.e., $r_x := r(x, \cdot, \cdot)$.

The existence of a measurable selector[8] is important in proving the optimality of the dynamic programming equation.

Lemma 1. *Assuming the function*

$$\sigma(r_x + \beta v, x, \delta \circ Q_x)$$

[8]See appendix for the definition of a measurable selector.

is upper semi-continuous (u.s.c) in δ, r is a non-positive valued function, and $\mathcal{P}(A(x))$ is compact-valued, then T maps $\mathbb{M}(\mathbb{X})^{-}$ into itself, i.e., for every v in $\mathbb{M}(\mathbb{X})^{-}$, Tv is also in $\mathbb{M}(\mathbb{X})^{-}$, and moreover, there exists a measurable selector $\psi : \mathbb{X} \to \mathcal{P}(\mathbb{A})$ with $\psi(x) \in \mathcal{P}(A(x))$ such that

$$\sigma(r_x + \beta v, x, \psi(x) \circ Q_x) = \max_{\delta \in \mathcal{P}(A(x))} \sigma(r_x + \beta v, x, \delta \circ Q_x), \ \forall x \in \mathbb{X}.$$

Proof. This follows from Proposition 6 in the appendix. □

The lemma above is important for ensuring the value function is measurable. The reader might notice that ψ can be used to construct a stationary policy $\pi = \{\psi, \psi, \dots\}$, which will be used to prove the optimality of stationary Markov polices. The following Lemma is used in the upcoming theorem.

Lemma 2. *If $u \in \mathbb{M}(\mathbb{X})^{-}$ is such that $u \leq Tu$, and r is a non-positive valued function, then $u \leq V^{*}$.*

Proof. Assuming $u \leq Tu$ and using Lemma 1, we write the following inequality:

$$u(x) \leq \sigma(r(x, \cdot, \cdot) + \beta u, x, \psi(x) \circ Q_x).$$

Iterating this inequality, we obtain

$$u(x) \leq \sigma\left(r(x, a_0^\pi, x_1^\pi) + \beta\sigma\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right.$$
$$+ \beta\sigma\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right.$$
$$\left.\left.\beta\sigma\left(r(x_{n-1}^\pi, a_{n-1}^\pi, x_n^\pi) + \beta u\left(x_n^\pi\right)\right)\cdots\right)\right), \ \forall n \geq 1, \ x \in \mathbb{X}, \quad (2.3.10)$$

where $\pi = \{\psi, \psi, \dots\}$. In the inequality above, we used the short-hand notation

$$\sigma\left(r + \beta u\right) := \sigma\left(r + \beta u, x, \psi\left(x\right) \circ Q_x\right),$$

and $\{x_t^\pi\}$ is the resulting process from applying the policy π. Applying the fact that

$$\beta u\left(x_n^\pi\right) \leq 0$$

to Eq. 2.3.10, we conclude the following inequality:

$$u(x) \leq \sigma\left(r(x, a_0^\pi, x_1^\pi) + \beta\sigma\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right.$$
$$+ \beta\sigma\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right.$$
$$\left.\left.\beta\sigma\left(r(x_{n-1}^\pi, a_{n-1}^\pi, x_n^\pi)\right)\cdots\right)\right), \ \forall n \geq 1, \ x \in \mathbb{X}.$$

By letting $n \to \infty$, the inequality above yields

$$u\left(x\right) \leq V\left(x, \pi\right) \leq V^*\left(x\right) \ \ \forall x \in \mathbb{X}.$$

□

Theorem 3. *Assume the following conditions hold:*

1) The stochastic kernels $Q_x : a \to Q(\cdot|x,a)$ are continuous $\forall x \in \mathbb{X}$;

2) The one-step dynamic reward measure ρ is Markov (see Definition 15), and a sequence of corresponding reward transition mappings $\sigma : m \to \sigma(\psi, x, m)$ is upper semi-continuous;

3) The function $r(\cdot,\cdot,\cdot)$ is bounded, measurable, and $a \to r(\cdot,a,\cdot)$ is upper semi-continuous in a;

4) For every $x \in \mathbb{X}$ the set $A(x)$ is compact;

5) $\beta \in (0,1)$.

Then a maximizer for the dynamic programing equation

$$v(x) = \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r(x,\cdot,\cdot) + \beta v(\cdot), x, \delta \circ Q_x\right) \ \forall x \in \mathbb{X}, \tag{2.3.11}$$

exists. Furthermore, an optimal policy, $\pi^* := \{\psi^*, \psi^* \cdots\}$ exists and each ψ^* is a maximizer for the right-hand side of Eq. *2.3.11. In addition, every bounded measurable solution of Eq. 2.3.11 is an optimal solution for Eq. 2.3.8.*

Proof. Since $A(x)$ is compact for every x, we know $\mathcal{P}(A(x))$ is also compact-valued. We want to show that the operator

$$Tv := \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r + \beta v, x, \delta \circ Q_x\right)$$

is a contraction on the space of bounded measurable functions endowed with the

supremum norm

$$\|v\| := \sup_{\mathbb{X}} |v| .$$

We first prove that a solution to Eq. 2.3.11 exists. By assumption 1, Q_x is a continuous stochastic kernel, which implies $\delta \circ Q_x : \mathcal{P}(\mathbb{A}) \to \mathcal{M}$ is continuous in δ. Here, $Q_x : \mathbb{A} \to \mathcal{P}(\mathbb{X})$ is the stochastic kernel parameterized by x at time t (see Definition 7). By assumption 2, we know that $\delta \to \sigma(\psi, x, \delta \circ Q_x)$ is upper semi-continuous in δ. Assumptions 3 and 4 imply that the set $\mathcal{P}(A(x))$ is weakly-compact; hence a maximizer exists for Eq. 2.3.11. This also proves the existence of $\pi^* := \{\psi^*, \psi^*, \cdots\}$.

Next, the mapping $T : \mathbb{M}(\mathbb{X})^- \to \mathbb{M}(\mathbb{X})^-$ satisfies:

$$\begin{aligned}
Tv :&= \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r + \beta v, x, \delta \circ Q_x\right) \\
&= \sigma\left(r + \beta v, x, \psi\left(x\right) \circ Q_x\right) \\
&= \sigma\left(r + \beta\left(v' + (v - v')\right), x, \psi\left(x\right) \circ Q_x\right) \\
&\leq \sigma\left(r + \beta v', x, \psi\left(x\right) \circ Q_x\right) + \beta \sup_{\mathbb{X}} |v - v'| \\
&\leq \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r + \beta v', x, \delta \circ Q_x\right) + \beta \sup_{\mathbb{X}} |v - v'| \\
&= Tv' + \beta \sup_{\mathbb{X}} |v - v'| \implies \\
Tv - Tv' &\leq \beta \sup_{\mathbb{X}} |v - v'| \ \ \forall x \in \mathbb{X} \implies \\
\sup_{\mathbb{X}} |Tv - Tv'| &\leq \beta \sup_{\mathbb{X}} |v - v'| ,
\end{aligned}$$

where ψ is an optimal measurable selector and its existence is ensured by Lemma 1.

Hence, by appealing to Banach's Fixed-Point Theorem for contraction mappings and assumptions 3 and 5 of this theorem, we conclude there exists a unique function $v^* \in \mathcal{V}$ such that $Tv^* = v^*$.

Finally, we need to prove that v^* is a measurable solution to Eq. 2.3.8. We need to prove this fact in two steps: $v^* \leq V^*$ and $v^* \geq V^*$. From the fact that $Tv^* = v^*$ and Lemma 2 we conclude that $v^* \leq V^*$. To prove the inequality in the other direction, we know from the finite-horizon case with the reward-to-go function denoted by R_n, we have

$$v_n \geq R_n(x, \pi) \geq V(x, \pi), \ \forall n \in \{0, 1, 2, \dots\}, \ \forall n, \ \forall \pi \in \Pi, \ \forall x \in \mathbb{X},$$

which implies

$$v_n \geq R_n(x, \pi^*) \geq V^*(x) \ \forall n, \ \forall x \in \mathbb{X},$$

where $\pi^* = \{\psi^*, \psi^*, \dots\}$ is constructed using the maximizer, ψ^*, of Eq. 2.3.11. The operator T is monotone, i.e., if u and u' are functions in $\mathbb{M}(\mathbb{X})^-$ and $u \geq u'$, then $Tu \geq Tu'$. Since $v_0 := 0$ and $v_n := Tv_{n-1}$ for $n \geq 1$, v_n form a non-increasing sequence in $\mathbb{M}(\mathbb{X})^-$ converging to some function $v^* \in \mathbb{M}(\mathbb{X})^-$. Since $v_n \downarrow v^*$ due to the monotone convergence theorem, and $v_n \geq V^*$, we conclude that $v^* \geq V^*$. The desired conclusion is reached given the fact the policy $\pi^* = \{\psi^*, \psi^*, \dots\} \in \Pi$. $\square$

If the per-stage reward is both positive and negative, we defer it to the later transient case; as the discounted infinite-horizon problem can be rewritten into a transient problem by adding an absorbing state. The requirement that the reward

function, r, be a bounded measurable function can be relaxed in the previous theorem. Of course, the proof for the relaxed case will be different. Due to space limitations, we do not explore alternatives with the relaxed assumption on the reward function here. For the interested reader, the proof for the existence of an optimal policy with the relaxed assumption on the reward function can be found in [48] for the standard expected value measure, which can be adapted for our case.

We present a numerical example, similar to the finite-horizon case, which will demonstrate the type of policies expected from the CPT-based reward measures. In the example below, we also compare an optimal solution of the expected value (standard reward measures) with that of the CPT-based reward measures.

Example 6. As in the finite-horizon case of Example 5, its infinite discounted counterpart tries to explain why people become entrepreneurs. The major difference here is we are no longer given a terminal reward function. The transition probabilities and the per-stage reward function used for this numerical example can be found from Example 5.

We calculate the discounted infinite-horizon counterpart with the nonlinear weighting function $w^+(F) := e^{-\delta(-ln(F))^\gamma}$, where $0 < \gamma < 1$ and $\delta > 0$. For the purpose of this example, we take γ to be 0.9 and δ to be 0.5. Furthermore, we assume the discount factor, β, to be 0.5.

The tables below summarize our numerical results. We notice that the value function given by the CPT measures is higher than that of the expected value. In addition, using CPT measures will yield a more risk-seeking optimal policy.

x_0	P(entrepreneur)	$v(x_0)$
poor	0	3.49999
middle	0	4.99999
upper-middle	0	5.99999
super-rich	0	200

(a) Expected Value

x_0	P(entrepreneur)	$v(x_0)$
poor	0.868488	11.5553
middle	0.861585	13.0742
upper-middle	0.881884	14.1858
super-rich	0	200

(b) CPT Expected Value

Table 2.3.4: An optimal solution for Ex. 6 (a value function and an optimal policy) .

Table 2.3.4 is similar to Table 2.3.3 in the finite-horizon case in the sense that they both produce randomized polices.

This example suggests that a middle-class person should be the least likely to pursue entrepreneurship. On the other hand, an upper-middle class person is mostly likely to start his/her own business. However, the difference in the probability of entrepreneurship for the states poor, middle and upper-middle is very small, which suggests in the long run we should all be entrepreneurial regardless of our current state unless one is already super-rich. Of course, in practice we need to calibrate the underlying Markov model with empirical data.

In the next section, we will examine the suitability of the dynamic programming method for the transient Markov control model case.

2.3.3 Transient Markov Control Model

In this section, we prove the optimality of the dynamic programming equation for transient Markov control models. Our main technical references are [18], [55] and [49]. Transient Markov control models require further specification in addition to the definitions provided in Section 2.2.1. Before we introduce the definition of a transient Markov control model, we need to define a few notations first. Given a norm weight function $w : \mathbb{X} \to [1, \infty)$, the w weighted-norm is denoted by $\|\cdot\|_w$. It is calculated for a substochastic kernel A as:

$$\|A\|_w = \sup_{\mathbb{X}} \frac{1}{w(x)} \int_{\mathbb{X}} w(y) A(dy|x).$$

Similarly, a w-norm can be defined for a measurable function $v : \mathbb{X} \to \mathbb{R}$ as:

$$\|v\|_w = \sup_{x \in \mathbb{X}} \frac{|v(x)|}{w(x)}.$$

It is the standard operator norm in the space $\mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$, of measurable functions v such that $\|v\|_w < \infty$. The reader can refer to [48] for a more complete discussion on weighted norms. At this point, the reader may be confused by the three functions w, w^+, and w^-. The first function is used in defining a weighted norm, and the latter two functions are used in CPT-based measures. The function used should be clear from the context.

Assumption 2. *The function $w \in \mathcal{V}$ (i.e., the integrable function space) is fixed*

with respect to the given Markov control model such that

$$\|Q_\psi\|_w < \infty,\ \forall \psi \in \Psi.$$

Furthermore, the per-stage reward function $r(\cdot,\cdot,\cdot)$ is measurable, w-bounded, i.e., there is a constant $\bar{r} \geq 0$ such that

$$\sup_{A(x)} |r\left(x, a, x'\right)| \leq \bar{r} w\left(x'\right),\ \forall x, x' \in \mathbb{X},$$

and $r : a \to r(\cdot, a, \cdot)$ is upper semi-continuous in a.

The assumption above is assumed to hold throughout this section. A transient Markov model has some absorbing state $x_A \in \mathbb{X}$, such that $Q\left(\{x_A\} | x_A, a\right) = 1$ and $r\left(x_A, a, x_A\right) = 0$ for all $a \in A\left(x\right)$. In other words, once an absorbing state is reached, no further rewards will be given. In addition, a transient Markov model reaches its absorbing state in finite amount of time, i.e.,

$$\sup_{\Pi, \mathbb{X}} \mathbb{E}\left[\tau_0^\pi | x\right] < \infty, \text{ where } \tau_0^\pi := \inf\left\{t \geq | x_t^\pi = x_A\right\}.$$

Without loss of generality, we assume the model only has one absorbing state, because the case of multiple absorbing states can be easily reduced to the single absorbing state case. We introduce some additional notations for clarity. We denote the effective state space by $\tilde{\mathbb{X}} = \mathbb{X} \setminus \{x_A\}$, and the effective controlled substochastic kernel by $\tilde{Q}$. The substochastic kernel $\tilde{Q}$ restricts its arguments to only allow the

effective states (i.e., $\tilde{Q}\left(B|x,a\right)=Q\left(B|x,a\right),\ \forall B\in\mathscr{B}\left(\tilde{\mathbb{X}}\right),\ \forall x\in\tilde{\mathbb{X}},\ \forall a\in A\left(x\right)$). We introduce the definition of a transient Markov control model below.

Definition 17. A randomized Markov policy $\pi\in\Pi^{RM}$ is *transient* with respect to a Markov control model, if there exists a constant k and a weight function $w:\mathbb{X}\rightarrow[1,\infty)$ such that

$$\left\|\sum_{t=0}^{\infty}\tilde{Q}_{\pi}^{t}\right\|_{w}\leq k, \tag{2.3.12}$$

where $Q_{\pi}^{t}:=Q_{0}Q_{1}\ldots Q_{t-1}$ and $Q_{\pi}^{0}\left(\cdot|x\right):=\delta_{x}\left(\cdot\right)$. If the inequality above is uniform for all Markov policies, then the model is called *uniformly transient* (i.e., Eq. 2.3.12 is true for all $\pi\in\Pi^{RM}$).

Since we are working with stationary transition probabilities (i.e., $Q_1=Q_2$), Eq. 2.3.12 implies that $Q\left(\tilde{\mathbb{X}}|x,a\right)\leq 1$ for all $x\in\mathbb{X}$ and $a\in A\left(x\right)$. Eq. 2.3.12 can also be written as:

$$\begin{aligned}\left\|\sum_{t=0}^{\infty}\tilde{Q}_{\pi}^{t}\right\|_{w} &= \sup_{\mathbb{X}}w\left(x\right)^{-1}\sum_{t=0}^{\infty}\int_{\mathbb{X}}w\left(y\right)\tilde{Q}_{\pi}^{t}\left(dy|x\right)\\ &= \sup_{\mathbb{X}}w\left(x\right)^{-1}\sum_{t=0}^{\infty}\mathbb{E}\left[w\left(x_{t}^{\pi}\right)|x\right].\end{aligned}$$

Hence, we can infer from Eq. 2.3.12 that $\mathbb{E}\left[w\left(x_{t}^{\pi}\right)|x\right]\rightarrow 0$ as $t\rightarrow\infty$. Eq. 2.3.12 is also known as the Pliska condition [69]. One major contribution of Çavuş and Ruszczyński [18] is to suggest a generalized version of the Pliska condition for coherent risk measures (i.e., convexity). Since we are interested in non-convex performance measures, we take a different approach to prove the optimality of dynamic programming equations.

We are interested in solving a more general version of the standard expected total reward problem, which searches for an optimal policy that maximizes the following expected value:

$$\mathbb{E}\left[\sum_{t=0}^{\infty} r\left(x_t, a_t, x_{t+1}\right)\right].$$

We know from [48] that we can apply dynamic programming to this problem and obtain an optimal stationary deterministic policy. In this section, we would like to explore the risk-sensitive version of this problem, especially when the conditional reward function ρ is not convex. Our goal is to prove that dynamic programming can still be applied to the non-convex risk-sensitive version of the expected total reward problem.

We are interested in finding the maximum of a total reward function of the form:

$$\begin{aligned} V(x_0, \pi) = \rho\left(r(x_0^\pi, a_0^\pi, x_1^\pi) + \rho\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right. \\ +\rho\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right. \\ \left.\left.\left.\rho\left(r(x_\infty^\pi, a_\infty^\pi, x_\infty^\pi)\right)\cdots\right)\right)\right). \end{aligned}$$

The corresponding optimization problem can be written as:

$$V^*\left(x_0\right) := \max_{\pi \in \Pi} V(x_0, \pi). \tag{2.3.13}$$

Without loss of generality, we restrict ourselves to randomized Markov policies (see

[49], Theorem 9.4.5), i.e.,

$$V^{*}\left(x_{0}\right)=\max_{\pi\in\Pi^{RM}}V\left(x_{0},\pi\right). \tag{2.3.14}$$

To solve Eq. 2.3.14, we start by finding an optimal solution for the simpler case of randomized stationary policies, i.e.,

$$V^{\dagger}\left(x_{0}\right):=\max_{\pi\in\Pi^{RS}}V(x_{0},\pi). \tag{2.3.15}$$

Later, the sufficiency of randomized stationary Markov polices is proven, i.e., the left-hand sides of Equ 2.3.14 and Eq. 2.3.15 are equivalent. We denote the reward-to-go function at time t by:

$$\begin{aligned}R_{t}(x_{0},\pi)=\rho\left(r(x_{t}^{\pi},a_{t}^{\pi},x_{t+1}^{\pi})+\rho\left(r(x_{t+1}^{\pi},a_{t+1}^{\pi},x_{t+2}^{\pi})\right.\right.\\+\rho\left(r(x_{t+2}^{\pi},a_{t+2}^{\pi},x_{t+3}^{\pi})+\cdots+\right.\\\left.\left.\rho\left(r(x_{\infty}^{\pi},a_{\infty}^{\pi},x_{\infty}^{\pi}))\cdots\right)\right)\right),\end{aligned}$$

and the t-stage total reward function by:

$$\begin{aligned}J_{t}(x_{0},\pi)=\rho\left(r(x_{0}^{\pi},a_{0}^{\pi},x_{1}^{\pi})+\rho\left(r(x_{1}^{\pi},a_{1}^{\pi},x_{2}^{\pi})\right.\right.\\+\rho\left(r(x_{2}^{\pi},a_{2}^{\pi},x_{3}^{\pi})+\cdots+\right.\\\left.\left.\rho\left(r(x_{t}^{\pi},a_{t}^{\pi},x_{t+1}^{\pi})\right)\cdots\right)\right)\right).\end{aligned}$$

We denote the optimal t-stage-reward value function by

$$J_t^*(x) := \max_{\pi \in \Pi} J_t(x, \pi),$$

and define the operator $T(\psi)$ on $\mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$ as

$$T(\psi) v(x) := \sigma(r_x + v, x, \psi(x) \circ Q_x).$$

In addition, the *dynamic programming operator* T is denoted by

$$Tv := \max_{\delta \in \mathcal{P}(A(x))} \sigma(r_x + v, x, \delta \circ Q_x).$$

Since σ is monotone, the operator $T(\psi)$ is also monotone, i.e.,

$$v \geq \bar{v} \implies T(\psi) v \geq T(\psi) \bar{v}, \ \forall v, \bar{v} \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X})), \psi \in \Psi.$$

Given a Markov policy $\pi = \{\psi_t\} \in \Pi$ and $T(\pi)^0 = I$, then for $k = 1, 2, \ldots$, the iterated operator $T^k(\pi)$ on $\mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$ is defined by

$$T^k(\pi) := T(\psi_0) T(\psi_1) \cdots T(\psi_{k-1}).$$

We denote the total-reward function with respect to a policy π as $T \to \infty$ by

$$V(x_0, \pi) = \lim_{T \to \infty} J_T(x_0, \pi) \ \forall x \in \tilde{\mathbb{X}}.$$

For the rest of this section, we make the following assumption.

Assumption 3. *The following conditions hold.*

1. *There exists a* $k \geq 1$ *such that* $T^k(\pi)$ *is a contraction mapping for all transient stationary polices* π*; i.e.,*

$$\exists \gamma < 1 \ s.t. \ \left\| T^k(\pi) v - T^k(\pi) \bar{v} \right\|_w \leq \gamma \left\| v - \bar{v} \right\|_w$$
$$\forall v_1, v_2 \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X})), \ \forall \ transient \ \pi \in \Pi^{RS};$$

2. *The reward transition mapping* $\sigma : \psi \to \sigma(\psi, x, m)$ *is continuous;*

3. $V^* \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$;

4. *The Markov control model is uniformly transient.*

Condition 4 in Assumption 3 can be relaxed (see [55]) at the expense of additional assumptions. From condition 3 in Assumption 3, we can trivially deduce the following lemma.

Lemma 3. *If Assumption 3(3) holds, then* $J_t(\pi, x)$ *and* $J_t^*(x)$ *are both* w*-bounded for all* $x \in \mathbb{X}, \ \pi \in \Pi$*,* $t = 1, 2, \ldots$.

Proof. Proof by contradiction: If $J_t(\pi, x)$ and J_t^* are not w-bounded, then $V^* \notin \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$. This contradicts Assumption 3(3). □

The lemma above justifies for writing $J_t(\pi, \cdot)$ and $J_t^*(\cdot)$ as arguments of $\rho(\cdot)$. Assumption 3(1) ensures the convergence of the operator T^k , which is stated in the following lemma.

Lemma 4. *For any transient stationary policy $\pi \in \Pi^{RS}$, if Assumption 3(1) holds, then for any $v \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$*

$$\lim_{k\to\infty} T^k(\pi)\, v(x) = V(x,\pi) = \lim_{k\to\infty} J_k(x,\pi) \;\; \forall x \in \mathbb{X}.$$

Proof. Using Assumption 3(1) and the Banach fixed-point theorem, noting the fact that

$$V(x,\pi) = \lim_{k\to\infty} T^k(\pi)\, 0 < \infty,$$

the proof follows. □

The following theorem proves the optimality criteria for Eq. 2.3.15.

Theorem 4. *Let Assumptions 2 and 3 hold. For a transient Markov controlled model, a Markov reward transition mapping $\sigma(\cdot,\cdot,\cdot)$, and a randomized stationary Markov policy $\pi = \{\psi, \psi, \dots\}$, a bounded measurable function $v : \tilde{\mathbb{X}} \to \mathcal{R}$ (i.e., $\|v\|_w < \infty$) satisfies the equation*

$$\begin{aligned} v(x) &= \sigma(r_x + v, x, \psi(x) \circ Q_x), \;\; x \in \tilde{\mathbb{X}} \\ v(x_A) &= 0, \end{aligned} \tag{2.3.16}$$

if and only if $v(x) = V(\pi, x)$ for all $x \in \mathbb{X}$.

Proof. Let $v(\cdot)$ be a bounded measurable solution of Eq. 2.3.16. Since $\|v\|_w < \infty$ and $w \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$, we know that $v \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$. By assumption, $r(\cdot,\cdot,\cdot)$ is w-bounded, and thus $r_x \in \mathbb{B}_w(\mathbb{X}, \mathscr{B}(\mathbb{X}))$. Consequently, the right-hand side of Eq.

2.3.16 is well defined and can be iterated, which results in the following equation

$$v(x_0) = \rho\left(r(x_0^\pi, a_0^\pi, x_1^\pi) + \rho\left(r(x_1^\pi, a_1^\pi, x_2^\pi)\right.\right.$$
$$+\rho\left(r(x_2^\pi, a_2^\pi, x_3^\pi) + \cdots +\right.$$
$$\left.\left. v\left(x_{T+1}\right)\cdots\right)\right)) \ \forall x_0 \in \tilde{\mathbb{X}}.$$

Since $v\left(\cdot\right)$ is a w-bounded function, we conclude by evoking Lemma 4 and taking k to infinity that

$$\begin{aligned} v\left(x\right) &= \lim_{k\to\infty} T^k\left(\pi\right) v\left(x\right) = V\left(\pi, x\right) \ \forall x \in \tilde{\mathbb{X}} \\ v\left(x_A\right) &= 0 = V\left(\pi, x_A\right). \end{aligned}$$

The converse is proved by writing down the equation:

$$J_T\left(\pi, x_0\right) = \rho\left(r\left(x_0, \psi\left(x_0\right), x_1\right) + J_{T-1}\left(\pi, x_0\right)\right).$$

Taking the limit as $T \to \infty$ on both sides, we arrive at the following equation:

$$\lim_{T\to\infty} J_T(\pi, x_0) = \lim_{T\to\infty} \rho\left(r\left(x_0, \psi\left(x_0\right), x_1\right) + J_{T-1}\left(\pi, x_0\right)\right).$$

Since $\rho\left(\cdot\right)$ is continuous by assumption, we conclude that

$$\lim_{T\to\infty} J_T\left(\pi, x_0\right) = \rho\left(r\left(x_0, \psi\left(x_0\right), x_1\right) + \lim_{T\to\infty} J_{T-1}\left(\pi, x_0\right)\right).$$

Using the fact that

$$\lim_{T\to\infty} J_T(\pi, x_0) = V(\pi, x_0) = v(x_0), \ \forall x_0 \in \tilde{\mathbb{X}}$$

as $T \to \infty$, we rewrite the previous equation as:

$$v(x_0) = \rho\left(r\left(x_0, \psi(x_0), x_1^{\pi}\right) + v\left(x_1^{\pi}\right)\right) = \sigma\left(r_{x_0} + v, x_0, \psi(x_0) \circ Q_{x_0}\right), \ \forall x_0 \in \tilde{\mathbb{X}},$$

which is the same as Eq. 2.3.16. Furthermore, $V(\pi, x_A) = 0 = v(x_A)$ by definition.

□

Theorem 5. *Assume the following conditions hold for a uniformly transient Markov control model:*

1) The stochastic kernels $Q_x : a \to Q(\cdot|x, a)$ are continuous $\forall x \in \mathbb{X}$;

2) The one-step dynamic reward measure ρ is Markov (see Definition 15), and a sequence of corresponding reward transition mappings $\sigma : m \to \sigma(\psi, x, m)$ is upper semi-continuous;

3) The assumptions in Theorem 4 are satisfied;

4) For every $x \in \mathbb{X}$ the set $A(x)$ is compact;

Then a maximizer for the dynamic programing equation

$$v(x) = \max_{\delta \in \mathcal{P}(A(x))} \sigma\left(r(x, \cdot, \cdot) + v(\cdot), x, \delta \circ Q_x\right) \ \forall x \in \tilde{\mathbb{X}}$$

$$v(x_A) = 0, \tag{2.3.17}$$

exists. Furthermore, an optimal randomized stationary policy, $\pi^* := \{\psi^*, \psi^*, \cdots\}$ *exists and each* ψ^* *is a maximizer for the right-hand side of Eq. 2.3.17; In addition, a bounded measurable function v, (i.e.,* $\|v\|_w < \infty$*) is a solution of Eq. 2.3.17 if and only if it equals* $V^\dagger$ *in Eq. 2.3.15, i.e.,* $v(x) = V^\dagger(x),\ \forall x \in \mathbb{X}$.

Proof. Since the set of all policy sequences of the form $\pi = \{\lambda, \pi, \pi, \dots\}$ contains Π^{RM}, we write down the inequality

$$V^\dagger(x_0) \leq \sup_{\lambda \in \mathcal{P}(A(x_0)), \pi \in \Pi^{RM}} \rho\left(r(x_0, a_0, x_1) + V(\pi, x_1)\right),$$

where $V^\dagger$ is defined by Eq. 2.3.14. Because ρ is monotone, we move the supremum operator inside:

$$\begin{aligned} V^\dagger(x_0) &\leq \sup_{\lambda \in \mathcal{P}(A(x_0))} \rho\left(r(x_0, a_0 x_1) + \sup_{\pi \in \Pi^{RM}} V(\pi, x_1)\right) \\ &\leq \sup_{\lambda \in \mathcal{P}(A(x_0))} \rho\left(r(x_0, a_0, x_1) + V^\dagger(x_1)\right). \end{aligned}$$

By Assumption 3(3), i.e., $\left\|V^\dagger\right\|_w < \infty$, the right-hand side is well defined. Thus $V^\dagger$ satisfies the inequality

$$V^\dagger(x) \leq \sup_{\lambda \in \mathcal{P}(A(x))} \sigma\left(r_x + V^\dagger, x, \lambda \circ Q_x\right),\ x \in \mathbb{X}. \tag{2.3.18}$$

Since the existence of a solution for Eq. 2.3.17 is assured by the semi-continuity of the mapping $\sigma : \lambda \to \sigma(r_x + v, x, \lambda \circ Q_x)$ and the weak compactness of the set

$\mathcal{P}(A(x))$, we conclude that

$$V^{\dagger}(x) \leq \sigma\left(r_x + V^{\dagger}, x, \psi^*(x) \circ Q_x\right), \ x \in \mathbb{X}.$$

Here, ψ^* is a solution to the optimization problem represented by the right-hand side of Eq. 2.3.18. By iterating the inequality above, appealing to the monotonicity property of σ, and applying the policy $\pi^* = \{\psi^*, \psi^*, \dots\}$, we obtain the fact that

$$V^{\dagger}(x_0) \leq V(\pi^*, x_0).$$

Since by assumption $V^{\dagger}(\cdot)$ is the optimal value function, $V^{\dagger}(\cdot) \geq V(\pi^*, \cdot)$, which along with the previous inequality, imply $V^{\dagger}(\cdot) = V(\pi^*, \cdot)$. Using Theorem 4, we conclude $V^{\dagger}(\cdot)$ satisfies the dynamic programming equation.

To prove the converse, we first suppose $v(\cdot)$ satisfies Eq. 2.3.17, and $\|v\|_w < \infty$. Since the mapping $\sigma : \lambda \to \sigma(r_x + v, x, \lambda \circ Q_x)$ is continuous and the set $\mathcal{P}(A(x))$ is weakly compact, an optimal control function, $\hat{\psi}$, exists. Furthermore, $\hat{\psi}$ is the maximizer for the right-hand size of the dynamic programing equation. This enables us to write

$$v(x) = \sigma\left(r_x + v, x, \hat{\psi}(x) \circ Q_x\right), \ x \in \mathbb{X}. \tag{2.3.19}$$

Using Theorem 4, we conclude that

$$v(x) = V(\hat{\pi}, x) \leq V^{\dagger}(x), \ x \in \mathbb{X}, \tag{2.3.20}$$

where $\hat{\pi} = \left\{\hat{\psi}, \hat{\psi}, \dots\right\}$. On the other hand, it follows from Eq. 2.3.17 that the control function $\hat{\psi}$ satisfies

$$v(x) \geq \sigma\left(r_x + v, x, \hat{\psi}(x) \circ Q_x\right), \ x \in \tilde{\mathbb{X}}.$$

Using the monotonicity property of σ, we iterate the above inequality and arrive at

$$v(x) \geq \rho_{0,T}(0, Z_1, \dots, Z_T + v(x_T)),$$

where Z_t is the reward sequence resulting from applying the policy $\hat{\pi}$. By taking $T \to \infty$ for the equation above, we conclude that

$$v(x) \geq V(\hat{\pi}, x) = V^{\dagger}(x), \ x \in \tilde{\mathbb{X}}.$$

The last inequality, together with Eq. 2.3.20 and the fact that $v(\cdot) = V^{\dagger}(\cdot)$ imply the stationary policy $\hat{\psi}$ that satisfies Eq. 2.3.19 is optimal, i.e., $V(\hat{\pi}, x) = V(\pi^*, x) = V^{\dagger}(x), \ \forall x \in \mathbb{X}$. In addition, we know $v(x_A) = V^{\dagger}(x_A) = 0$ from the definition of transition Markov model. □

In the theorem above, we provide the optimality criteria for the case of randomized stationary Markov policies. Next, we prove the sufficiency of randomized stationary Markov policies as optimal policies.

Theorem 6. *Assume the assumptions in the Theorem 5 are satisfied. Then a w-bounded measurable function $v : \mathbb{X} \to \mathcal{R}$, with $\|v\|_w < \infty$, satisfies Eq. 2.3.17 if and*

only if $v(x) = V^*(x)$ *for all* $x \in \mathbb{X}$. *Moreover, a maximizer* ψ^* *exists for Eq. 2.3.17 and defines an optimal randomized stationary Markov policy* $\pi^* = \{\psi^*, \psi^*, \psi^*, \dots\}$.

Proof. We denote a Markov policy by $\pi^1 = \{\psi_1, \psi_2, \dots\}$. Given the monotonicity and continuity of $\rho(\cdot)$, we have

$$\begin{aligned}
V^*(x_0) &= \sup_{\lambda_0,\lambda_1,\dots} \limsup_{T\to\infty} \rho\left(r(x_0,a_0,x_1) + J_{T-1}\left(\pi^1, x_1\right)\right) \\
&\leq \sup_{\lambda_0,\lambda_1,\dots} \limsup_{T\to\infty} \rho\left(r(x_0,a_0,x_1) + \sup_{\tau\geq T-1} J_\tau\left(\pi^1, x_1\right)\right) \\
&= \sup_{\lambda_0,\lambda_1,\dots} \lim_{T\to\infty} \rho\left(r(x_0,a_0,x_1) + \sup_{\tau\geq T-1} J_\tau\left(\pi^1, x_1\right)\right) \\
&= \sup_{\lambda_0,\lambda_1,\dots} \rho\left(r(x_0,a_0,x_1) + \limsup_{T\to\infty} J_{T-1}\left(\pi^1, x_1\right)\right) \\
&= \sup_{\lambda_0,\lambda_1,\dots} \rho\left(r(x_0,a_0,x_1) + V\left(\pi^1, x_1\right)\right).
\end{aligned}$$

By appealing to the monotonicity property of $\rho(\cdot)$, we can move the supremum inside the argument

$$\begin{aligned}
V^*(x_0) &\leq \sup_{\lambda_0} \rho\left(r(x_0,a_0,x_1) + \sup_{\Lambda^1} V\left(\pi^1, x_1\right)\right) \\
&= \sup_{\lambda_1} \rho\left(r(x_0,a_0,x_1) + V^*(x_2)\right).
\end{aligned}$$

Thus $V^*(\cdot)$ satisfies the inequality

$$V^*(x) \leq \sup_{\lambda\in\mathcal{P}(A(x))} \sigma\left(r_x + V^*, x, \lambda\circ Q_x\right),\ x \in \mathbb{X}. \tag{2.3.21}$$

Appealing to the monotonicity property of σ, iterating the above inequality and

letting ψ^* be the maximizer from the equation above we conclude that

$$V^*(x) \leq V(\pi^*, x), \ x \in \mathbb{X},$$

where $\pi^* = \{\psi^*, \psi^*, \dots\}$ is a stationary Markov policy that maximizes Eq. 2.3.21. Therefore, optimization with respect to stationary Markov policies is sufficient, and the result follows from Theorem 5. □

We need to ensure the first three conditions in Assumption 3 are satisfied by all CPT-inspired reward measures, which have the form:

$$\begin{aligned}\sigma_t(r_x + v, x, \psi(x) \circ Q_x) = &\int_0^\infty w^+\left(\psi(x) \circ Q_x\left((r_x + v)_+ > s\right)\right) ds \\ &- \int_0^\infty w^-\left(\psi(x) \circ Q_x\left((r_x + v)_- > s\right)\right) ds. \end{aligned} \tag{2.3.22}$$

With this specific form, we can write down the operator T with respect to a transient stationary policy $\pi = \{\psi, \psi, \dots\}$ as:

$$\begin{aligned} T(\pi) v(x): \ = \ &\int_0^\infty w^+\left(\psi(x) \circ \tilde{Q}_x\left((r_x + v)_+ > s\right)\right) ds \\ &- \int_0^\infty w^-\left(\psi(x) \circ \tilde{Q}_x\left((r_x + v)_- > s\right)\right) ds \\ = \ &\int_0^\infty (r_x + v)_+ \, d\left(\left(\psi(x) \circ \tilde{Q}_x\right)^{w^+, r_x + v}\right) \\ &- \int_0^\infty (r_x + v)_- \, d\left(\left(\psi(x) \circ \tilde{Q}_x\right)^{w^-, r_x + v}\right). \end{aligned}$$

The second equality is due to the fact that the transformed measures $\left(\psi(x) \circ \tilde{Q}_x\right)^{w^+, r_x + v}$

and $\left(\psi(x) \circ \tilde{Q}_x\right)^{w^-, r_x+v}$ are absolutely continuous with respect to $\psi(x) \circ \tilde{Q}_x$; hence a Radon-Nikodym derivative exists. The lemma below will be used in Theorem 7.

Lemma 5. *If a uniformly transient Markov model is given, i.e.,*

$$\lim_{k\to\infty} \left\| \sum \left(\pi_k(x) \circ \tilde{Q}_x\right)^k \right\|_w \leq \kappa_1, \ \forall \pi = \{\pi_1, \pi_2, \dots\} \in \Pi^{RM},$$

then there exists a $\tilde{k} > 0$ such that

$$\left\| \left(\left(\pi_k(x) \circ \tilde{Q}_x\right)^{w^-, v} \right)^k \right\|_w < 1, \ \forall k \geq \tilde{k}, \ \forall \pi = \{\pi_1, \pi_2, \dots\} \in \Pi^{RM}, \ v \in \mathcal{V}.$$

Proof. Note that if

$$\lim_{k\to\infty} \left\| \sum \left(\pi_k(x) \circ \tilde{Q}_x\right)^k \right\|_w$$

is finite, then there exists a $\tilde{k}$ such that for all $k \geq \tilde{k}$

$$\left\| \left(\left(\pi_k(x) \circ \tilde{Q}_x\right)^{w^-, v} \right)^k \right\|_w < 1.$$

Since $w^-(p)$ equal to 1 if and only if $p = 1$, the assertion follows. □

A similar statement (Lemma 5) can be made about

$$\left\| \left(\left(\pi_k(x) \circ \tilde{Q}_x\right)^{w^+, v} \right)^k \right\|_w.$$

Theorem 7. *Suppose the Markov control model is uniformly transient, and the following assumptions are satisfied:*

1) w^+ and w^- are continuous non-decreasing functions;

2) There is a constant $\bar{r}$ such that

$$\sup_{A(x)} |r(x, a, x')| \leq \bar{r} w(x'), \ \forall x, x' \in \mathbb{X};$$

Then Eq. 2.3.22 satisfies Assumption 3.

Proof. 1) Letting $\pi = \{\psi, \psi, \dots\}$, we know the following

$$\begin{aligned}
& \left\| T(\pi)^k v_1 - T(\pi)^k v_2 \right\|_w \\
\leq & \left\| \int_{-\infty}^{\infty} (v_1 - v_2) \, d \left(\left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^+, v_1} \right)^k + \left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^-, v_1} \right)^k \right) \right\|_w \\
& + \epsilon_k \|v_1 - v_2\|_w + \tilde{\epsilon}_k \|v_1 - v_2\|_w \\
\leq & \left(\left\| \left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^+, v_1} \right)^k \right\|_w + \left\| \left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^-, v_1} \right)^k \right\|_w + \epsilon_k + \tilde{\epsilon}_k \right) \\
& \cdots \|v_1 - v_2\|_w ,
\end{aligned} \tag{2.3.23}$$

where $\left(\left\| \left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^+, v_1} \right)^k \right\|_w + \left\| \left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^-, v_1} \right)^k \right\|_w + \epsilon_k + \tilde{\epsilon}_k \right)$ can be chosen to be less than 1. The appropriate k is found by appealing to Lemma 5. Since the Markov control model is uniformly transient (i.e., the probability weight on the non-absorbing states decreases as k increases), ϵ_k and $\tilde{\epsilon}_k$ can be made arbitrarily small as $k \to \infty$. Here, $\tilde{\epsilon}_k$ captures the difference between the distorted measures, of the the previous finite number of per-stage rewards, induced by v_1 and v_2. As k increases, the measure distortions induced by v_1 and v_2 on the initial steps disappear. The first inequality in Eq. 2.3.23 is due to the fact that the transformed measures $\left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^+, v_1} \right)^k$ and $\left(\left(\psi(x) \circ \tilde{Q}_x \right)^{w^-, v_1} \right)^k$ are absolutely continuous with respect to $\left(\psi(x) \circ \tilde{Q}_x \right)^k$; hence a Radon-Nikodym derivative ex-

ists. Furthermore, we used the fact that

$$\begin{aligned}&\int f_1(x_1)\,dP(x_1) - \int f_2(x_1)\,dQ(x_1)\\ =&\int (f_1 - f_2)(x_1) dP(x_1) + \int f_2(x_1)\left(dP(x_1) - dQ(x_1)\right),\end{aligned}$$

where f_1 and f_2 are two w-bounded measurable functions, and P and Q are two σ-finite measures. In Eq. 2.3.23, ϵ_k represents the difference between the v_1 and v_2 distorted measures (i.e., $\epsilon_k = \frac{\|v_2\|_w}{\|v_1 - v_2\|_w} \|P - Q\|_w$, where P is distorted by v_1 and Q is distorted by v_2). In other words, ϵ_k captures the difference in the distorted measures at the k-th stage induced by v_1 and v_2.

2) To prove condition 2, we appeal to the continuity property of w^+ and w^-, which implies the following inequalities:

$$\begin{aligned}&\|\sigma_t(z_1, x, m) - \sigma_t(z_2, x, m)\|_w\\ =&\Bigg\| \int_0^\infty w^+\left(m\left((z_1)_+ > s\right)\right) - w^+\left(m\left((z_2)_+ > s\right)\right) ds\\ &- \left(\int_0^\infty w^-\left(m\left((z_1)_- > s\right)\right) - w^-\left(m\left((z_2)_- > s\right)\right) ds\right)\Bigg\|_w\\ \leq&\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,\end{aligned}$$

where for brevity the measure $\psi(x) \circ \tilde{Q}_x$ is denoted by m.

From the continuity property of the functions w^+ and w^-, we know that there

exist δ_1 and δ_2 such that

$$\begin{aligned} & \left\| m\left((z_1)_+ > s\right) - m\left((z_2)_+ > s\right) \right\|_w \leq && \delta_1 \\ \implies \quad & \left\| w^+\left(m\left((z_1)_+ > s\right)\right) - w^+\left(m\left((z_2)_+ > s\right)\right) \right\|_w && \leq \frac{\epsilon}{2} \end{aligned}$$

and

$$\begin{aligned} & \left\| m\left((z_1)_- > s\right) - m\left((z_2)_- > s\right) \right\|_w \leq && \delta_2 \\ \implies \quad & \left\| w^+\left(m\left((z_1)_- > s\right)\right) - w^+\left(m\left((z_2)_- > s\right)\right) \right\|_w && \leq \frac{\epsilon}{2}. \end{aligned}$$

Since

$$\|z_1 - z_2\|_w \to 0,$$

implies

$$\left\| m\left((z_1)_+ > s\right) - m\left((z_2)_+ > s\right) \right\|_w \to 0$$

and

$$\left\| m\left((z_1)_- > s\right) - m\left((z_2)_- > s\right) \right\|_w \to 0,$$

we conclude that there exists a δ_3 such that

$$\begin{aligned} \|z_1 - z_2\|_w &< \delta_3 \\ \left\| m\left((z_1)_+ > s\right) - m\left((z_2)_+ > s\right) \right\|_w &\leq \delta_1 \\ \left\| m\left((z_1)_- > s\right) - m\left((z_2)_- > s\right) \right\|_w &\leq \delta_2. \end{aligned}$$

3) Now, we prove condition 3 of Assumption 3. Since

$$\sup_{A(x)} |r(x,a,x')| \leq \bar{r} w(x'), \ \forall x, x' \in \mathbb{X},$$

and the Markov control model is assumed to be uniformly transient, we can show, by induction, that $V(\pi, x)$ is also w-bounded for any transient policy $\pi = \{\lambda, \lambda, \dots\} \in \Pi^{RM}$. We start by writing down the one-stage reward function:

$$\begin{aligned} \sigma\left(|r_x|, x, \lambda \circ \tilde{Q}_x\right) &= \int_0^\infty |r_x| \, d\left(\left(\lambda \circ \tilde{Q}_x\right)^{w^+, |r_x|}\right) \\ &\leq \bar{r} \int_0^\infty w d\left(\left(\lambda \circ \tilde{Q}_x\right)^{w^+, |r_x|}\right). \end{aligned}$$

The two-stage reward function is written as

$$\begin{aligned} &\sigma\left(|r_x| + \sigma\left(|r_{x_1}|, x_1, \lambda \circ \tilde{Q}_{x_1}\right), x, \lambda \circ \tilde{Q}_x\right) \\ &\leq \bar{r}\left(\int_0^\infty w d\left(\left(\lambda \circ \tilde{Q}_x\right)^{w^+, |r_x|}\right) + \int_0^\infty w d\left(\left(\lambda \circ \tilde{Q}_x\right)^{w^+, |r_x| + |r_x|}\right)^2\right). \end{aligned}$$

By iterating the inequality above and appealing to Lemma 5, we arrive at the conclusion that

$$V(\pi, \cdot) \leq \bar{r} k w(\cdot),$$

where the k is found in Eq. 2.3.12. □

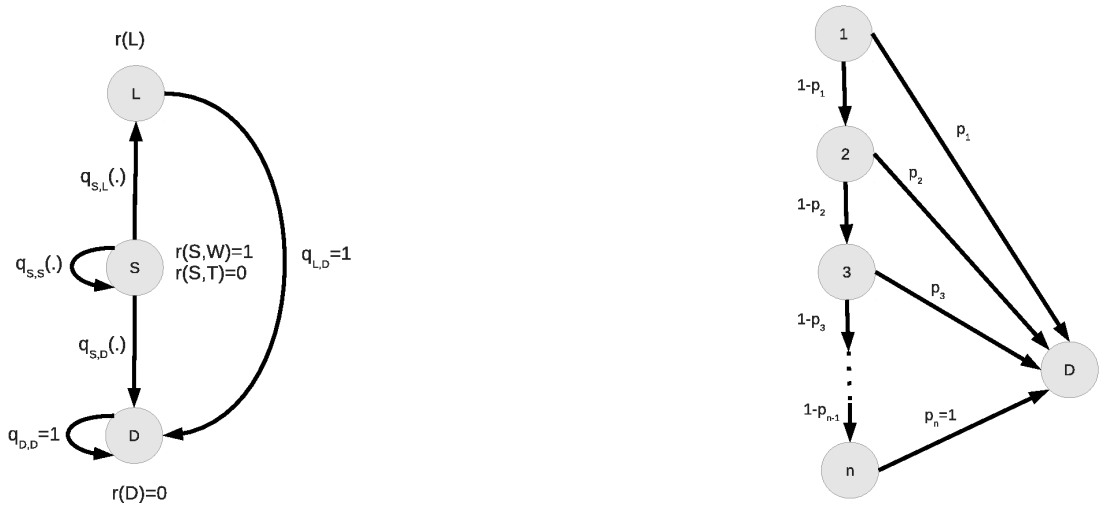
In the next section, we present a numerical example to explore the structure of optimal policies for the transient Markov case.

2.3.4 The Organ Transplant Example: A Comparative Analysis

We will compare numerically the type of polices obtained from CPT-based measures against some of the other risk-sensitive approaches.

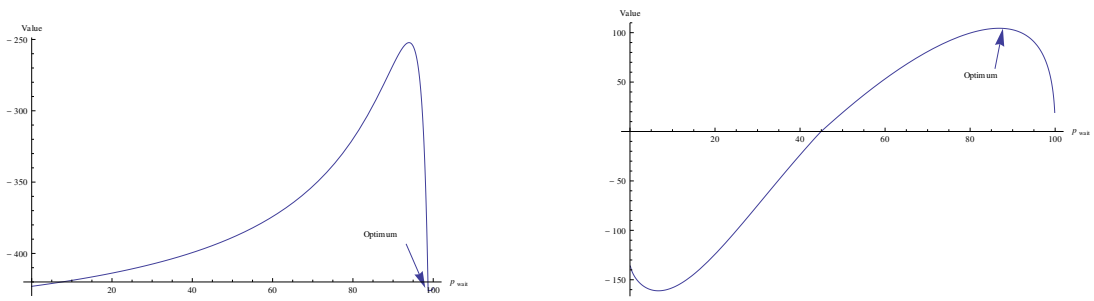
Example 7. The following example is from [18], which is a simplified version of the organ transplant problem discussed in [1]. The problem considers the discrete-time absorbing Markov chain depicted in Fig. 2.3.1a. The initial state S (i.e., sick) represents a patient waiting on an organ transplant due to sickness. The state L (i.e., live) represents the state where the patient lives after a successful transplant. The state D, an absorbing state, represents death. There are two possible actions to take in state S: 1) one can wait (W), in which case the next state could either be D or S probabilistically; 2) one can choose to transplant (T), which concentrates the transition probability on states L and D (i.e., states L and D are the only two possible next states). The probability of death is lower for W than for T, but a successful transplant may result in a longer life. In other two states, only the action continue is allowed. The reward collected at each time step is months of life. In state S, a reward equal to 1 is collected if the control is W; otherwise, the immediate reward is 0. In state L, the reward r(L) is collected representing the certainty equivalent of the random length of life after the transplant. In state D the reward is 0.

The states where there is only one possible action allowed have a deterministic reward function (i.e., L and D). In particular, the equivalent length of life at the state L is $r(L)$. However, this value is generated by taking on certain assumptions, which are the focus of the following discussion. The state L is in fact an aggregation



(a) Organ Transplant Transitions (b) The Survival Model

Figure 2.3.1: Organ Transplant State Transitions & Rewards



(a) Semideviation [18] (b) CPT-based

Figure 2.3.2: Optimal Policy Comparison of the Organ Transplant Example

of n states in a survival model representing months of life after the transplant, as depicted in Fig. 2.3.1b. At the state i, $i = 1, \ldots, n$, the patient dies with probability p_i and survives with probability $1 - p_i$. The patient will die for sure in the state n (i.e., $p_n = 1$). The reward collected at each state i is equal to 1. In Çavuş and Ruszczyński [18], the problem is stated as a minimization problem. However, we desire a maximization problem, thus we compare our results to that of Çavuş and Ruszczyński's [18] by negating the rewards.

In [18], $r\,(L)$ is calculated from the survival model using the transition mapping

of the form:

$$\sigma(\varphi, i, m) = \underbrace{\mathbb{E}_m[\varphi]}_{expected\ value} + \kappa \underbrace{\mathbb{E}\left[(\varphi - \mathbb{E}_m[\varphi])_+\right]}_{semideviation}. \tag{2.3.24}$$

In Eq. 2.3.24, the measure m is the transition kernel at the current state i, and the function $\varphi(\cdot)$ is the reward collected at the current state and action plus the value function at the next state (i.e., cost-to-go). At each state $i = 1, \ldots, n-1$, two transitions are possible: 1) transition to the state D with probability p_i and $\varphi = -1$; 2) transition to the state $i+1$ with probability $1 - p_i$ and $\varphi = -1 + v_{i+1}(i+1)$. At the state $i = n$, the transition to D occurs with probability 1, and $\varphi = -1$. Therefore, $v_n(n) = -1$.

The survival problem is now a finite-horizon problem, which can be expressed as in Eq. 2.3.1. Since there is only one action allowed, the minimization operation is eliminated. The equation has the form:

$$v_i(i) = \sigma(\varphi, i, Q_i), \ i = 1, \ldots, n-1,$$

with φ and Q_i being the reward and the transition probability respectively. The values of φ and Q_i are explained in the previous paragraph. By induction, $v_i(i) \leq 0$, for $i = n-1, n-2, \ldots, 1$.

The mean and semideviation components of Eq. 2.3.24 at the states $i =$

$1, \ldots, n-1$ can be calculated as follows:

$$\mathbb{E}_{Q_i}[\varphi] = -p_i + (1-p_i)(-1 + v_{i+1}(i+1)) = -1 + (1-p_i)\, v_{i+1}(i+1),$$

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

where the last equality in the equation above is implied by the fact that $v_{i+1}(i+1) \leq 0$.

For $i = 1, \ldots, n-1$, the dynamic programming equation for the optimization problem stated in Eq. 2.3.1 takes the form:

$$v_i(i) = \underbrace{-1 + (1-p_i)\, v_{i+1}(i+1)}_{expected\ value} - \kappa \underbrace{p_i(1-p_i)\, v_{i+1}(i+1)}_{semideviation},\ i = n-1, n-2, \ldots, 1.$$

The value $v(1)$ is the negative of the risk-adjusted length of life with the new transplanted organ. For $\kappa = 0$, the above formula gives the negative of the expected length of life with the new organ. In the calculations below, we use the transition data from Table 2.3.5. They have been chosen for illustrative purposes only, and do not correspond to any real-life medical data. For the survival model, the distribution function, $F(x)$, of lifetime of the American population is suggested by Jasiulewicz

Action	S	L	D
W	0.99882	0	0.00118
T	0	0.90782	0.09218

Table 2.3.5: Transition Probabilities From State S

Distribution	Parameters	Weights
Weibull	$\delta = 0.297, \beta = 0.225$	$w_1 = 0.0170$
Lognormal	$m = 3.11, \sigma = 0.218$	$w_2 = 0.0092$
Gompertz	$b = 0.0000812, \alpha = 0.0844$	$w_3 = 0.9737$

Table 2.3.6: Organ transplant example: parameters for $F(x)$.

[58]. It is a mixture of Weibull, lognormal, and Gompertz distribution:

$$F(x) := w_1 \left(1 - e^{-\left(\frac{x}{\delta}\right)^{\beta}}\right) + w_2 \Phi\left(\frac{\log x - m}{\sigma}\right) + w_3 \left(1 - e^{-\frac{b}{a}(e^{\alpha x} - 1)}\right), \ x \geq 0.$$

The values of the parameters and weights, provided by Jasiulewicz [58], are given in Table 2.3.6.

Using the information provided above, the probability of dying in the k-th month can be calculated using the equation:

$$p_k = \frac{F\left(\frac{k}{12} + \frac{1}{24}\right) - F\left(\frac{k}{12} - \frac{1}{24}\right)}{1 - F\left(\frac{k}{12} - \frac{1}{24}\right)}, \ k = 1, 2, \ldots.$$

The maximum lifetime of the patient is assumed to be 1200 months, and the post-transplant survival probabilities for the patient starts from $k = 300$. Hence, a total of 900 steps, n=900, is used in the survival model to calculate $r(L)$.

If we let $\lambda = (\lambda_W, \lambda_T)$ be a randomized policy in the state S and let $\Lambda = \{\lambda \in \mathbb{R}^2 : \lambda_W + \lambda_T = 1, \lambda \geq 0\}$, then the dynamic programming equation at the

state S has the form:

$$v(S) = \min_{\lambda \in \Lambda} \{\lambda_W [q_{S,S}(W)(v(S)-1) + q_{S,D}(W)(v(D)-1)]$$
$$+\lambda_T [q_{S,L}(T) v(L) + q_{S,D}(T) v(D)]$$
$$+\kappa \left(\lambda_W \left[q_{S,S}(W)(v(S)-1-\mu)_+ + q_{S,D}(W)(v(D)-1-\mu)_+\right]\right.$$
$$\left.+\lambda_T \left[q_{S,L}(T)(v(L)-\mu)_+ + q_{S,D}(T)(v(D)-\mu)_+\right]\right)\}.$$

Here, $\kappa = 1$. If λ is held fixed in the equation above, then we can solve for $v(S)$. By varying $\lambda \in (0,1)$, we obtain Fig. 2.3.2a. We can compute the value function of the CPT-based reward measure as follows:

$$v(S) = \max_{\lambda \in \Lambda} \int_0^\infty w^+ \left(\lambda_W \left(q_{S,S}(W) 1\left\{(v(S)+1-\mu)_+ > s\right\}\right.\right.$$
$$+q_{S,D}(W) 1\left\{(v(D)+1-\mu)_+ > s\right\}\Big)$$
$$+\lambda_T \left(q_{S,L}(T) 1\left\{(v(L)-\mu)_+ > s\right\}\right.$$
$$\left.\left.+q_{S,D}(T) 1\left\{(v(D)-\mu)_+ > s\right\}\right)\right) ds$$
$$-\int_0^\infty w_- \left(\lambda_W \left(q_{S,S}(W) 1\left\{(v(S)+1-\mu)_- > s\right\}\right.\right.$$
$$+q_{S,D}(W) 1\left\{(v(D)+1-\mu)_- > s\right\}\Big)$$
$$+\lambda_T \left(q_{S,L}(T) 1\left\{(v(L)-\mu)_- > s\right\}\right.$$
$$\left.\left.+q_{S,D}(T) 1\left\{(v(D)-\mu)_- > s\right\}\right)\right) ds.$$

In the equation above, $w^+(\cdot) = w^-(\cdot) = \exp(-0.5(-\ln(\cdot)))$, and μ is the expected value without probability weighting. The numerical results of the three

Method	$r(L)$	Optimum Value	Optimal λ_W
Expected Value	610.46	846.611	1.000000
Semideviation	515.33	426.139	0.987236
CPT	702.32	104.438	0.868232

Table 2.3.7: Organ Transplant Optimal Value and Policy Comparison

solution methods (i.e., expected value, semideviation, and CPT) [9] are listed in Table 2.3.7. Furthermore, the value functions of the semideviation and the CPT performance measures are plotted in Fig. 2.3.2.

We calculated $r(L)$ for the CPT method using the following equation:

$$\sigma(\varphi, i, m) = \int_0^\infty w^+(m(\varphi_+ > s))ds.$$

As is evident from Table 2.3.7, the CPT performance measure produces a more randomized optimal policy than the other two approaches. The λ_W value of 0.99 is very close to the deterministic policy of W (i.e., to wait). In fact, obtaining a very randomized policy is difficult using semideviation. The ease with which the CPT performance measure is able to obtain an optimal randomized policy can be explained by the fact that the probability weighting function is applied to the control. Intuitively, the need for randomized policies stems from the nonlinear transformation of the uncertainty in the system, which renders deterministic optimal policies insufficient.

[9] In the table, some values is negated to be positive for the purpose of comparison.

2.4 Reward Measures and Optimal Policies

From the previous sections, we have learned that we can solve finite-horizon non-convex optimization problems with reward functions of the form:

$$\begin{aligned} V_t(x_0, \pi) = \rho\left(r_t(x_t^\pi, a_t^\pi, x_{t+1}^\pi) + \rho\left(r_{t+1}(x_{t+1}^\pi, a_{t+1}^\pi, x_{t+2}^\pi)\right.\right. \\ + \rho\left(r_{t+2}(x_{t+2}^\pi, a_{t+2}^\pi, x_{t+3}^\pi) + \cdots +\right. \\ \left.\left.\left.\rho\left(r_{T-1}(x_{T-1}^\pi, a_{T-1}^\pi, x_T^\pi) + r_T\left(x_T^\pi\right)\right)\cdots\right)\right)\right), \end{aligned}$$

where V_t is known as the reward-to-go function. The equation is analogous to the expanded form of the standard expected value measure:

$$\begin{aligned} \tilde{V}_t\left(x_t, \pi\right) = \mathbb{E}\left[r_t\left(x_t, a_t^\pi, x_{t+1}^\pi\right) + \mathbb{E}\left[r_{t+1}\left(x_{t+1}^\pi, a_{t+1}^\pi, x_{t+2}^\pi\right)\right.\right. \\ + \mathbb{E}\left[r_{t+2}\left(x_{t+2}^\pi, a_{t+2}^\pi, x_{t+3}^\pi\right) + \cdots +\right. \\ \left.\left.\mathbb{E}\left[r_{T-1}\left(x_{T-1}^\pi, a_{T-1}^\pi, x_T^\pi\right) + r_T\left(x_T^\pi\right)|x_{T-1}^\pi\right]\cdots|x_{t+2}^\pi\right]|x_{t+1}^\pi\right] x_t\right]. \end{aligned}$$

One of the advantages of the standard expected value measure is that it can be written more compactly as

$$\tilde{V}_t\left(x_0, \pi\right) = \mathbb{E}\left[\sum_{i=t}^{T-1} r_i\left(x_i^\pi, a_i^\pi, x_{i+1}^\pi\right) + r_T\left(x_T^\pi\right)\right].$$

We would like to write down a similar simplified counterpart in the CPT Markov conditional reward measure case for V_t. In other words, we would like to

study the reward measures:

$$V_t(x_t, \pi) = \int_0^\infty w_t\left(\pi_t(x_t) \circ Q_{x_t,t}\left(r_t\left(x_t, a_t^\pi, x_{t+1}\right) + V_{t+1}\left(x_{t+1}, \pi\right)\right) > s\right) ds_t,$$

$$\text{where } V_T(x_T, \pi) = r_T(x_T). \tag{2.4.1}$$

For simplicity, we only consider bounded non-negative rewards (i.e., $r_t \geq 0$ and $r_t \leq M,\ M > 0\ \forall t \geq 0$). The inclusion of negative rewards is a straightforward exercise. We can see that Eq. 2.4.1 is a complicated sequence of nested integrals. We would like to simplify this expression by introducing some new notations.

We note that the transformed measure on the measurable space $(\mathbb{X} \times \mathbb{A}, \mathcal{B}(\mathbb{X} \times \mathbb{A}))$

$$\mathbb{P}^{w,\psi}(\psi > s) := w\left(\mathbb{P}(\psi > s)\right) \tag{2.4.2}$$

is absolutely continuous with respect to $\mathbb{P}(\psi > s)$. By the Radon-Nikodym theorem, there exists a measurable function such that

$$\mathbb{P}^{w,\psi}(B) := \int_B \frac{d\mathbb{P}^{w,\psi}}{d\mathbb{P}} d\mathbb{P},$$

where $\frac{d\mathbb{P}^{w,\psi}}{d\mathbb{P}}$ is a Radon-Nikodym derivative.

Using Radon-Nikodym derivatives of the form $\frac{d\mathbb{P}^{w,\psi}}{d\mathbb{P}}$, we can rewrite Eq.2.4.1

as

$$V_t(x_t,\pi) := \int_{\mathbb{X}\times\mathbb{A}} r_t\left(x_t, a_t^{\pi}, x_{t+1}\right) + \int_{\mathbb{X}\times\mathbb{A}} r_{t+1}\left(x_{t+1}^{\pi}, a_{t+1}^{\pi}, x_{t+1}^{\pi}\right)$$

$$\cdots$$

$$\frac{d\left(\pi\left(x_{t+1},t+1\right)\circ Q_{x_{t+1},t+1}\right)^{w_{t+1},r_{t+1}+v_{t+2}}}{d\pi\left(x_{t+1},t+1\right)\circ Q_{x_{t+1},t+1}} d\pi\left(x_{t+1},t+1\right)\circ Q_{x_{t+1},t+1}$$

$$\frac{d\left(\pi\left(x_t,t\right)\circ Q_{x_t,t}\right)^{w_t,r_t+v_{t+1}}}{d\pi\left(x_t,t\right)\circ Q_{x_t,t}} d\pi\left(x_t,t\right)\circ Q_{x_t,t},$$

where v_t is recursively defined as

$$\begin{aligned} v_t\left(x\right) &= \int_0^{\infty} w_t\left(\pi_t\left(x\right)\circ Q_{x,t}\left(r_t + v_{t+1} > s\right)\right) ds \\ v_T\left(x\right) &= r_T\left(x\right). \end{aligned}$$

In the equation above, the mapping $r_t+v_{t+1} : \mathbb{X}\times\mathbb{A}\times\mathbb{X} \to \mathbb{R}$ is used to transform the probability measure $\pi_t\left(x\right)\circ Q_{x,t}$, which is a probability measure on the probability space $\left(\mathbb{X}\times\mathbb{A}, \mathcal{B}\left(\mathbb{X}\times\mathbb{A}\right)\right)$. According to Eq. 2.4.2, the function used to transform the measure $\pi_t\left(x\right)\circ Q_{x,t}$ should be a $\mathcal{B}\left(\mathbb{X}\times\mathbb{A}\right)$-measurable function. It is obvious that $r_t + v_{t+1}$ is a $\mathcal{B}\left(\mathbb{X}\times\mathbb{A}\right)$-measurable function if x is held fixed (i.e., $r_t + v_{t+1}$ is treated as $r_t\left(x,\cdot\right) + v\left(\cdot\right)$ in the equation above). In the sequel, whether or not x is held fixed for the reward function $r_t + v_{t+1}$ should be obvious from the context.

Using the Radon-Nikodym derivative notation, V_t can be written more compactly as:

$$V_t(x_t, \pi) = \mathbb{E}_t^{\pi}\left[\left(\sum_{i=t}^{T-1} r_t\left(x_i^{\pi}, a_i^{\pi}, x_{i+1}^{\pi}\right) + r_T\left(x_T^{\pi}\right)\right)\prod_{i=t}^{T-1}\frac{d\left(\pi\left(x_i, i\right)\circ Q_{x_i,i}\right)^{w_i, r_i+v_{i+1}}}{d\pi\left(x_i, i\right)\circ Q_{x_i,i}}\right]$$

$$v_t(x) = \int_0^{\infty} w_t\left(\pi_t(x)\circ Q_{x,t}\left(r_t + v_{t+1} > s\right)\right) ds$$

$$v_T(x) = r_T(x). \tag{2.4.3}$$

Now, we can easily see that the difficulty of solving the optimization problem stated in Eq. 2.3.1 is due to the appearance of the value functions $\{v_i\}_t^{T-1}$ in the calculation of V_t. The following proposition aggregates several fundamental properties of Eq. 2.4.3.

Proposition 1. *The value function, V_t, in Eq. 2.4.3 has the following properties:*

1) As $\sup_x |w_t(x) - x| \to 0 \ \forall t \in [0, T]$, a solution $\tilde{\pi}^*$ for the standard (i.e., risk-neutral) optimization problem:

$$\max_{\pi} \tilde{V}_t(x_t, \pi),$$

also solves the optimization problem:

$$\max_{\pi} V_t(x_t, \pi);$$

2) If w_t is such that it puts all weights on the highest possible reward value, then an

optimal policy for the optimization problem:

$$\max_{\pi} V_t(x_t, \pi)$$

is obtained by considering only deterministic policies:

$$\pi_t^*(x) = \arg\max_{a \in A(x)} r_t(x, a) + v_{t+1}(x_{t+1}).$$

Proof. The proof for (1) is trivial since $\tilde{V}_t(x_t, \pi) \to V_t(x_t, \pi)$ point-wise when

$$\sup_x |w_t(x) - x| \to 0 \ \forall t \in [0, T].$$

The proof for (2) is also straightforward. By placing all probability weights on the highest reward value, it is always optimal to pick the deterministic action with the highest reward value. □

Remark 11. When there exists an action a^* such that

$$Q_{x,t}(\psi_{x,a^*} > s | x, a^*) \geq Q_{x,t}(\psi_{x,a} > s | x, a)$$
$$\forall s \in [0, \infty),\ \forall a \in A(x),\ \forall x \in \mathbb{X},\ \forall t \in [0, T],$$

where $\psi_{x,a}$ denotes the reward function with x and a fixed, then there exists a deterministic optimal policy (i.e., the policy that takes action a^* for state x at time t). Deterministic optimal policies is obtained trivially when the complement CDF of an action dominates all other complement CDFs (i.e., If $1 - F_{a_1}(x) \geq 1 - F_{a_2}(x)\ \forall a_1 \neq a_2$,

then a_1 is an optimal deterministic action). The next example demonstrates some of the difficulties in analyzing even the simplest probability weighting function. We prove the uniqueness of the optimal policy for some special cases afterwards.

Example 8. For this example, we will use the probability weighting function:

$$w(x) = 1 - (1-x)^b \; b > 1.$$

For simplicity, we deal with a two-state-two-action problem. We write the probability transition matrix as

$$Q_1 := \begin{bmatrix} Q[1,1,1] & Q[1,1,2] \\ Q[2,1,1] & Q[2,1,2] \end{bmatrix}$$

and similarly we denote the transition probability matrix of taking action 2 (i.e., a_2) as:

$$Q_2 := \begin{bmatrix} Q[1,2,1] & Q[1,2,2] \\ Q[2,2,1] & Q[2,2,2] \end{bmatrix}.$$

In other words, $Q[1,2,1]$ is the probability of starting and arriving at state 1 by taking action 2. Using the reward function $\psi(x,a,x) := x + a + x$, we study the optimization problem:

$$\max_{p_1} \int_0^\infty 1 - (1 - (\mathbb{P}(\psi > s|x,a_1) \times p_1 + \mathbb{P}(\psi > s|x,a_1) \times (1-p_1)))^2 \, ds. \quad (2.4.4)$$

By differentiating Eq. 2.4.4 with respect to p_1, we obtain

$$\begin{aligned}
&-Q[1,2,2]w'\left(-Q[1,2,2]\left(-1+p_1\right)\right)\\
&+(Q[1,1,2]-Q[1,2,1]-Q[1,2,2])\\
&\times w'\left(-(Q[1,2,1]+Q[1,2,2])\left(-1+p_1\right)+Q[1,1,2]p_1\right)\\
&+3(Q[1,1,1]+Q[1,1,2]-Q[1,2,1]-Q[1,2,2])\\
&\times w'\left(-(Q[1,2,1]+Q[1,2,2])\left(-1+p_1\right)+(Q[1,1,1]+Q[1,1,2])p_1\right).
\end{aligned}$$

Substituting $w'(p) = 2(1-p)$, we obtain an affine equation in p_1:

$$\begin{aligned}
g(p_1) = &-2\left(-3Q[1,1,1]-4Q[1,1,2]+4Q[1,2,1]\right.\\
&+3Q[1,1,1]Q[1,2,1]+4Q[1,1,2]Q[1,2,1]\\
&-4Q[1,2,1]^2+5Q[1,2,2]+3Q[1,1,1]Q[1,2,2]\\
&\left.+4Q[1,1,2]Q[1,2,2]-8Q[1,2,1]Q[1,2,2]-5Q[1,2,2]^2\right)\\
&-2\left(3Q[1,1,1]^2+6Q[1,1,1]Q[1,1,2]+4Q[1,1,2]^2\right.\\
&-6Q[1,1,1]Q[1,2,1]-8Q[1,1,2]Q[1,2,1]+4Q[1,2,1]^2\\
&-6Q[1,1,1]Q[1,2,2]-8Q[1,1,2]Q[1,2,2]\\
&\left.+8Q[1,2,1]Q[1,2,2]+5Q[1,2,2]^2\right)p_1.
\end{aligned}$$

Although the equation above is affine in p_1, its coefficients depend on the transition probabilities (Q_1 and Q_2) and the reward function ψ. This dependence makes any generalized results on the structure of the optimal policies difficult.

We observe from the example above that any optimal solution of the problem stated in Eq. 2.4.4 must satisfy $g(p_1) = 0$, which has a unique solution p_1^*. In the next theorem, we will prove that this is true in general.

Theorem 8. *Assume w_t is a strictly concave function, then the function*

$$\int_0^\infty w_t \left(\delta \circ Q_{x,t} \left(r_t + v_{t+1} > s\right)\right) ds$$

is strictly concave in δ. Furthermore, there is a unique maximizer for the optimization problem:

$$\max_{\delta \in \mathcal{P}(A(x))} \int_0^\infty w_t \left(\delta \circ Q_{x,t} \left(r_t + v_{t+1} > s\right)\right) ds.$$

Proof. We prove the concavity of the function of interest:

$$\begin{aligned}
&\int_0^\infty w_t \left(\left(\alpha\delta_1 + (1-\alpha)\,\delta_2\right) \circ Q_{x,t} \left(r_t + v_{t+1} > s\right)\right) ds \\
&= \int_0^\infty w_t \left(\int Q_{x,t} \left(r_t + v_{t+1} > s\right) d\left(\alpha\delta_1 + (1-\alpha)\,\delta_2\right)\right) ds \\
&> \alpha \left(\int_0^\infty w_t \left(\delta_1 \circ Q_{x,t} \left(r_t + v_{t+1} > s\right)\right)\right) ds \\
&+ (1-\alpha) \left(\int_0^\infty w_t \left(\delta_x \circ Q_{x,t} \left(r_t + v_{t+1} > s\right)\right)\right) ds.
\end{aligned}$$

Since the right hand side of the second equation in the theorem is strictly concave, the uniqueness of maximizer follows. □

We are only able to prove the uniqueness of the optimal policy in the positive reward case. For the more general case of both positive and negative rewards, where

we have a convex-concave measure, we were unable to prove the uniqueness of the optimal policy.

2.5 Conclusion

Non-convex reward measures are useful for modeling many real-life problems. More specifically, CPT reward measures are derived from experimental data and have been proven to model several key characteristics of human behaviors well. This inspired us to start building a rigorous theoretical foundation for their application to dynamic problems. Our effort has resulted in proving the applicability of dynamic programming equations for non-convex reward measures.

In relation to Çavuş and Ruszczyński [18], we relaxed their assumptions on the performance measures to monotonicity and positive homogeneity. In the finite-horizon case, the monotonicity assumption is important in the proof for the suitability of the dynamic programming method for the non-convex case. One of our contributions in the finite-horizon case is the assumptions on the weighting functions such that the monotonicity assumption of the performance measures is satisfied. In the discounted infinite-horizon case, the suitability of the dynamic programming method for non-convex performance measures is proven using the monotonicity assumption and the fact that they are contractions. In the transient case, the proofs are more difficult and require the utilization of k-step contractions.

In this chapter, we have established a rigorous mathematical foundation for using dynamic programming to solve Markov Decision Problems with CPT-based

reward measures. In our new framework, CPT-based reward measures, unlike the existing work where CPT-based reward measures are only applied statically or to special cases, can be applied to a larger class of problems. The development of these new MDPs is especially useful for modeling dynamic human decision making processes. Through numerical examples, we demonstrated the properties of optimal policies obtained by solving these problems. The optimal polices obtained from these problems are different from the standard polices; they are randomized rather than deterministic. This finding, perhaps not too surprising due to previous work done by Ruszczyński, suggests that deterministic optimal policies are insufficient for obtaining the optimal value function when humans are involved. It is our hope that these new proposed models will be utilized to model human risk-sensitive behaviors in dynamic settings.

Chapter 3

Cumulative Weighting Optimization

Global optimization problems are relevant in many fields (e.g., control systems, operations research, economics). There are many approaches to solving these problems. One particular approach is model-based methods, which are a class of random search methods. A model-based method iteratively updates its probability density function. At each step, additional weight is given to solution subspaces that are more likely to yield an optimal objective value. Model-based methods can be analyzed by writing down a corresponding system of differential equations similar to the well known Fokker-Planck equation, which models the evolution of probability density functions for diffusions. We propose an innovative model-based method, Cumulative Weighting Optimization (CWO), which can be proven to converge to an optimal solution. Using this rigorous theoretical foundation, we design a class of CWO-based numerical algorithms for solving global optimization problems. Interestingly, the well-known cross-entropy method is a special case of CWO-based numerical algorithms.

3.1 Introduction

Many problems in engineering and science can be formulated as global optimization problems. These problems are challenging when their objective functions are

nonlinear (e.g., non-convex, multi-modal, or badly scaled). If we are only interested in finding their local extrema and they are differentiable, then the standard local optimization method (i.e., first derivative being zero) would suffice. If there are only a few local extrema, then we can easily find a global optimal solution by evaluating all of them. However, this approach does not work on objective functions with absence of structural information (e.g., non-differentiable), or in the presence of many local extrema. Approaches developed to solve these problems can be divided into two categories: deterministic and random search algorithms. Random search algorithms can be further divided into instance-based (e.g., simulated annealing, genetic algorithm, tabu search, nested partitions, generalized hill climbing, and evolutionary programming) and model-based algorithms (e.g., annealing-adaptive search, cross-entropy (CE), model reference adaptive search (MRAS), and estimation of distribution algorithms (EDAs)). A more recent addition to model-based algorithms is model-based evolutionary optimization [88]. For the interested reader, Hu et al. have a recent survey paper on model-based methods [54], which also contains references to instance-based methods mentioned in this paragraph

We propose a new addition, inspired by Cumulative Prospect Theory (CPT), to the class of model-based methods. The new CWO-based algorithms have an intuitive connection with the risk-sensitive nature of the human decision making process.

In the rest of this chapter, we will proceed in the following sequence. In Section 3.2, we present the problem statement. In Section 3.3, we introduce the concept of

probability weighting functions. In Section 3.4, we will work with the case when

$$\mathcal{X} = \{1, 2, 3, 4, \dots\}$$

and provide the reader some insight into the construction of our probability updating equation. Later in the same section, we will prove the convergence properties for the equation. In Section 3.5, we will generalize the results of Section 3.4 to Polish spaces. Finally, in Section 3.6, we will outline our numerical algorithms and present some simulation results.

3.2 Problem

In many engineering applications, we are looking for a "best" solution based on some criterion. For example, in the well known traveling salesman problem (TSP), we are looking for the cheapest route that visits all cities and terminates at the starting point. Problems of this nature can be formulated as the following optimization problem:

$$x^* \in \arg\max_{x \in \mathcal{X}} H(x), \tag{3.2.1}$$

where x^* is an optimal solution to the problem and $\mathcal{X}$ is a non-empty, often compact, solution space (in many applications $\mathcal{X} \subset \mathcal{R}^n$). $H : \mathcal{X} \to \mathcal{R}$, the objective function, is a bounded deterministic measurable function with many local extrema. In the rest of this chapter we assume the following.

Assumption 4. *There exists a global optimal solution to Eq. 3.2.1, i.e.,* $\exists x^* \in \mathcal{X}$

such that $H(x) \leq H(x^*)\ \forall x \neq x^*,\ \forall x \in \mathcal{X}$.

In practice, this assumption is true for many optimization problems. For example, the assumption holds trivially when $\mathcal{X}$ is a finite discrete solution space. Generally, we do not assume any other structural information about the objective function (i.e., convexity, differentiability).

We can introduce a measurable strictly increasing *fitness function,* $\phi : \mathcal{R} \rightarrow \mathcal{R}^+$, and reformulate Eq. 3.2.1 as:

$$x^* \in \arg\max_{x \in \mathcal{X}} \phi\left(H(x)\right). \tag{3.2.2}$$

Since the reformulated problem guarantees the range of the new fitness-objective function (i.e., $\phi\left(H\left(\cdot\right)\right)$) will always be non-negative, and it is equivalent to the original problem, we will solve Eq. 3.2.2 in place of Eq. 3.2.1. A similar problem statement can be found in Hu et al. [54].

3.3 Probability Weighting Functions

Probability weighting functions have many applications in science and engineering. In this thesis, we are most concerned with using them to re-weight the probabilities of outcomes. Weighting is suggested by Cumulative Prospect Theory (CPT) as an important part of the human decision making process. Prospect Theory (PT), the predecessor to CPT, was suggested in the 1970s by Kahneman and Tversky [59]. They were unsatisfied with PT due to its violation of second order stochastic

dominance, which was remedied by CPT in the 1990s [83]. CPT improves PT by re-weighting the outcome cumulative probability function instead of the outcome probability density function. This new approach can also be useful for global optimization problems. The purpose of this section is to familiarize the reader with probability weighting functions, which will be used later for iteratively updating probability measures. We first introduce several definitions to assist us in our discussion.

Definition 18. A *weighting function,* $w : [0,1] \to [0,1]$, is a monotonically increasing and Lipschitz continuous function with $w(0) = 0$ and $w(1) = 1$.

We are interested in weighting functions with the additional property of optimal-seeking.

Definition 19. A weighting function, $w : [0,1] \to [0,1]$, is *optimal-seeking if*

$$w\left(\alpha x + (1-\alpha)\, y\right) > \alpha w\left(x\right) + (1-\alpha)\, w(y),\ \forall \alpha \in (0,1),\ x \neq y \in [0,1].$$

Optimal-seeking is called risk-seeking in fields modeling risk-sensitivity. From the definitions above, we can prove the following proposition.

Proposition 2. *An optimal-seeking weighting function satisfies the inequality*

$$w(z) > z,\ \forall z \in (0,1)\,.$$

Proof. Let $x = 1$, $y = 0$, and $\alpha = z$ in the definition for optimal-seeking weighting functions. The proof follows trivially. □

Next, we will be more concrete and provide several examples of risk-seeking probability weighting functions.

Assumption 5. *w is an optimal-seeking weighting function.*

Example 9. A simple *polynomial* weighting function has the form:

$$w(p) = 1 - (1-p)^b, \ b > 1. \tag{3.3.1}$$

Another more complicated weighting function involving *exponentials* has the form:

$$w(p) = \frac{e^{cp} - 1}{e^c - 1}, \tag{3.3.2}$$

where $c < 0$. There are other parametric weighting functions, which can be found in [29].

An optimal-seeking weighting function tends to place more weight on highly unlikely, yet highly rewarding outcomes. In particular, it is used to overweight the probabilities of unlikely events (i.e., events with small probabilities) and underweight the probabilities of highly likely events. More specifically, we can apply an optimal-seeking weighting function to the cumulative distribution function of the outcomes, as in the example below.

Example 10. We are given a die and asked to roll it once. We are told that we will be given a payoff that is equivalent to the outcome of the roll. For example, if we rolled a 1, then we would be given a \$1 reward. We assume the die is fair, and

calculate the expected payoff for both the risk-neutral and optimal-seeking cases. We use R to denote the random variable associated with the outcomes of this game.

The risk-neutral expected payoff is calculated as:

$$\mathbb{E}[R] = \sum_{n=1}^{6} 1 - F(n) = \sum_{n=1}^{6} \frac{n}{6} = \frac{21}{6} \approx 3.5,$$

where $F(n)$ is the cumulative distribution function at outcome n. Using the polynomial weighting function in Eq. 3.3.1 with $b = 2$, we have our optimal-seeking re-weighted expected payoff:

$$\mathbb{E}^w[R] = \sum_{n=1}^{6} w(1 - F(n)) = \sum_{n=1}^{6} w(\frac{n}{6}) = \sum_{n=1}^{6} 1 - (1 - \frac{n}{6})^2 = \frac{161}{36} \approx 4.47222.$$

Remark 12. The reader should note that the re-weighting is applied to the good-news function[1] (i.e., the probability of the outcome exceeding a threshold). In other words, the unlikely events are events whose payoff exceeded some threshold. It should be noted that the optimal-seeking re-weighted expected payoff is greater than that of the risk-neutral. This will be an important feature in proving the convergence of the CWO method.

[1] In other fields, the good-news function is also known as the survival function or the reliability function.

3.4 Discrete Solution Space

We are trying to find an solution for Eq. 3.2.2 assuming that

$$\mathcal{X} := \{1, 2, 3, 4, \dots\}.$$

We further assume the discrete topology for $\mathcal{X}$. The discrete solution space case should offer the reader some intuitive insight into the workings of the CWO method. In the next section, we will present analogous results on Polish spaces.

We denote the set of optimal measures on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ by

$$\mathcal{P}^*_{\mathcal{X}} := \{\mathbb{P} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}(\mathcal{X}^*) = 1\},$$

where $\mathcal{X}^*$ is the set of all optimal solutions, i.e.,

$$\mathcal{X}^* := \{i^* \in \mathcal{X} | H(i) \leq H(i^*),\ \forall i \in \mathcal{X}\},$$

and $\mathcal{P}_{\mathcal{X}}$ is the set of all possible probability measures over $\mathscr{B}(\mathcal{X})$. It should not surprise the reader that if we can find an element of $\mathcal{P}^*_{\mathcal{X}}$, then we have found a solution to the global optimization problem stated in Eq. 3.2.2. We assume $\mathcal{X}^*$ has only a finite number of elements.

Assumption 6. *The objective function, $H : \mathcal{X} \to \mathcal{R}$, has a finite number of optimal*

solutions, i.e., the set

$$\mathcal{X}^* := \{i^* \in \mathcal{X} | H(i) \leq H(i^*),\ \forall i \in \mathcal{X}\}$$

has a finite number of elements.

Proposition 3. *$\mathcal{P}^*_{\mathcal{X}}$ and $\mathcal{P}_{\mathcal{X}}$ are both non-empty sets.*

Proof. Using Assumption 4 we know that a Dirac measure concentrated at any $i^* \in \mathcal{X}^*$ is in $\mathcal{P}^*_{\mathcal{X}}$, which is a subset of $\mathcal{P}_{\mathcal{X}}$. □

Our objective is to restrict the temporal evolution of a probability measure such that it will eventually concentrate its probability density at one of the optimal solutions. This evolution can be defined on a measurable space, $(\mathcal{X} \times \mathcal{R}^+, \mathscr{B}(\mathcal{X} \times \mathcal{R}^+))$, where $\mathcal{X}$ is the given solution space and $\mathcal{R}^+$ represents time. If the evolution happens in discrete time or iteration steps, then $\mathcal{R}^+$ can be replaced by $\{0, 1, 2, 3, \dots\}$. To solve Eq. 3.2.1, we want to find a probability measure $\mathbb{P}$ and a $t^* \in \mathcal{R}^+$ such that

$$\mathbb{P}(\{(t, i) | i \in \mathcal{X}^*\}) = \mathbb{P}(\{(t, i) | i \in \mathcal{X}\}),\ \forall t \geq t^*. \tag{3.4.1}$$

In other words, $\mathbb{P}$ at some finite time t^* is a member of $\mathcal{P}^*_{\mathcal{X}}$. We denote the resulting probability space by [2]

$$\left(\mathcal{X} \times \mathcal{R}^+, \mathscr{B}\left(\mathcal{X} \times \mathcal{R}^+\right), \mathbb{P}\right).$$

[2]For simplicity, we choose to work with $(\mathcal{X} \times \mathcal{R}^+, \mathscr{B}(\mathcal{X} \times \mathcal{R}^+), \mathbb{P})$ instead of $\left(\mathcal{X}^{\mathcal{R}^+}, \mathscr{B}\left(\mathcal{X}^{\mathcal{R}^+}\right), \tilde{\mathbb{P}}\right)$, which is the common practice for defining a stochastic process. It should not be hard for the reader to see that $\left(\mathcal{X}^{\mathcal{R}^+}, \mathscr{B}\left(\mathcal{X}^{\mathcal{R}^+}\right), \tilde{\mathbb{P}}\right)$ can induce a measure $\mathbb{P}$ on the measurable space $(\mathcal{X} \times \mathcal{R}^+, \mathscr{B}(\mathcal{X} \times \mathcal{R}^+))$.

At each time t, $\mathbb{P}$ induces a probability measure on the measurable space $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$,

$$\mathbb{P}_t(B_{\mathcal{X}}) := \mathbb{P}(\{(t,i)|i \in B_{\mathcal{X}}\}),\ \forall B_{\mathcal{X}} \in \mathscr{B}(\mathcal{X}), \tag{3.4.2}$$

resulting in a probability space $(\mathcal{X}, \mathscr{B}(\mathcal{X}), \mathbb{P}_t)$. Conversely, if we know $\mathbb{P}_t$ at all times, then we can construct a $\mathbb{P}$ that satisfies Eq. 3.4.2[3]. The coordinate random variable is denoted by X (i.e., $X(i) = i,\ i \in \mathcal{X}$). Similarly, the outcome random variable is denoted by Y, where $Y = \phi(H(X))$. We denote the set of all possible outcomes from evaluating $\phi(H(\cdot))$ over $\mathcal{X}$ by

$$\mathcal{Y} := \left\{y \in \mathcal{R}^+ | \exists i \in \mathcal{X}\ s.t.\ y = \phi(H(i))\right\}.$$

A sensible next step is to write down the dynamics of $\mathbb{P}_t$ with respect to time (i.e., $\dot{\mathbb{P}}_t$), and interestingly we will be able to prove that

$$\mathbb{E}_t[\phi(H(X))] := \int_{\mathcal{X}} \phi(H(i))\, d\mathbb{P}_t := \int_{\mathcal{Y}} y d\mathbb{P}_t^{\phi(H(X_t))} \tag{3.4.3}$$

is a strictly increasing function with increasing t. In the equation above, $\mathbb{P}_t^{\phi(H(X_t))}$ is the probability measure of the random variable $\phi(H(X_t))$ induced by X_t. Of course, probability weighting functions from Section 3.3 play a key role in the equations for $\dot{\mathbb{P}}_t$[4]. Using Eq. 3.4.3 along with Lyapunov stability analysis, we will conclude the convergence of $\mathbb{P}_t$ to optimal solutions (i.e., elements of $\mathcal{P}_{\mathcal{X}}^*$). The following examples

[3]We can construct such a measure by using the definition: $\mathbb{P}(B_{\mathcal{X}} \times B_{\mathcal{T}}) = \int_{B_{\mathcal{X}} \times B_{\mathcal{T}}} \mathbb{P}_t(dx)\, dt,\ B_{\mathcal{X}} \times B_{\mathcal{T}} \in \mathscr{B}(\mathcal{X} \times \mathcal{R}^+)$.

[4]We opted for the notation $\mathbb{P}_t$ instead of $\mathbb{P}_t^w$ for simplicity, but the reader should be mindful of $\mathbb{P}_t$'s dependence on w.

illustrate our approach.

Example 11. (Distinct Outcomes)

We are given a finite solution space $\mathcal{X} = [1, 2, 3]$, and its corresponding outcome space $\mathcal{Y} = [\phi(H(1)), \phi(H(2)), \phi(H(3))] \subset \mathcal{R}^+$. In addition, we assume the outcomes are distinct (i.e., $\phi(H(3)) > \phi(H(2)) > \phi(H(1)) \geq 0$). We denote the vector of probabilities for non-intersecting outcome events,

$$y_i(t) = \mathbb{P}_t(\phi(H(i)) \leq \phi(H(X)) < \phi(H(i+1)),\ i \in \mathcal{X} \text{ with } \phi(H(4)) = \infty,$$

by $[y_1(t), y_2(t), y_3(t)]$. Furthermore, we denote the probabilities on elements of the solution space by $[x_1(t), x_2(t), x_3(t)]$, respectively.

To avoid confusion, we will refer to $[y_1(t), y_2(t), y_3(t)]$ as the *outcome probability vector*, and $[x_1(t), x_2(t), x_3(t)]$ as the *solution probability vector*. Note, $\mathbb{P}_t$ in the previous discussion is equivalent conceptually to the solution probability vector.

Our goal is to write down the dynamic equation for the solution probability vector. However, in order to do that, we need to write down the dynamic equation for the outcome probability vector. Using an optimal-seeking weighting function, w, the dynamics of the outcome probability vector can be written as:

$$\begin{aligned}
\frac{dy_1}{dt} &= (w(1) - w(y_2 + y_3)) - y_1 \\
\frac{dy_2}{dt} &= (w(y_2 + y_3) - w(y_3)) - y_2 \\
\frac{dy_3}{dt} &= w(y_3) - y_3.
\end{aligned}$$

We define the matrix G, which in the nonlinear Markov processes literature is called a generator (see [61, 62]), as

$$G(y_1, y_2, y_3) := \begin{bmatrix} -w(y_2+y_3) & w(y_2+y_3)-w(y_3) & w(y_3) \\ 1-w(y_2+y_3) & w(y_2+y_3)-w(y_3)-1 & w(y_3) \\ 1-w(y_2+y_3) & w(y_2+y_3)-w(y_3) & w(y_3)-1 \end{bmatrix}.$$

Using the matrix G, we can write down the outcome and solution probability vector equations as:

$$\begin{bmatrix} \dot{} \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} G(y_1, y_2, y_3)$$

$$\begin{bmatrix} \dot{} \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} G(x_1, x_2, x_3).$$

From Proposition 2 we see that $w(y) > y \ \forall y \in (0,1)$, which implies

$$\frac{dy_3}{dt} > 0, \frac{dx_3}{dt} > 0 \ \forall x_3, y_3 \in (0,1) \text{ and } \frac{dy_3}{dt} = 0, \frac{dx_3}{dt} = 0 \ x_3, y_3 = \{0,1\}.$$

Since the best outcome, $x = 3$, will monotonically increase in its probability weight, and the increase in probability weight has to come from the non-best

solutions, the non-best solutions will eventually die out, i.e.,

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \; as \; t \to \infty.$$

We will prove our convergence assertion more rigorously later in this section.

What would happen if two or more members in the solution space might map to the same outcome?

Example 12. (Non-Distinct Outcomes)

Consider the case when there are more than one solution mapping to the same outcome. We are given an optimal-seeking probability weighting function, w, and a solution space $\mathcal{X} = [1, 2, 3, 4]$. Assume we know that $\phi(H(4)) = \phi(H(3)) > \phi(H(2)) > \phi(H(1)) \geq 0$. Now, the outcome space, $\mathcal{Y} = [\phi(H(3)), \phi(H(2)), \phi(H(1))]$, has fewer elements than the solution space. Following the similar line of logic as in the previous example, the outcome probability vector equation is written as :

$$\begin{aligned} \frac{dy_1}{dt} &= (w(1) - w(y_2 + y_3)) - y_1 \\ \frac{dy_2}{dt} &= (w(y_2 + y_3) - w(y_3)) - y_2 \\ \frac{dy_3}{dt} &= w(y_3) - y_3, \end{aligned}$$

where y_i is defined as

$$\begin{aligned} y_1(t) &= \mathbb{P}_t\left(\phi\left(H(1) \leq \phi\left(H\left(X\right)\right) < \phi\left(H(2)\right)\right.\right. \\ y_2(t) &= \mathbb{P}_t\left(\phi\left(H\left(2\right)\right) \leq \phi\left(H\left(X\right)\right) < \phi\left(H(3)\right)\right. \\ y_3(t) &= \mathbb{P}_t\left(\phi\left(H\left(3\right)\right) \leq \phi\left(H\left(X\right)\right) < \infty\right). \end{aligned}$$

We define the matrix G_y as:

$$G_y(y_1, y_2, y_3) := \left[\begin{array}{ccc} -w(y_2+y_3) & w(y_2+y_3)-w(y_3) & w(y_3) \\ 1-w(y_2+y_3) & w(y_2+y_3)-w(y_3)-1 & w(y_3) \\ 1-w(y_2+y_3) & w(y_2+y_3)-w(y_3) & w(y_3)-1 \end{array}\right].$$

As in the distinct outcome case, the solution probability vector is written as:

$$\begin{aligned} \frac{dx_1}{dt} &= (w(1) - w(x_2+x_3+x_4)) - x_1 \\ \frac{dx_2}{dt} &= (w(x_2+x_3+x_4) - w(x_3+x_4)) - x_2 \\ \frac{dx_3}{dt} &= \beta w(x_3+x_4) - x_3 \\ \frac{dx_4}{dt} &= (1-\beta)w\left(x_3+x_4\right) - x_4,\ \beta \in [0,1]\,, \end{aligned}$$

and its corresponding generator is

$$G_{x,\beta}(x_1, x_2, x_3, x_4) :=$$

$$\begin{bmatrix} -w(x_2+x_3+x_4) & w(x_2+x_3+x_4)-w(x_3+x_4) & \beta w(x_3+x_4) & (1-\beta)\,w(x_3+x_4) \\ 1-w(x_2+x_3+x_4) & w(x_2+x_3+x_4)-w(x_3+x_4)-1 & \beta w(x_3+x_4) & (1-\beta)\,w(x_3+x_4) \\ 1-w(x_2+x_3+x_4) & w(x_2+x_3+x_4)-w(x_3+x_4) & \beta w(x_3+x_4)-1 & (1-\beta)\,w(x_3+x_4) \\ 1-w(x_2+x_3+x_4) & w(x_2+x_3+x_4)-w(x_3+x_4) & \beta w(x_3+x_4) & (1-\beta)\,w(x_3+x_4)-1 \end{bmatrix}.$$

We can write the dynamic equation for the outcome and solution probability vectors more compactly as:

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix} G_y(y_1, y_2, y_3)$$

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} G_{x,\beta}(x_1, x_2, x_3).$$

From Proposition 2 we see that $w(y) > y,\ \forall y \in (0,1)$, which implies

$$\frac{dy_3}{dt} > 0, \frac{dx_3}{dt} + \frac{dx_4}{dt} > 0,\ \forall x_3, x_4, y_3 \in (0,1)$$

and

$$\frac{dy_3}{dt} = 0, \frac{dx_3}{dt} = 0, \frac{dx_4}{dt} = 0,\ x_3, x_4, y_3 = \{0,1\}.$$

Ultimately, we have the best solutions for the problem, $x = 3$ and $x = 4$,

monotonically increasing in their probability weights. Since the increase in probability weights has to come from the non-best solutions, they will eventually die out, i.e.,

$$\begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \\ x_4(t) \end{bmatrix} \Rightarrow \begin{bmatrix} 0 \\ 0 \\ \alpha_3 \\ \alpha_4 \end{bmatrix} \quad as\ t \to \infty,\ \alpha_3, \alpha_4 \geq 0,\ \alpha_3 + \alpha_4 = 1.$$

Remark 13. A key feature of the non-distinct outcomes case is the appearance of the β-function (later, this will be called a distribution rule). For instance, in the example above, the β-function could be:

$$\mathbb{P}_t\left(X = i | Y = \phi\left(H(i)\right)\right) = \frac{x_i(t)}{\sum_{j:\phi(H(j))=\phi(H(i))} x_j(t)}.$$

An interesting observation from the examples above is that the solution and outcome probability vector equations fall into a category of equations called the nonlinear Fokker-Planck equation (see [61, 42]). In addition, the examples suggest that by propagating the solution and outcome probability vectors appropriately, we can concentrate the probability weights on optimal solutions. The key step forward is writing down the general equations for the solution and outcome probability vectors.

The generalized solution probability vector equation has the form:

$$\begin{aligned}\frac{dx_i(t)}{dt} &= \beta_i(t)\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))} x_j(t)\right) - w\left(\sum_{j:\phi(H(j))>\phi(H(i))} x_j(t)\right)\right) \\ &\cdots \quad - x_i(t)\ \forall i \in \mathcal{X} \qquad (3.4.4)\end{aligned}$$

$$\sum_{\phi(H(i))=y} \beta(i,y,t) = 1\ \forall y \in \mathcal{Y}\ \forall t \in \mathcal{R}^+, \qquad (3.4.5)$$

where $x_i : \mathcal{R}^+ \to [0,1]$ is the probability measure assigned to an element $i \in \mathcal{X}$, and $\beta_i(t) := \beta(i, \phi(H(i)), t)$ is a distribution rule defined below. In Eq. 3.4.4, the difference between the first w distorted term and the second w distorted term is the event $\phi(H(j)) = \phi(H(i))$. Wang et al. (see [88]) have an alternative set of evolution equations, also nonlinear Fokker-Planck equations, motivated by evolutionary game theory. As the reader will see later, we reach the same convergence results as Wang et al. in [88] with a modified approach.

Definition 20. A *distribution rule* with respect to a given objective function, $\phi(H(\cdot))$, is a mapping $\beta : \mathcal{X} \times \mathcal{Y} \times \mathcal{R}^+ \to [0,1]$ such that

$$\sum_{\phi(H(i))=y} \beta(i,y,t)\,dx = 1\ \forall y \in \mathcal{Y}\ \forall t \in \mathcal{R}^+.$$

Connecting this equation with the discussion at the beginning of this section, the reader should note that

$$\mathbb{P}_t(X = i) = x_i(t)\ \forall i \in \mathcal{X}.$$

The generalized outcome probability vector equation has the form:

$$\frac{dy_z(t)}{dt} = w\left(\sum_{j:j\geq z} y_j(t)\right) - w\left(\sum_{j:j>z} y_j(t)\right) - y_z(t) \;\; \forall z \in \mathcal{Y}.$$

We pay special attention to the best outcome equation:

$$\frac{dy_*(t)}{dt} = w\left(\sum_{j:j\geq *} y_j(t)\right) - y_*(t),$$

where $* := \phi(H(i^*)) \;\; i^* \in \mathcal{X}^*$.

In the rest of this section, we want to study the convergence properties of Eq. 3.4.4. Furthermore, we want to understand the stability properties, in the Lyapunov sense, of its limit points. The first step in understanding Eq. 3.4.4 is to understand the existence and uniqueness of its solutions. The outline of our proof for the next theorem follows [65] and [51].

Theorem 9. *For each $x(0) \in \mathcal{P}_{\mathcal{X}}$, the ordinary differential equation 3.4.4 has a unique solution for $t \in \mathcal{R}^+$. Here, $\beta : \mathcal{X} \times \mathcal{Y} \times \mathcal{R}^+ \to [0,1]$ is a distribution rule*[5].

Proof. We use the total variation norm, $\|\cdot\|$, on a σ-finite signed measure space over $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$:

$$\|x(t)\| = \sup_{A\in\mathscr{B}(\mathcal{X})} \sum_{i\in A} |x_i(t)|.$$

Since $x(t) \in \mathcal{P}_{\mathcal{X}}$ is a probability measure $\forall t$, and $|\beta_i| \leq 1$, we know the following

[5]We will provide the definition for distribution rules in more general spaces in the next section. For the proof of this theorem, we are only using the fact that it is a bounded function. In the future, β could depend on both $i \in \mathcal{X}$ and $x(t) \in \mathcal{P}_{\mathcal{X}}$.

inequalities hold:

$$\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\beta_i(t)\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j(t)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j(t)\right)\right)-x_i(t)\right|$$
$$\leq\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\beta_i(t)\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j(t)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j(t)\right)\right)\right|$$
$$+\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}|x_i(t)|\leq 2.$$

Hence, we conclude the right hand side of Eq. 3.4.4 is bounded by 2. Next, we need

to prove that the right hand side of Eq. 3.4.4 is Lipschitz continuous.

$$
\begin{aligned}
&\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\left(\beta_i\left(t\right)\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^1\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^1\left(t\right)\right)\right)-x_i^1\left(t\right)\right)\right.\\
&\left.-\left(\beta_i\left(t\right)\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^2\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^2\left(t\right)\right)\right)-x_i^2\left(t\right)\right)\right|\\
&\leq\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^1\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^1\left(t\right)\right)\right)-x_i^1\left(t\right)\right.\\
&\quad\left.-\left(\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^2\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^2\left(t\right)\right)\right)-x_i^2\left(t\right)\right)\right|\\
&\quad\leq\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^1\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^1\left(t\right)\right)\right.\\
&\qquad\left.-\left(w\left(\sum_{j:\phi(H(j))\geq\phi(H(i))}x_j^2\left(t\right)\right)-w\left(\sum_{j:\phi(H(j))>\phi(H(i))}x_j^2\left(t\right)\right)\right)\right|\\
&\qquad+\left|x_i^1(t)-x_i^2(t)\right|\\
&\qquad\leq K\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\sum_{j:\phi(H(j))=\phi(H(i))}x_j^1\left(t\right)-\sum_{j:\phi(H(j))=\phi(H(i))}x_j^2\left(t\right)\right|\\
&\qquad+\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|x_i^1(t)-x_i^2(t)\right|\\
&\quad\leq K\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|\sum_{j:\phi(H(j))=\phi(H(i))}x_j^1\left(t\right)-x_j^2\left(t\right)\right|+\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|x_i^1(t)-x_i^2(t)\right|\\
&\qquad\leq K\sup_{A\in\mathscr{B}(\mathcal{X})}\sum_{i\in A}\left|x_i^1\left(t\right)-x_i^2\left(t\right)\right|+\left\|x^1(t)-x^2(t)\right\|\\
&\qquad\leq K\left\|x^1(t)-x^2(t)\right\|+\left\|x^1(t)-x^2(t)\right\|\leq\left(K+1\right)\left\|x^1(t)-x^2(t)\right\|.
\end{aligned}
$$

Hence, the right hand side of Eq. 3.4.4 is Lipschitz continuous in x with the constant $K+1$, where K is the Lipschitz constant for w (w is defined to be Lipschitz

continuous; see Definition 18). Using [90, Corollary 3.9], we conclude that Eq. 3.4.4 with an initial measure $x(0) \in \mathcal{P}_{\mathcal{X}}$ has a unique solution $x(t)\ \forall t \in \mathcal{R}^{+}$. □

The next Lemma is needed in Theorem 12, which shows the monotonically increasing nature of $\mathbb{E}_t\left[\phi\left(H\left(X\right)\right)\right]$.

Lemma 6. *Given an optimal-seeking weighting function, w, there exists a $\tilde{\zeta}$ such that*

$$\sum_{\zeta \in \mathcal{Y}} \zeta \left(w\left(\sum_{j:Y \geq \zeta} y_j\left(t\right) \right) - w\left(\sum_{j:Y > \zeta} y_j\left(t\right) \right) - y_\zeta\left(t\right) \right) \tag{3.4.6}$$

can be decomposed into the sum of its non-negative and negative parts:

$$\underbrace{\sum_{\zeta \geq \tilde{\zeta}} \zeta \left(w\left(\sum_{j:Y \geq \zeta} y_j\left(t\right) \right) - w\left(\sum_{j:Y > \zeta} y_j\left(t\right) \right) - y_\zeta\left(t\right) \right)}_{non-negative}$$
$$\underbrace{\sum_{\zeta < \tilde{\zeta}} \zeta \left(w\left(\sum_{j:Y \geq \zeta} y_j\left(t\right) \right) - w\left(\sum_{j:Y > \zeta} y_j\left(t\right) \right) - y_\zeta\left(t\right) \right)}_{negative}.$$

In other words, we can write Eq. 3.4.6 as the sum of its non-negative and negative parts.

Proof. Since w is a monotonically increasing function, it satisfies

$$\frac{w\left(\sum_{j:Y \geq \zeta} y_j\left(t\right)\right) - w\left(\sum_{j:Y > \zeta} y_j\left(t\right)\right)}{y_\zeta\left(t\right)} \geq 0.$$

Furthermore, since w is an optimal-seeking function we have

$$\frac{w\left(\sum_{j:Y\geq\zeta_1} y_j(t)\right) - w\left(\sum_{j:Y>\zeta_1} y_j(t)\right)}{y_{\zeta_1}(t)} > \frac{w\left(\sum_{j:Y\geq\zeta_2} y_j(t)\right) - w\left(\sum_{j:Y>\zeta_2} y_j(t)\right)}{y_{\zeta_2}(t)} \quad \forall \zeta_1 \geq \zeta_2 \in \mathcal{Y}. \quad (3.4.7)$$

In addition, since $w(0) = 0$ and $w(1) = 1$, we know that

$$\frac{w\left(\sum_{j:Y\geq\zeta} y_j(t)\right) - w\left(\sum_{j:Y>\zeta} y_j(t)\right)}{y_{\zeta}(t)} > 1$$

for some $\zeta \in \mathcal{Y}$. From Eq. 3.4.7, we know that if ζ_2 satisfies the above inequality, then so does $\zeta_1 \geq \zeta_2 \in \mathcal{Y}$. Hence, we conclude that $\tilde{\zeta}$ is the smallest such ζ. □

At the beginning of this section, we stated implicitly that if we can find an element of $\mathcal{P}^*_{\mathcal{X}}$, then we have found a solution to the global optimization problem stated in Eq. 3.2.2. The theorems below present a blueprint, through the use of Eq. 3.4.4, to obtain an element of $\mathcal{P}^*_{\mathcal{X}}$. In Theorem 9, an initial point can be any element of $\mathcal{P}_{\mathcal{X}}$. As we have discovered, $\mathcal{P}_{\mathcal{X}}$ is too large a set to initialize Eq. 3.4.4 to guarantee as $t \to \infty$ the solution probability vector, $x(t)$, will be an element of $\mathcal{P}^*_{\mathcal{X}}$. Hence, we need to constrain our initial points to a smaller set.

Definition 21. We denote the set of all $x(0)$ for which there exists an optimal solution, $i^* \in \mathcal{X}^*$, such that $x_{i^*}(0) > 0$ by $\mathcal{O}$.

In other words, $\mathcal{O}$ contains all initial probability vectors with nonzero weights on at least one optimal solution. The next theorem proves the total probability

measure on the optimal solution set will converge to 1 as $t \to \infty$. On the other hand, the total probability measure on the non-optimal solution set will converge to 0 as $t \to \infty$.

Theorem 10. *If $x(t)$ is a solution for Eq. 3.4.4, then it satisfies the following with initial points in $\mathcal{O}$ (i.e., $x(0) \in \mathcal{O}$):*

1) The total probability weight on the optimal solutions, $\sum_{i \in \mathcal{X}^*} x_i(t)$, is a monotonically increasing function of t. In fact, it converges to 1 as $t \to \infty$;

2) The probability of any non-optimal solution, $x_i(t) : \mathcal{R}^+ \to [0,1],\ i \notin \mathcal{X}^*$, approaches zero as $t \to \infty$.

Proof. We know that

$$\sum_{i \in \mathcal{X}^*} x_i = y_*,$$

hence we only need to prove y_* is a monotonically increasing function of t. Writing down the equation for y_*,

$$\frac{dy_*(t)}{dt} = w\left(\sum_{j:j \geq *} y_j(t)\right) - y_*(t),$$

and from Proposition 2:

$$w\left(\sum_{j:j \geq *} y_j(t)\right) > y_*(t),$$

we conclude

$$\frac{dy_*(t)}{dt} > 0\ \forall y_*(t) \neq 1,\ \textit{and}\ \frac{dy_*(t)}{dt} = 0 \text{ when } y_*(t) = 1.$$

Since $x(0) \in \mathcal{O}$ implies $y_*(0) > 0$, the first claim is proved.

The second claim follows from the first claim. Since $y_*(\infty) = 1$, and x is a solution probability vector (i.e., sum of x_is is 1), we can conclude the following:

$$\begin{aligned} \lim_{t\to\infty} \sum_{i\in\mathcal{X}} x_i(t) &= \lim_{t\to\infty} \sum_{i\in\mathcal{X}^*} x_i(t) + \sum_{i\notin\mathcal{X}^*} x_i(t) = 1 + \lim_{t\to\infty} \sum_{i\notin\mathcal{X}^*} x_i(t) = 1 \\ \implies \lim_{t\to\infty} \sum_{i\notin\mathcal{X}^*} x_i(t) &= 0 \implies \lim_{t\to\infty} x_i(t) = 0 \ \forall i \notin \mathcal{X}^*. \end{aligned}$$

□

We are interested in finding the limit points of Eq. 3.4.4. Ideally, these limit points should be elements in $\mathcal{P}^*_{\mathcal{X}}$. This is accomplished by picking the initial point set more carefully.

Definition 22. We define the *limit set* of a differential equation starting from an element $x(0) \in \mathcal{I}$ as

$$\mathcal{E}_{\mathcal{I}} := \left\{ x_\infty \in \mathcal{P}_{\mathcal{X}} | x_\infty = \lim_{t\to\infty} x(t),\ x(0) \in \mathcal{I} \right\}.$$

Remark 14. The limit set is invariant with respect to Eq. 3.4.4. More specifically, once x enters the set, it will not exit the set under Eq. 3.4.4. The author is not the first to introduce the concept of an initial point dependent limit set (cf. [87]).

We characterize the limit set of Eq. 3.4.4 when $x(0) \in \mathcal{O}$ in the following theorem.

Theorem 11. *The limit set of Eq. 3.4.4 started in $\mathcal{O}$ is $\mathcal{P}_{\mathcal{X}}^{*}$, i.e.,*

$$\mathcal{E}_{\mathcal{O}} = \mathcal{P}_{\mathcal{X}}^{*} := \left\{ x \in \mathcal{P}_{\mathcal{X}} | \sum_{i^{*} \in \mathcal{X}^{*}} x_{i^{*}} = 1 \right\}.$$

Proof. To prove the first claim, we will first prove $\mathcal{E}_{\mathcal{O}} \supset \mathcal{P}_{\mathcal{X}}^{*}$, then we will prove $\mathcal{E}_{\mathcal{O}} \subset \mathcal{P}_{\mathcal{X}}^{*}$. The first case, $\mathcal{E}_{\mathcal{O}} \supset \mathcal{P}_{\mathcal{X}}^{*}$, can be trivially proved by taking an element $x \in \mathcal{P}_{\mathcal{X}}^{*}$, we notice that $x \in \mathcal{O}$, and by definition of $\mathcal{E}_{\mathcal{O}}$ (i.e., the limit set of Eq. 3.4.4 starting from $\mathcal{O}$), we conclude $x \in \mathcal{E}_{\mathcal{O}}$.

Now we proceed to prove $\mathcal{E}_{\mathcal{O}} \subset \mathcal{P}_{\mathcal{X}}^{*}$. We prove by contradiction. Assume there is an element $e \in \mathcal{E}_{\mathcal{O}}$, but not in $\mathcal{P}_{\mathcal{X}}^{*}$ such that:

$$\begin{aligned} \dot{e}_i(t) \quad &= \beta_i(t) \left(w\left(\textstyle\sum_{j:\phi(H(j)) \geq \phi(H(i))} e_j(t) \right) - w\left(\textstyle\sum_{j:\phi(H(j)) > \phi(H(i))} e_j(t) \right) \right) \quad - e_i(t) \\ & e_i(0) \geq 0, \ \lim_{t \to \infty} e_i(t) > 0 \quad i \notin \mathcal{X}^{*}. \end{aligned}$$

This contradicts the second claim of Theorem 10, where $e_i(\infty) = 0$. □

The next theorem shows the monotonically increasing nature of $\mathbb{E}_t[\phi(H(X))]$, which will be useful later in proving some stability properties of Eq. 3.4.4.

Theorem 12. *Let $x(t)$ be a solution of the dynamics represented by Eq. 3.4.4 with an initial point in $\mathcal{O}$. Then the following statements hold:*

1) The expected outcome, i.e., $\mathbb{E}_t[\phi(H(X))] := \sum_{i \in \mathcal{X}} \phi(H(i)) x_i(t)$, is monotonically increasing with t;

2) If $x(t) \notin \mathcal{E}_{\mathcal{O}}$ for any $t \in \mathcal{R}^{+}$, then $\mathbb{E}_t[\phi(H(X))]$ is strictly increasing with t.

Proof. We start our proof by differentiating the average outcome function:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_t\left[\phi\left(H\left(X\right)\right)\right] &= \frac{d}{dt}\mathbb{E}_t\left[Y\right] \\
&= \sum_{\zeta\in\mathcal{Y}}\zeta\frac{dy_\zeta(t)}{dt} \\
&= \sum_{\zeta\in\mathcal{Y}}\zeta\left(w\left(\sum_{j:Y\geq\zeta}y_j\left(t\right)\right)-w\left(\sum_{j:Y>\zeta}y_j\left(t\right)\right)-y_\zeta\left(t\right)\right) \\
&= \sum_{\zeta\geq\tilde{\zeta}}\zeta\left(w\left(\sum_{j:Y\geq\zeta}y_j\left(t\right)\right)-w\left(\sum_{j:Y>\zeta}y_j\left(t\right)\right)-y_\zeta\left(t\right)\right) \\
&+ \sum_{\zeta<\tilde{\zeta}}\zeta\left(w\left(\sum_{j:Y\geq\zeta}y_j\left(t\right)\right)-w\left(\sum_{j:Y>\zeta}y_j\left(t\right)\right)-y_\zeta\left(t\right)\right) \quad \text{(Lemma. 6)} \\
&\geq \tilde{\zeta}\sum_{\zeta\in\mathcal{Y}}\left(w\left(\sum_{j:Y\geq\zeta}y_j\left(t\right)\right)-w\left(\sum_{j:Y>\zeta}y_j\left(t\right)\right)-y_\zeta\left(t\right)\right) \\
&= \tilde{\zeta}\times 0=0.
\end{aligned}$$

Here, the last equality holds because

$$\sum_{\zeta\in\mathcal{Y}}w\left(\sum_{j:Y\geq\zeta}y_j\left(t\right)\right)-w\left(\sum_{j:Y>\zeta}y_j\left(t\right)\right)=1, \text{ and } \sum_{\zeta\in\mathcal{Y}}y_\zeta\left(t\right)=1.$$

The first claim is proved.

The second claim is proved by contradiction. We assume that $x(t)$ is not in the limit set, and

$$\frac{d}{dt}\mathbb{E}_t\left[\phi\left(H\left(X\right)\right)\right]=0.$$

Along with Theorem 10, the equality above implies that $\phi\left(H\left(X\right)\right)$ is equal to a constant $C=\sup_{i\in\mathcal{X}}\phi\left(H\left(i\right)\right)$, which means $x\left(t\right)$ has all its probability mass on the optimal solutions. From Theorem 11, we know a limit point has its probability mass

on the optimal solutions. However, we assumed $x(t)$ is not a limit point, hence we reach a contradiction. □

We will now proceed to prove some stability properties of Eq. 3.4.4, but first we need to introduce our definitions of stability given a metric d.

Definition 23. Let $\mathcal{L}$ be a subset of $\mathcal{P}_{\mathcal{X}}$. For a point $x(t) \in \mathcal{P}_{\mathcal{X}}$, we define the distance between $x(t)$ and $\mathcal{L}$ as

$$d\left(x(t), \mathcal{L}\right) := \inf\left\{d\left(x(t), q\right),\ \forall q \in \mathcal{L}\right\}.$$

$\mathcal{L}$ is called *Lyapunov stable* if for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$d\left(x\left(0\right), \mathcal{L}\right) < \delta \Rightarrow d\left(x\left(t\right), \mathcal{L}\right) < \epsilon,\ \forall t > 0.$$

Lyapunov was also interested in other stronger types of stability.

Definition 24. Let $\mathcal{L}$ be a subset of $\mathcal{P}_{\mathcal{X}}$. $\mathcal{L}$ is called *asymptotically stable* if $\mathcal{L}$ is Lyapunov stable, and there exists a $\delta > 0$ such that

$$d\left(x\left(0\right), \mathcal{L}\right) < \delta \Rightarrow d\left(x\left(t\right), \mathcal{L}\right) \to 0$$

as $t \to \infty$.

The next theorem is the main result of this section. It states $\mathcal{E}_{\mathcal{O}}$ is compact and asymptotically stable.

Theorem 13. *$\mathcal{E}_{\mathcal{O}}$ is a compact set and it is asymptotically stable.*

Proof. We need to first prove that $\mathcal{E}_{\mathcal{O}}$ is a compact set. Since from our Theorem 11, we have $\mathcal{E}_{\mathcal{O}} = \mathcal{P}^*_{\mathcal{X}}$ and can instead prove

$$\mathcal{P}^*_{\mathcal{X}} := \left\{ x \in \mathcal{P}_{\mathcal{X}} | \sum_{i^* \in \mathcal{X}^*} x_{i^*} = 1 \right\}$$

is compact. It is easy to see that $\mathcal{P}^*_{\mathcal{X}}$ is tight (see Appendix, Definition 31) due to Assumption 6. Furthermore, we can prove it is a closed set by contradiction. Assume there exists a sequence $\{x^n\} \in \mathcal{P}^*_{\mathcal{X}}$ such that $x^n \to \hat{x} \notin \mathcal{P}^*_{\mathcal{X}}$. This implies $\exists N$ such that $\forall n > N$ we have $\sum_{i^* \in \mathcal{X}^*} x^n_{i^*} < 1$, and $\sum_{i \notin \mathcal{X}^*} x^n_i > 0$, which contradicts the second claim of Theorem 10. Hence, $\mathcal{P}^*_{\mathcal{X}} = \mathcal{E}_{\mathcal{O}}$ is a compact set.

Consider the Lyapunov function

$$V(x_t) = \mathbb{E}\left[\phi\left(H\left(X^*\right)\right)\right] - \mathbb{E}_t\left[\phi\left(H\left(X\right)\right)\right],$$

where $x^* \in \mathcal{P}^*_{\mathcal{O}}$ and X^* is the corresponding random variable. Note that $V(x_t)$ is positive for all $x_t \in \mathcal{P}_{\mathcal{X}} \backslash \mathcal{P}^*_{\mathcal{X}}$, and $V(x_t) = 0$ for $x_t \in \mathcal{P}^*_{\mathcal{X}} = \mathcal{E}_{\mathcal{O}}$. From Theorem 12 we have $\dot{V}(x_t) < 0$ for all $t > 0$ and $x_t \notin \mathcal{P}^*_{\mathcal{O}}$. Furthermore, we know $\mathcal{E}_{\mathcal{O}}$ is a compact set. Applying a generalized version of Lyapunov's theorem (see [12, Chapter 5]), the desired conclusion is reached. □

Remark 15. Chapter V of [12] presented a generalized version of Lyapunov's theorem on a general metric space. In the proof of Theorem 13, we applied this generalized Lyapunov's theorem on the Banach Space of σ-finite signed measures over

$(\mathcal{X}, \mathscr{B}(\mathcal{X}))$, where we used the total variation distance

$$d\left(x^1, x^2\right) = \sup_{A \in \mathscr{B}(\mathcal{X})} \sum_{j \in A} \left| x_j^1(t) - x_j^2(t) \right|.$$

Conclusion 1. We have proven so far that if we start Eq. 3.4.4 in $\mathcal{O}$, then the possible limit points are elements of $\mathcal{E}_{\mathcal{O}} = \mathcal{P}_{\mathcal{X}}^*$. Furthermore, the set $\mathcal{E}_{\mathcal{O}}$ is asymptomatically stable.

The use of the Lyapunov function in proving the stability of the limit set can also be found in Wang [87], which applied the generalized version of Lyapunov's theorem to an infinite dimensional space. In the next section, we will apply a similar approach to prove the stability of the limit set when the solution space is a Polish space.

3.5 Polish Space

In this section, we try to find a solution for Eq. 3.2.2 given that $\mathcal{X}$ is a Polish space with the Prohorov topology (see Appendix C.1). Polish spaces include finite-dimensional real spaces (i.e., $\mathcal{R}^n$), which are important in many engineering applications. The Polish space of probability measures on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ is denoted by $\mathcal{P}_{\mathcal{X}}$, which also has the Prohorov topology.

We will alter our notations in this section from the discrete space case. The symbol x in this section is an element of $\mathcal{X}$ (i.e., $x \in \mathcal{X}$). We will use the notation $\mathbb{P}_{X,t}(\{x\})$ to represent the probability measure of x at time t. This is different

from the discrete space case, where $\mathbb{P}_t(i) = x_i(t)\ \forall i \in \mathcal{X}$[6]. We write X for the solution random variable (i.e., $X(x) = x$), and denote the outcome random variable by $Y = \phi(H(X))$. On top of the assumptions we made for the discrete space case, we will make the following assumption.

Assumption 7. *w is differentiable, and has a bounded first derivative which is denoted by w'.*

The initial probability space, $(\mathcal{X}, \mathscr{B}(\mathcal{X}), \mathbb{P}_{X,0})$, with an initial distribution $\mathbb{P}_{X,0}$ induces a probability measure $\mathbb{P}_{Y,0}$ on the measurable space $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$, where

$$\mathcal{Y} := \left\{ y \in \mathcal{R}^+ | y = \phi(H(x))\ \ \exists x \in \mathcal{X} \right\} \subset \mathcal{R}^+$$

$$\mathbb{P}_{Y,0}(B) = \mathbb{P}_{X,0}(\{x | \phi(H(x)) = y\ \exists y \in B\})\ \ \forall B \in \mathscr{B}(\mathcal{Y}), \tag{3.5.1}$$

and $\mathscr{B}(\mathcal{Y})$ denotes the Borel σ-algebra for $\mathcal{Y}$. The space of probability measures on $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$ is denoted by $\mathcal{P}_{\mathcal{Y}}$. From examples 11 and 12, the generalized (i.e., when $\mathcal{X}$ is a Polish space) outcome probability measure, on the measurable space $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$, satisfies:[7]

$$\dot{\mathbb{P}}_{Y,t}(B) = \int_B w'(\mathbb{P}_{Y,t}(\{Y \geq y\}))\, d\mathbb{P}_{Y,t} - \mathbb{P}_{Y,t}(B)\ \ \forall B \in \mathscr{B}(\mathcal{Y}). \tag{3.5.2}$$

The initial condition for $\mathbb{P}_{Y,t}$ is given by Eq. 3.5.1. We pay special attention to the

[6] In the discrete space case, we used $x_i(t)$ to denote the probability of obtaining the i-th solution at time t

[7] $\mathbb{P}_{Y,t}(\{y\})$ is equal to $y_i(t)$ in the discrete space case.

best outcome equation:

$$\dot{\mathbb{P}}_{Y,t}(y*) = w\left(\mathbb{P}_{Y,t}\left(\{Y \geq y*\}\right)\right) - \mathbb{P}_{Y,t}(y*),$$

where $y^* = \max_{x \in \mathcal{X}} \phi(H(x))$.

Next, we write down the the Polish space counterpart to Eq. 3.4.4

$$\mathbb{P}_{X,t}(A) = \int_{A \times \mathcal{Y}} d\mathbb{P}_{X|Y,t} d\mathbb{P}_{Y,t} \ \forall A \in \mathscr{B}(\mathcal{X}) \ \ \forall t \in \mathcal{R}^+ \setminus \{0\}, \tag{3.5.3}$$

where $\mathbb{P}_{X|Y,t}$, the probability of X conditioned on Y, is the generalized β function. In other words, given a fixed $\mathbb{P}_{X|Y,t}$, $\mathbb{P}_{X,t}$ is determined by $\mathbb{P}_{Y,t}$, which is a solution of Eq. 3.5.2. $\mathbb{P}_{Y,t}$ can be determined, without knowing $\mathbb{P}_{X|Y,t}$, from $\mathbb{P}_{X,t}$ at any $t \in \mathcal{R}^+$ by the following equation:

$$\mathbb{P}_{Y,t}(B) = \mathbb{P}_{X,t}\left(\{x | \phi(H(x)) = y \ \exists y \in B\}\right) \ \forall B \in \mathscr{B}(\mathcal{Y}). \tag{3.5.4}$$

Assumption 8. *$\mathbb{P}_{X|Y,t}$ is a given fixed conditional probability measure.*

We denote the set of optimal measures on $(\mathcal{X}, \mathscr{B}(\mathcal{X}))$ by

$$\mathcal{P}^*_{\mathcal{X}} := \{\mathbb{P} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}(\mathcal{X}^*) = 1\},$$

where $\mathcal{X}^*$ is the set of all optimal solutions, i.e.,

$$\mathcal{X}^* := \{x^* \in \mathcal{X} | H(x) \leq H(x^*) \ \forall x \in \mathcal{X}\}.$$

Similarly, we denote the set of optimal measures on $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$ by

$$\mathcal{P}_{\mathcal{Y}}^{*} := \{\mathbb{P}_Y \in \mathcal{P}_{\mathcal{Y}} | \mathbb{P}(y^*) = 1\},$$

where $y^* = \max_{x \in \mathcal{X}} \phi(H(x))$. Obtaining an element $\mathbb{P}$ of $\mathcal{P}_{\mathcal{X}}^{*}$ is equivalent to solving the optimization problem stated in Eq. 3.2.2.

We will now prove the existence and uniqueness of a solution for Eq. 3.5.2. This is important for building a solid theoretical foundation for our approach.

Theorem 14. *For each $\mathbb{P}_{Y,0} \in \mathcal{P}_{\mathcal{Y}}$, the ordinary differential equation (3.5.2) has a unique solution for $t \in \mathcal{R}^{+}$.*

Proof. The outline of our proof follows [65] and [51]. We use the total variation norm, $\|\cdot\|$, on a σ-finite signed measure space over $(\mathcal{Y}, \mathscr{B}(\mathcal{Y}))$ at the time t:

$$\|\mathbb{P}_t\| = \sup_{g} \left| \int_{\mathcal{Y}} g(y) d\mathbb{P}_t \right|,$$

where the sup is taken over all measurable functions $g : \mathcal{Y} \to \mathcal{R}$ and

$$\sup_{y \in \mathcal{Y}} |g(y)| \leq 1.$$

We simplify our notations by introducing the following definition:

$$\mathscr{C}(\mathbb{P}_{Y,t})(B) := \int_B w'(\mathbb{P}_{Y,t}(\{Y \geq y\})) d\mathbb{P}_{Y,t} - \mathbb{P}_{Y,t}(B),$$

,

where $B \in \mathscr{B}(\mathcal{Y})$. Since $\mathbb{P}_{Y,t} \in \mathcal{P}_{\mathcal{Y}}$ is a probability measure, we note that

$$\begin{aligned}
\|\mathscr{C}(\mathbb{P}_{Y,t})\| &= \sup_g \left| \int_{\mathcal{Y}} g(y)\, \mathscr{C}(\mathbb{P}_{Y,t}) \right| \\
&\leq \sup_g \left| \int_{\mathcal{Y}} g(y)\, w'(\mathbb{P}_{Y,t}(\{Y \geq y\}))\, d\mathbb{P}_{Y,t} \right| + \sup_g \left| \int_{\mathcal{Y}} g(y)\, d\mathbb{P}_{Y,t} \right| \\
&\leq \int_{\mathcal{Y}} w'(\mathbb{P}_{Y,t}(\{Y \geq y\}))\, d\mathbb{P}_{Y,t} + \int_{\mathcal{Y}} d\mathbb{P}_{Y,t} \\
&\leq K \int_{\mathcal{Y}} d\mathbb{P}_{Y,t} + \int_{\mathcal{Y}} d\mathbb{P}_{Y,t} \\
&\leq (K+1),
\end{aligned}$$

which proves the boundedness of $\mathscr{C}(\mathbb{P}_{Y,t})$, with K being the Lipschitz constant for w.

Next, we need to prove that the right hand side of Eq. 3.5.2 is Lipschitz continuous. We know that

$$\begin{aligned}
&\|\mathscr{C}(\mathbb{P}_{Y,t}) - \mathscr{C}(\mathbb{Q}_{Y,t})\| \\
&\leq \sup_g \left| \int_{\mathcal{Y}} g(y)\, d\left(\mathscr{C}(\mathbb{P}_{Y,t}) - \mathscr{C}(\mathbb{Q}_{Y,t})\right) \right|. \qquad (3.5.5)
\end{aligned}$$

Furthermore, we know that

$$
\begin{aligned}
&\left|\mathscr{C}\left(\mathbb{P}_{Y,t}\right)-\mathscr{C}\left(\mathbb{Q}_{Y,t}\right)\right|(B) \\
&=\left|\left(\int_{B} w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{P}_{Y,t}-\mathbb{P}_{Y,t}(B)\right)\right. \\
&\left.-\left(\int_{B} w'\left(\mathbb{Q}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{Q}_{Y,t}-\mathbb{Q}_{Y,t}(B)\right)\right| \\
&\leq\left|\int_{B} w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{P}_{Y,t}-\int_{B} w'\left(\mathbb{Q}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{Q}_{Y,t}\right| \\
&\quad+\left|\mathbb{P}_{Y,t}(B)-\mathbb{Q}_{Y,t}(B)\right| \\
&\leq K\left|\int_{B} d\left(\mathbb{P}_{Y,t}-\mathbb{Q}_{Y,t}\right)\right|+\left|\int_{B} d\left(\mathbb{P}_{Y,t}-\mathbb{Q}_{Y,t}\right)\right| \\
&\leq(K+1)\left|\int_{B} d\left(\mathbb{P}_{Y,t}-\mathbb{Q}_{Y,t}\right)\right| . \qquad (3.5.6)
\end{aligned}
$$

Substituting Eq. 3.5.6 into Eq. 3.5.5, we have

$$
\begin{aligned}
&\left\|\mathscr{C}\left(\mathbb{P}_{Y,t}\right)-\mathscr{C}\left(\mathbb{Q}_{Y,t}\right)\right\| \\
&\leq \sup_{g}(K+1)\left|\int_{\mathcal{Y}} g(y) d\left(\mathbb{P}_{Y,t}-\mathbb{Q}_{Y,t}\right)\right| \\
&=(K+1)\left\|\mathbb{P}_{Y,t}-\mathbb{Q}_{Y,t}\right\| .
\end{aligned}
$$

Hence, the right hand side of Eq. 3.5.2 is Lipschitz continuous in $\mathbb{P}_{Y,t}$ with the constant $K+1$, where K is the Lipschitz constant for w (w is assumed to be Lipschitz continuous). Using Corollary 3.9 of [90], we conclude that Eq. 3.5.2 with an initial measure $\mathbb{P}_{Y,0} \in \mathcal{P}_{\mathcal{Y}}$ has a unique solution $\mathbb{P}_{Y,t}$. □

It should not be surprising that $\mathbb{P}_{Y,t}$ is a probability measure for all t.

Lemma 7. *Given that $\mathbb{P}_{Y,0}$ is a probability measure, then a solution $\mathbb{P}_{Y,t}$ of Eq. 3.5.2 at each time $t > 0$ is a probability measure, i.e.,*

$$\mathbb{P}_{Y,t}(B) \geq 0 \ \forall B \in \mathscr{B}(\mathcal{Y})$$

$$\mathbb{P}_{Y,t}(\mathcal{Y}) = 1 \ \forall t \in \mathcal{R}^{+}$$

$$\mathbb{P}_{Y,t}(\cup B_i) = \sum \mathbb{P}_{Y,t}(B_i),$$

where $\{B_i\}$ is any countable collection of pairwise disjoint elements of $\mathscr{B}(\mathcal{Y})$.

Proof. If we can prove that

$$\dot{\mathbb{P}}_{Y,t}(\mathcal{Y}) = 0,$$

and

$$\dot{\mathbb{P}}_{Y,t}(\cup B_i) = \sum \dot{\mathbb{P}}_{Y,t}(B_i)$$

then we have obtained our desired result.

From Eq. 3.5.2 we know

$$\begin{aligned} \dot{\mathbb{P}}_{Y,t}(\mathcal{Y}) &= \int_{\mathcal{Y}} w'(\mathbb{P}_{Y,t}(\{Y \geq y\}))\, d\mathbb{P}_{Y,t} - \mathbb{P}_{Y,t}(\mathcal{Y}) \\ &= \int_{\mathcal{Y}} (w'(\mathbb{P}_{Y,t}(\{Y \geq y\})) - 1)\, d\mathbb{P}_{Y,t}. \end{aligned}$$

Furthermore, that fact that $\int_0^1 w'(s)\, ds = 1$ (i.e., we assumed $w(1) = 1$) implies

$$\int_{\mathcal{Y}} (w'(\mathbb{P}_{Y,t}(\{Y \geq y\})) - 1)\, d\mathbb{P}_{Y,t} = 0.$$

Next, we need to prove that

$$\dot{\mathbb{P}}_{Y,t}\left(\cup B_i\right) = \sum \dot{\mathbb{P}}_{Y,t}\left(B_i\right),$$

because

$$\begin{aligned}
\int \dot{\mathbb{P}}_{Y,t}\left(\cup B_i\right) dt &= \int \sum \dot{\mathbb{P}}_{Y,t}\left(B_i\right) dt \\
\int \dot{\mathbb{P}}_{Y,t}\left(\cup B_i\right) dt &= \sum \int \dot{\mathbb{P}}_{Y,t}\left(B_i\right) dt \\
\implies \mathbb{P}_{Y,t}\left(\cup B_i\right) &= \sum \mathbb{P}_{Y,t}\left(B_i\right).
\end{aligned}$$

Using the fact that w' is bounded, along with dominated convergence theorem we conclude that

$$\begin{aligned}
\dot{\mathbb{P}}_{Y,t}\left(\cup B_i\right) &= \int_{\cup B_i} w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{P}_{Y,t} - \mathbb{P}_{Y,t}\left(\cup B_i\right) \\
&= \sum \int_{B_i} w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) d\mathbb{P}_{Y,t} - \mathbb{P}_{Y,t}\left(B_i\right) \\
&= \sum \dot{\mathbb{P}}_{Y,t}\left(B_i\right).
\end{aligned}$$

□

The next Lemma is needed in Theorem 18, which shows the monotonically increasing nature of $\mathbb{E}_t\left[Y\right]$.

Lemma 8. *Given an optimal-seeking weighting function, w, there exists a $\tilde{y}$ such*

that

$$\int_{\mathcal{Y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{\mathcal{Y}} y d\mathbb{P}_{Y,t} \tag{3.5.7}$$

can be decomposed into the sum of its non-negative and negative parts:

$$\underbrace{\int_{Y \geq \tilde{y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{Y \geq \tilde{y}} y d\mathbb{P}_{Y,t}}_{non-negative}$$

$$\underbrace{\left(\int_{Y < \tilde{y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{Y < \tilde{y}} y d\mathbb{P}_{Y,t}\right)}_{negative}.$$

In other words, we can write Eq. 3.5.7 as the sum of its non-negative and negative parts.

Proof. Since w is a monotonically increasing function, it satisfies

$$w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) \geq 0 \ \forall y \in \mathcal{Y}.$$

Furthermore, since w is an optimal-seeking function we have

$$w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y_1\}\right)\right) > w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y_2\}\right)\right) \ \forall y_1 \geq y_2 \in \mathcal{Y}. \tag{3.5.8}$$

In addition, since $w\left(0\right) = 0$ and $w\left(1\right) = 1$, we know that

$$w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right) > 1$$

for some $y \in \mathcal{Y}$. From Eq. 3.5.8 we know if y_2 satisfies the above inequality, then

so does $y_1 \geq y_2 \in \mathcal{Y}$. Hence, we can conclude that $\tilde{y}$ is the smallest of such y. $\square$

At the beginning of this section, we stated implicitly that if we can find an elements of $\mathcal{P}^*_{\mathcal{X}}$, then we have found a solution to the global optimization problem stated in Eq. 3.2.2. The theorems below present a blueprint, through the use of Eq. 3.5.3, to obtain an element of $\mathcal{P}^*_{\mathcal{X}}$. In Theorem 14, an initial condition can be any element of $\mathcal{P}_{\mathcal{Y}}$, which implies there is no restriction on the initial condition $\mathbb{P}_{X,0} \in \mathcal{P}_{\mathcal{X}}$. As we will discover later, $\mathcal{P}_{\mathcal{X}}$ is too large of a set to start Eq. 3.5.3 to guarantee as $t \to \infty$ the solution probability measure, $\mathbb{P}_{X,t}$, will be an element of $\mathcal{P}^*_{\mathcal{X}}$. Hence, we need to constrain our potential initial points to a smaller set.

Definition 25. The set of all *optimal initial solution probability measures* is denoted by:

$$\mathcal{O}_X := \{\mathbb{P}_{X,0} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}_{X,0}(\mathcal{X}^*) > 0\}.$$

Furthermore, the set of all *optimal initial outcome probability measures* is denoted by:

$$\mathcal{O}_Y := \{\mathbb{P}_{Y,0} \in \mathcal{P}_{\mathcal{Y}} | \mathbb{P}_{Y,0}(B) = \mathbb{P}_{X,0}(\{x | \phi(H(x)) = y \ \exists y \in B\})$$
$$\exists \mathbb{P}_{X,0} \in \mathcal{O}_X, \ \forall B \in \mathscr{B}(\mathcal{Y})\}.$$

Proposition 4. $\mathcal{O}_Y = \{\mathbb{P}_{Y,0} \in \mathcal{P}_{\mathcal{Y}} | \mathbb{P}_{Y,0}(y*) > 0\}$ *and* $\mathcal{P}^*_{\mathcal{X}} \subset \mathcal{O}_X$.

Proof. The first claim is a direct result of the definition above and Eq. 3.5.1. The second claim holds because how $\mathcal{P}^*_{\mathcal{X}}$ is defined. $\square$

The next theorem proves the probability measure on the optimal outcome set

will converge to 1 as $t \to \infty$. On the other hand, the total probability measure on the non-optimal outcome set will converge to 0 as $t \to \infty$.

Theorem 15. *If $\mathbb{P}_{Y,t}$ is a solution of Eq. 3.5.2 with initial points in $\mathcal{O}_Y$ (i.e., $\mathbb{P}_{Y,0} \in \mathcal{O}_Y$), then the following statements hold:*

1) $\mathbb{P}_{Y,t}(\{y^*\})$ is a monotonically increasing function of t. In fact, it converges to 1 as $t \to \infty$;

2) $\mathbb{P}_{Y,t}(\mathcal{Y} \backslash \{y^*\})$ approaches zero as $t \to \infty$.

Proof. We first write down the equation for $\mathbb{P}_{Y,t}(*)$:

$$\dot{\mathbb{P}}_{Y,t}(*) = w\left(\mathbb{P}_{Y,t}(\{Y \geq *\})\right) - \mathbb{P}_{Y,t}(*).$$

From Proposition 2, we conclude that

$$w\left(\mathbb{P}_{Y,t}(\{Y \geq *\})\right) > \mathbb{P}_{Y,t}(*).$$

From the two equations above, we conclude that

$$\dot{\mathbb{P}}_{Y,t}(*) > 0 \ \forall \mathbb{P}_{Y,t}(*) \in (0,1), \text{ and } \dot{\mathbb{P}}_{Y,t}(*) = 0 \text{ when } \mathbb{P}_{Y,t}(*) = 1.$$

Since $\mathbb{P}_{Y,0} > 0$, the first claim is proved.

The second claim follows from the first claim. Since $\lim_{t\to\infty} \mathbb{P}_{Y,t}(\{\mathbf{y}^*\}) = 1$ we

can conclude the following:

$$\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\mathcal{Y}\right)=\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\{\mathbf{y}^{*}\}\right)+\mathbb{P}_{Y,t}\left(\mathcal{Y}\backslash\{\mathbf{y}^{*}\}^{*}\right)$$

$$=1+\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\mathcal{Y}\backslash\{\mathbf{y}^{*}\}\right)=1\implies\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\mathcal{Y}\backslash\{\mathbf{y}^{*}\}\right)=0.$$

The second claim is proved. □

The next theorem connects the properties of $\mathbb{P}_{Y,t}$ with those of $\mathbb{P}_{X,t}$ as $t\to\infty$. This is an important step for understanding the evolution of Eq. 3.5.3.

Theorem 16. *Assuming* $\mathbb{P}_{X,0}\in\mathcal{O}_{X}$ *and*

$$\mathbb{P}_{Y,0}\left(B\right)\;=\;\mathbb{P}_{X,0}\left(\{x|\phi\left(H\left(x\right)\right)=y\;\exists y\in B\}\right)\;\forall B\in\mathscr{B}\left(\mathcal{Y}\right),$$

the following statements hold:

1) $\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\{y^{*}\}\right)=\lim_{t\to\infty}\mathbb{P}_{X,t}\left(\mathcal{X}^{*}\right)$*=1;*

2) $\lim_{t\to\infty}\mathbb{P}_{Y,t}\left(\mathcal{Y}\backslash\left\{y^{*}\right\}\right)=\lim_{t\to\infty}\mathbb{P}\left(\mathcal{X}\backslash\mathcal{X}^{*}\right)$*=0.*

Proof. Since we know

$$\mathbb{P}_{X,t}\left(A\right)=\int_{A\times\mathcal{Y}}d\mathbb{P}_{X|Y,t}d\mathbb{P}_{Y,t}\;\forall A\in\mathscr{B}\left(\mathcal{X}\right)\;\forall t\in\mathcal{R}^{+}\backslash\left\{0\right\},$$

we prove the first claim by writing down the following equation:

$$\begin{aligned}\lim_{t\to\infty}\mathbb{P}_{X,t}\left(\mathcal{X}^*\right) &= \lim_{t\to\infty}\int_{\mathcal{X}^*\times\mathcal{Y}} d\mathbb{P}_{X|Y,t}d\mathbb{P}_{Y,t} \\ &= \int_{\mathcal{X}^*\times\{y*\}} d\mathbb{P}_{X|Y,t}\lim_{t\to\infty} d\mathbb{P}_{Y,t} \\ &= 1.\end{aligned}$$

For claim 2, we can write down a similar equation:

$$\begin{aligned}\lim_{t\to\infty}\mathbb{P}_{X,t}\left(\mathcal{X}\backslash\mathcal{X}^*\right) &= \lim_{t\to\infty}\int_{\mathcal{X}\backslash\mathcal{X}^*\times\mathcal{Y}} d\mathbb{P}_{X|Y,t}d\mathbb{P}_{Y,t} \\ &= \int_{\mathcal{X}\backslash\mathcal{X}^*\times\{y^*\}} d\mathbb{P}_{X|Y,t}\lim_{t\to\infty} d\mathbb{P}_{Y,t} \\ &= 0.\end{aligned}$$

□

We are interested in finding the limit points of Eq. 3.5.2. Ideally, these limit points should be elements in $\mathcal{P}^*_{\mathcal{X}}$. In order to guarantee this, we need to restrict the potential initial points to $\mathcal{O}_X$.

Definition 26. We define a *limit set* starting from an element $\mathbb{P}_0 \in \mathcal{I}$ as

$$\mathcal{E}_{\mathcal{I}} := \left\{\mathbb{P}_{X,\infty} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}_{X,\infty}\left(A\right) = \lim_{t\to\infty}\mathbb{P}_{X,t}\left(A\right)\ \ \exists\mathbb{P}_{X,0} \in \mathcal{I},\ \forall A \in \mathscr{B}\left(\mathcal{X}\right)\right\}$$

We are particularly interested in the limit set $\mathcal{E}_{\mathcal{O}_X}$. The next theorem describes the elements in this set.

Theorem 17. *The limit set of Eq. 3.5.3 started in $\mathcal{O}_X$ is $\mathcal{P}^*_{\mathcal{X}}$, i.e.,*

$$\mathcal{E}_{\mathcal{O}_X} \;=\; \mathcal{P}^*_{\mathcal{X}} := \{\mathbb{P} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}(\mathcal{X}^*) = 1\}.$$

Proof. We will first prove $\mathcal{E}_{\mathcal{O}_X} \supset \mathcal{P}^*_{\mathcal{X}}$, then we will prove $\mathcal{E}_{\mathcal{O}_X} \subset \mathcal{P}^*_{\mathcal{X}}$. In the first case, $\mathcal{E}_{\mathcal{O}_X} \supset \mathcal{P}^*_{\mathcal{X}}$, can be proved by taking an element $\mathbb{P}_X \in \mathcal{P}^*_{\mathcal{X}} \subset \mathcal{O}_X$. By the definition of $\mathcal{E}_{\mathcal{O}_X}$ (i.e., the limit set of Eq. 3.5.3 starting from $\mathcal{O}_X$), equations 3.5.3 and 3.5.2, we conclude that $\mathbb{P}_X \in \mathcal{E}_{\mathcal{O}_X}$.

The second claim, $\mathcal{E}_{\mathcal{O}_X} \subset \mathcal{P}^*_{\mathcal{X}}$, can be proven by contradiction. We assume there is an element $\mathbb{Q}_X \in \mathcal{E}_{\mathcal{O}_X}$ but not in $\mathcal{P}^*_{\mathcal{X}}$, which implies

$$\mathbb{Q}_X(A) = \lim_{t\to\infty} \mathbb{Q}_{X,t}(A) = \int_{A\times\mathcal{Y}} d\mathbb{P}_{X|Y,t} \lim_{t\to\infty} d\mathbb{P}_{Y,t} \;\forall A \in \mathscr{B}(\mathcal{X})$$
$$s.t\; \mathbb{Q}_{X,0}(\mathcal{X}^*) > 0,\; \mathbb{Q}_X(\mathcal{X}\backslash\mathcal{X}^*) > 0.$$

The first line is due to the fact that $\mathbb{Q}_X \in \mathcal{E}_{\mathcal{O}_X}$. The second line in the equation above, along with Eq. 3.5.4, imply $\lim_{t\to\infty} \mathbb{P}_{Y,t}(\mathcal{Y}\backslash\{y^*\}) > 0$, which contradicts the second claim of Theorem 15, where $\lim_{t\to\infty} \mathbb{P}_{Y,t}(\mathcal{Y}\backslash\{y^*\}) = 0$. □

The following theorem shows the monotonically increasing nature of $\mathbb{E}_t[Y]$, which will be useful later in proving some stability properties for Eq. 3.5.3.

Theorem 18. *Let $\mathbb{P}_{Y,t}$ be a solution for Eq. 3.5.2 with its initial point in $\mathcal{O}_Y$. Then the following statements are true:*

1) The expected outcome, i.e., $\mathbb{E}_t[Y]$ is monotonically increasing with t;

2) If $\mathbb{P}_{Y,\tilde{t}} \notin \mathcal{E}_{\mathcal{O}_Y}$ for any $\tilde{t} \in \mathcal{R}^+$, then $\mathbb{E}_{\tilde{t}}[Y]$ is strictly increasing with $\tilde{t}$.

Proof. We start our proof by differentiating the average outcome function:

$$\begin{aligned}
\frac{d}{dt}\mathbb{E}_t[Y] &= \int_{\mathcal{Y}} y\dot{\mathbb{P}}_{Y,t}(dy) \\
&= \int_{\mathcal{Y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{\mathcal{Y}} y d\mathbb{P}_{Y,t} \\
&= \int_{Y \geq \tilde{y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{Y \geq \tilde{y}} y d\mathbb{P}_{Y,t} \\
&+ \left(\int_{Y < \tilde{y}} y\left(w'\left(\mathbb{P}_{Y,t}\left(\{Y \geq y\}\right)\right)\right) d\mathbb{P}_{Y,t} - \int_{Y < \tilde{y}} y d\mathbb{P}_{Y,t}\right) \\
&\geq \tilde{y}\int_{\mathcal{Y}} d\dot{\mathbb{P}}_{Y,t} = \tilde{y} \times 0 = 0.
\end{aligned}$$

The $\tilde{y}$ variable is used to decompose the expected outcome function into non-negative and negative parts (see Lemma 8). The last line of the inequality is true because $\mathbb{P}_{Y,t} \in \mathcal{P}_{\mathcal{Y}}\ \forall t \in \mathcal{R}^+$ (see Lemma 7). The first claim is proved.

The second claim is proved by contradiction. If there exists a $\mathbb{P}_{Y,\tilde{t}}$, not a limit point, with

$$\frac{d}{dt}\mathbb{E}_{\tilde{t}}[Y] = 0.$$

Along with Theorem 15, the equality above implies that Y is equal to a constant $C = \underset{x \in \mathcal{X}}{sup}\ \phi(H(x))$. This implies $\mathbb{P}_{Y,\tilde{t}}$ is a Dirac measure concentrated at C, which is a limit point (see Theorem 17). □

The metric function d in the following definitions is defined in Appendix C (see Definition 32).

Definition 27. Let $\mathcal{L}$ be a subset of $\mathcal{P}_{\mathcal{X}}$. For a point $\mathbb{P} \in \mathcal{P}_{\mathcal{X}}$, we define the distance

between $\mathbb{P}$ and $\mathcal{L}$ as

$$d(\mathbb{P}, \mathcal{L}) := \inf \{ d(\mathbb{P}, \mathbb{Q}), \ \forall \mathbb{Q} \in \mathcal{L} \}.$$

$\mathcal{L}$ is called *Lyapunov stable,* with respect to a sequence of measures $\{\mathbb{P}_t\}$, if for all $\epsilon > 0$, there exists a $\delta > 0$ such that

$$d(\mathbb{P}_t, \mathcal{L}) < \delta \Rightarrow d(\mathbb{P}_t, \mathcal{L}) < \epsilon, \ \forall t > 0.$$

$\mathcal{L}$ is called *asymptotically stable*, with respect to a sequence of measures $\{\mathbb{P}_t\}$, if $\mathcal{L}$ is Lyapunov stable, and there exists a $\delta > 0$ such that

$$d(\mathbb{P}_0, \mathcal{L}) < \delta \Rightarrow d(\mathbb{P}_t, \mathcal{L}) \to 0$$

as $t \to \infty$.

The next theorem is the main result of this section. It tells us that if we start Eq. 3.5.3 in the set $\mathcal{O}_X$, then $\mathcal{E}_{\mathcal{O}_X}$ will coincide with $\mathcal{P}^*_{\mathcal{X}}$. Furthermore, $\mathcal{E}_{\mathcal{O}_X}$ is asymptotically stable.

Theorem 19. *$\mathcal{E}_{\mathcal{O}_X}$ is a compact set and it is asymptotically stable.*

Proof. We need to first prove that $\mathcal{E}_{\mathcal{O}_X}$ is a compact set. Since from Theorem 17, we have $\mathcal{E}_{\mathcal{O}_X} = \mathcal{P}^*_{\mathcal{X}}$, we can instead prove

$$\mathcal{P}^*_{\mathcal{X}} := \{ \mathbb{P} \in \mathcal{P}_{\mathcal{X}} | \mathbb{P}(\mathcal{X}^*) = 1 \}$$

is compact. It is easy to see that $\mathcal{P}^*_{\mathcal{X}}$ is tight (see Appendix, Definition 31) due to Assumption 6. Furthermore, we can prove it is a closed set by contradiction. Assume there exists a sequence $\{\mathbb{P}^n\} \in \mathcal{P}^*_{\mathcal{X}}$ such that $\mathbb{P}^n \to \hat{\mathbb{P}} \notin \mathcal{P}^*_{\mathcal{X}}$. This implies $\exists N$ such that $\forall n > N$ we have $\mathbb{P}^n(\mathcal{X}^*) < 1$, and $\mathbb{P}^n(\mathcal{X} \backslash \mathcal{X}^*) > 0$, which implies

$$\lim_{n\to\infty} \mathbb{P}^n_Y(\mathcal{Y} \backslash \{y^*\}) > 0.$$

This contradicts the second claim of Theorem 15. Hence, $\mathcal{P}^*_{\mathcal{X}} = \mathcal{E}_{\mathcal{O}}$ is a compact set.

Using the Lyapunov function

$$V(\mathbb{P}_{X,t}) = y * -\mathbb{E}_t[\phi(H(X))] = V(\mathbb{P}_{Y,t}) = y * -\mathbb{E}_t[Y],$$

note that $V(\mathbb{P}_{X,t}) > 0$ for all $\mathbb{P}_{X,t} \in \mathcal{P}_{\mathcal{X}} \backslash \mathcal{P}^*_{\mathcal{X}}$, and $V(\mathbb{P}_{X,t}) = 0$ for $\mathbb{P}_{X,t} \in \mathcal{P}^*_{\mathcal{X}} = \mathcal{E}_{\mathcal{O}_X}$. From Theorem 18 we have $\dot{V}(\mathbb{P}_{X,t}) = \dot{V}(\mathbb{P}_{Y,t}) < 0$ for all $t > 0$ if $\mathbb{P}_{X,t} \notin \mathcal{P}^*_{\mathcal{X}}$. Furthermore, we know $\dot{V}(\mathbb{P}_{X,t}) = \dot{V}(\mathbb{P}_{Y,t}) = 0\ \forall t > 0$ if $\mathbb{P}_{X,t} \in \mathcal{P}^*_{\mathcal{X}}$. Using $V(\mathbb{P}_{X,t})$ as the Lyapunov function, and the fact that $\mathcal{E}_{\mathcal{O}_X}$ is a compact set, we can appeal to a generalized version of Lyapunov's theorem (see [12, Chapter 5]). The desired conclusion is reached. □

The use of the Lyapunov function for proving the asymptotic stability of the limit set can be found previously in Wang's dissertation [87].

3.6 Numerical Algorithms

In this section, we present a few numerical algorithms based on the CWO (Cumulative Weighting Optimization) method we presented in the previous sections. These algorithms attempt to find an optimal solution iteratively. Each iteration consists of 5 stages: generation, quantile-update, parameter-update, weight-update, and projection. Generation, quantile-update and projection stages remain the same for all variations of the generic algorithm (i.e., Algorithm 1). We propose several approaches for constructing the weight-update stage. These algorithms build on the theoretical results using the same types of modifications as are found in the CE and MRAS (see [75, 53, 88]).

Algorithm 1 Generic CWO Algorithm

1. Initialization: Select a number N_0 as the total initial number of candidate solutions generated at each iteration and an initial g_{θ_0} (a parameterized probability density distribution) defined on $\mathbb{X}$. Pick an initial quantile $\rho_0 \in (0,1)$, $\epsilon \geq 0$, $\alpha > 0$, $\lambda \in (0,1)$.

2. Generation: Generate N_k i.i.d candidate solutions $\{x_k^i\}_{i=1}^N$ from

$$\tilde{g}_{\theta_k} = (1-\lambda)((1-\beta)g_{\theta_{k-1}} + \beta g_{\theta_k}) + \lambda U,$$

where U is the uniform distribution.

3. Quantile-Update:
Calculate the $(1-\rho_k)$-quantile, $\tilde{\gamma}_{k+1}(\rho_k, N_k) := \phi(H)_{(\lceil (1-\rho_k)N_k \rceil)}$, where $\lceil a \rceil$ is the smallest integer greater than a and $H_{(i)}$ is the i-th highest value for the sequence $\{\phi(H(x_k^i))\}_{i=1}^{N_k}$.

4. Parameter-Update:
If k=0 **or** $\tilde{\gamma}_{k+1}(\rho_k, N_k) \geq \bar{\gamma}_k + \frac{\epsilon}{2}$, **then**
$\rightarrow$ 4(a). Set $\bar{\gamma}_{k+1} = \tilde{\gamma}_{k+1}(\rho_k, N_k)$, $\rho_{k+1} = \rho_k$, $N_{k+1} = N_k$;
else
$\rightarrow$ Find the largest $\bar{\rho} \in (0, \rho_k)$ such that $\tilde{\gamma}_{k+1}(\bar{\rho}, N_k) \geq \bar{\gamma}_k + \frac{\epsilon}{2}$;
$\rightarrow$ **If** such a $\bar{\rho}$ exists **and** $\bar{\rho} > \rho_{min}$, **then**
$\rightarrow\rightarrow$ 4(b). $\bar{\gamma}_{k+1} = \tilde{\gamma}_{k+1}(\bar{\rho}, N_k)$, $\rho_{k+1} = \bar{\rho}$, $N_{k+1} = N_k$;
$\rightarrow$ **else**
$\rightarrow\rightarrow$ 4(c). $\bar{\gamma}_{k+1} = \bar{\gamma}_k$, $\rho_{k+1} = \rho_k$, $N_{k+1} = \lceil \alpha N_k \rceil$;

5. Weight-Update: Update the weights of the generated samples $\{x_k^i\}_{i=1}^N$ according to weight update methods based on Eq. 3.4.4, producing the p.m.f $p_{X,k+1} = \sum_{i=1}^N w_{k+1}^i \delta(x - x_k^i)$. w_{k+1}^i is the updated weight for x_k^i.

6. Density Projection: Construct $g_{\theta_{k+1}}$ by projecting the density $p_{X,k+1} = \sum_{i=1}^N w_{k+1}^i \delta(x - x_k^i)$ onto g_θ by solving

$$\theta_{k+1} = \arg\max_{\theta \in \Theta} \sum_{i=1}^N w_{k+1}^i \ln g_\theta\left(x_k^i\right);$$

7. Stop if some stopping criterion is satisfied; otherwise go to step 2 and $k = k+1$.

Since discrete-time, discrete-state equations are more suitable for the computations in the weight-update stage, we write down the probability density counterparts

of Eq. 3.5.2 and 3.5.3:

$$p_{Y,k+1}(y) = w\left(\sum_{s:\phi(H(s))\geq y} p_{X,k}(s)\right) - w\left(\sum_{s:\phi(H(s))>y} p_{X,k}(s)\right)$$
$$p_{X,k+1}(x) = \frac{p_{Y,k+1}(H(x))\, p_{X,k}(x)}{\sum_{s:\phi(H(s))=y} p_{X,k}(s)}, \tag{3.6.1}$$

where $w : [0,1] \to [0,1]$ is the probability weighting function. If we assume $w(\cdot)$ is differentiable, then Eq. 3.6.1 can be written more compactly as:

$$\begin{aligned}
p_{X,k+1}(x) &= w'(1 - F_{Y,k}(\phi(H(x)))) \, p_{Y,k}(\phi(H(x))) \, p_{X|Y,k}(x, H(\phi(H(x)))) \\
&= w'(1 - F_{Y,k}(\phi(H(x)))) \, p_{Y,k}(\phi(H(x))) \, \frac{p_{X,k}(x)}{\sum_{\{s:\phi(H(s))=\phi(H(x))\}} p_{X,k}(s)} \\
&= \left(\frac{w\left(\sum_{s:\phi(H(s))\geq\phi(H(x))} p_{X,k}(s)\right) - w\left(\sum_{s:\phi(H(s))>\phi(H(x))} p_{X,k}(s)\right)}{\sum_{\{s:\phi(H(s))=\phi(H(x))\}} p_{X,k}(s)} \right) p_{X,k}(x) \\
&= w'(1 - F_{Y,k}(\phi(H(x)))) \, p_{X,k}(x), \qquad (3.6.2)
\end{aligned}$$

where the second equality holds because the conditional density is taken to be

$$p_{X|Y,t}(x, H(\phi(H(x)))) = \frac{p_{X,t}(x)}{\sum_{\{s:\phi(H(s))=\phi(H(x))\}} p_{X,t}(s)}$$

and $F_{Y,k}(\cdot)$ is the cumulative distribution function for the outcome values. Other choices for $p_{X|Y,t}$ are also allowed; in particular, a uniform conditional density.

The second to last equality in Eq. 3.6.2 can be related to Eq. 3.6.1 by the

following equation:

$$p_{Y,k+1}(y) = w\left(\sum_{s:\phi(H(s))\geq y} p_{X,k}(s)\right) - w\left(\sum_{s:\phi(H(s))>y} p_{X,k}(s)\right)$$
$$= w'\left(1 - F_{Y,k}\left(\phi\left(H\left(x\right)\right)\right)\right) p_{Y,k}\left(\phi\left(H\left(x\right)\right)\right).$$

Algorithm 1 is the generic CWO algorithm using Eq. 3.6.1 as the weight update equation. Later, two algorithms with different ways of updating the density function are described. Although both weight-update methods will use Eq. 3.6.1 with a chosen distribution rule, they differ in their assignment of the sample weights.

The performance of the algorithms is measured using asymmetric traveling salesman problems, which we will introduce in the section below.

3.6.1 Numerical Examples: Asymmetric Traveling Salesman Problems (ATSPs)

We apply variations of Algorithm 1 to several asymmetric traveling salesman problems (ATSPs). They are taken from the website `http://www.iwr.uni-heidelberg.de/groups/comopt/software/TSPLIB95`. We follow a similar approach as in Hu [53], which is outlined below. The reader is reminded here that Algorithm 1 is designed for maximization problems, whereas we are searching for the minimum distances of ATSPs. The goal in each ATSP problem is to find the minimum length of a tour that connects N_{cities} cities with the same starting and ending cities. For an ATSP, we are given an N_{cities}-by-N_{cities} distance matrix D, whose (i,j)-th element

$D_{i,j}$ represents the distance from city i to city j. The problem can be mathematically stated as:

$$\min_{x \in \mathfrak{X}} \left\{ \sum_{i=1}^{N_{cities}-1} D_{x_i, x_{i+1}} + D_{x_{N_{cities}}, x_1} \right\},$$

where $x := \left\{ x_1, x_2, \ldots, x_{N_{cities}}, x_1 \right\}$ is an admissible tour and $\mathfrak{X}$ is the set of all admissible tours.

We use the same approach suggested by Rubinstein [75], and De Boer et al. [26] for solving these problems. Each distance matrix D is given an initial state probability transition matrix $\tilde{P}_0$, whose (i,j)-th element specifies the probability of transitioning from city i to city j. At each iteration of the algorithm, there are two important steps: 1) generate random admissible tours according to the probability transition matrix and evaluate the performance of each sampled tour; 2) update the probability transition matrix based on the tours generated from step 1. We denote the set of tours generated at the k-th iteration by $\{x_k^i\}$, where $i \in \{1, \ldots, N_k\}$. Without loss of generality, we will assume the samples are sorted according to their values (i.e., $\phi\left(H\left(x_k^i\right)\right) < \phi\left(H\left(x_k^j\right)\right)$ if and only if $i < j$).

A detailed discussion of the admissible tour generation process can be found in De Boer et al. [26]. The CWO algorithm differs from other algorithms in how it updates its transition matrix. At the k-th iteration of CWO, the probability density function, $p_k(\cdot, \theta_k)$, parametrized by the transition matrix θ_k is given by the equation below:

$$p_k(x, \theta_k) = \prod_{l=i}^{N_{cities}} \sum_{i,j}^{N_{cities}} \theta_k(i,j)\, I_{\{x \in \mathfrak{X}_{i,j}(l)\}},$$

where $\mathfrak{X}_{i,j}(l)$ is the set of all tours in $\mathfrak{X}$ such that the l-th transition is from city i

to city j. We can show that the new transition matrix is updated (i.e., stage 6 in Algorithm 1) as

$$\theta_{k+1}(i,j) = \sum_{l=1}^{N_k} \left(p_{k+1}^{w}\left(x_k^i\right)\right) I_{\{x_k^i \in \mathfrak{x}_{i,j}\}},$$

where we denote the updated density by $p_{k+1}^{w}(\cdot)$ and $\{x_{k+1}^i\}$ is generated from $p_k(\cdot, \theta_k)$ (i.e., a density function that is parameterized by θ_k). The superscript w is used to emphasize the dependence of the updated probability mass function on the probability weighting function w. The construction of $p_{k+1}^{w}(\cdot)$ depends on the specific weight-update method.

3.6.1.1 Weight-Update Methods

In this section, we present several different methods of obtaining $p_{k+1}^{w}(\cdot)$ from a collection of samples $\{x_k^i\}$ at the k-th step. The first method we introduce is called tilted weight update.

Tilted weight update (CWO_T): The tilted weight-update method is described in Algorithm 2. The key idea behind this variation is that we assign the initial weights of the samples according to their outcome values: the smaller the value, the more initial weight it gets (see stage 2 in Algorithm 2).

Algorithm 2 Tilted Weight Update

1. Remove all the non-elite samples, i.e., $\{\hat{x}_k^i\} := \{x_k^i : H(x_k^i) \leq \bar{\gamma}_k\}$, where $\{\hat{x}_k^i\}$ is the set of remaining elite samples;

2. Assign a weight to each element in Y according to the equation:

$$p_{Y,k}(y) = \frac{\max_Y y - y}{\sum_Y \max_Y y - y},$$

where $Y := \{H(x) | x \in \{\hat{x}_k^i\}\}$;

3. Assign the updated outcome weights to samples according the following equation:

$$w_{k+1}^i = \frac{1}{\hat{N}_{k,\hat{x}_k^i}} \left(w \left(\sum_{y:y \leq \phi(H(\hat{x}_k^i))} p_{Y,k}(y) \right) - w \left(\sum_{y:y < \phi(H(\hat{x}_k^i))} p_{Y,k}(y) \right) \right)$$
$$\forall i \in \{1, \ldots, \hat{N}_k\},$$

where $\hat{N}_k$ is the number of elite samples, and $\hat{N}_{k,\hat{x}_k^i}$ is the number of elements in $\{\hat{x}_k^i\}$ having the same outcome value as $\hat{x}_k^i$. (We remind the reader that $w : [0,1] \to [0,1]$ is a probability weighting function.)

We ran 30 independent experiments for seven ATSPs. In those experiments, we used the probability weighting function:

$$w(p) := 1 - (1-p)^2.$$

The trials are done using Algorithm 1 with the parameters $\rho_0 = \rho_{min} = 0.6$, $N_0 = 1000$, $\epsilon = 1$, $\alpha = 2$, $\lambda = 0.02$, $\beta = 0.7$ and the weight-update scheme in Algorithm 2. The results are summarized in Table 3.6.1. N_{cities} is the the number of cities for each problem; N_{Total} is the average number of total samples until the solutions stop changing; H_{best} is the best known solution; H_* is the worst algorithm solution from the repeated runs; H^* is the best algorithm solution from the repeated runs;

δ_* and δ^* are the percentage deviation of the worst and best algorithm solutions from the best known solution, respectively; δ is the average percentage deviation of the algorithm solutions from the best known solution. The important thing to note about the algorithm is its dependence on the actual outcome values. In the next weight-update method, this dependence is eliminated; instead we weight the samples uniformly.

ATSP	N_{cities}	$N_{Total}\,(Std.\ err.)$	H_{best}	H_*	H^*	δ_*	δ^*	$\delta\,(Std.\ err.)$
ftv33	34	$6.59e4\,(1.81e4)$	1,286	1,379	1,286	0.0723	0.0000	0.0396(0.0279)
ftv35	36	$6.79e4\,(1.63e4)$	1,473	1581	1473	0.0733	0.0000	0.0195(0.0172)
ftv38	39	$8.81e4\,(3.26e4)$	1,530	1651	1536	0.0791	0.0039	0.0243(0.0190)
p43	43	$2.80e5\,(1.04e5)$	5,620	5,636	5,622	0.0028	0.0004	0.0011(0.0007)
ry48p	48	$4.65e5\,(2.30e5)$	14,422	18,725	14,618	0.2984	0.0136	0.0744(0.0676)
ft53	53	$3.24e5\,(1.23e5)$	6,905	7844	7059	0.1360	0.0223	0.0590(0.0247)
ft70	70	$7.02e5\,(3.32e5)$	38,673	39,738	38,760	0.0275	0.00225	0.0130(0.0050)

Table 3.6.1: Performance of CWO_T on various ATSP problems based on 30 independent replications

Uniform Weight Update(CWO_U): Tilting assigns the initial weights of the samples $\{x_k^i\}$ using their values. Uniform weighting updating differs from tilting by assuming uniform distribution over the samples. Another major difference from the above approach is that we no longer only consider elites samples. Instead, we use a carefully chosen probability weighting function that smoothly re-weights the samples. More specifically in stage 5, we assume a uniform initial density and use the weighting function

$$w\,(p) := \frac{10p\sigma + \ln\,(1+e^{-\sigma}) - \ln\left(1+e^{(-1+10p)\sigma}\right)}{10\theta + \ln\,(1+e^{-\rho}) - ln\,(1+e^{9\sigma})}, \tag{3.6.3}$$

where σ is the optimal-seeking factor and ρ is the quantile threshold. Eq. 3.6.3 is chosen as the weighting function due to its connection with the cross-entropy algorithm.

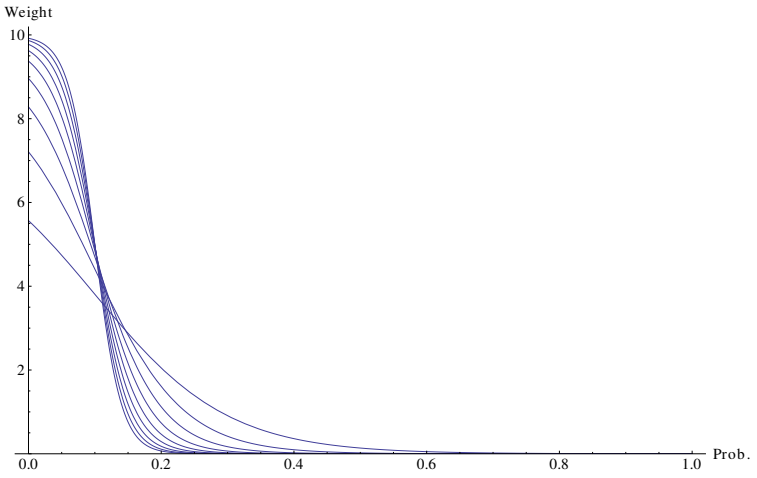


Figure 3.6.1: Derivatives of Eq. 3.6.3 as $\sigma \to \infty$

Using equations 3.6.2 and 3.6.3, we modify the generic CWO algorithm by altering the way the sample weights are updated. The algorithm has a strong connection with the traditional cross-entropy method, which is explained below.

Algorithm 3 CWO_U Weight Update Algorithm

---

1. Calculate the outcome cumulative distribution function (CDF), $F_{Y,k}\left(\phi\left(H\left(x\right)\right)\right)$, for $\{x_k^i\}$, assuming a uniform density $\left(i.e.,\ p_{X,k}\left(x\right) = \frac{1}{N_k}\right)$;

2. Assign the updated weights according the following equation:

$$w_{k+1}^i = \frac{1}{N_{k,x_k^i}} \left(w\left(\sum_{s:\phi(H(s))\leq\phi\left(H\left(x_k^i\right)\right)} p_{X,k}\left(s\right) \right) - w\left(\sum_{s:\phi(H(s))<\phi\left(H\left(x_k^i\right)\right)} p_{X,k}\left(s\right) \right) \right)$$
$$\forall i \in \{1, \ldots, N_k\},$$

where N_k is the number of samples, and N_{k,x_k^i}, is the number of elements in $\{x_k^i\}$ having the same value as x_k^i. (We remind the reader that $w : [0,1] \to [0,1]$ is a probability weighting function.)

We remind the reader that the density update equation for cross entropy is

$$\begin{aligned} p_{X,k+1}^{CE}(x) &= \frac{1\{\phi(H(s)) > \gamma\}}{l} p_{X,k}^{CE}(x) \\ &\propto 1\{\phi(H(s)) > \gamma\} p_{X,k}^{CE}(x), \end{aligned} \tag{3.6.4}$$

where an indicator function is used to select the elite samples. In fact, the cross-entropy equation is just the limiting case[8] of the CWO_U algorithm. As we increase the optimal-seeking factor, the derivative of Eq. 3.6.3 will approach a step function (i.e., Eq. 3.6.4) with its discontinuity occurring at $\rho = 0.1$ (see Fig. 3.6.1).

Table 3.6.2 contains the results from running 20 trials of CWO_U and CE algorithms with the parameters $\Delta = 0.01$, $\rho_0 = 0.1$, $\rho_{min} = 0.001$, $N_0 = 1000$, $\epsilon = 0$, $\alpha = 1$, $\lambda = 0.01$, and $\beta = 0.7$. Here, N_0 is the initial sample size.

ATSP	N_{cities}	N_{Total} (*Std. err.*)	H_{best}	H_*	H^*	δ_*	δ^*	δ (*Std. err.*)
ft53	53	90,450(6.0e3)	6,905	7,679	7,037	0.112	0.0191	0.060(0.0244)
ce_ft53	53	65,100(5.7e3)	6,905	7,676	7,088	0.111	0.0265	0.075(0.0276)

Table 3.6.2: CWO_U and CE performance Results

We plot the sorted minimum tour distances obtained from the 20 trials of CE and CWO_U algorithms in Fig. 3.6.2. We observe from Fig. 3.6.2 that compared with the standard cross-entropy method, our approach does better in every percentile. For example, the $\frac{19}{20}$th percentile would contain the lowest optimal solution obtained among the 20 trials. The $\frac{18}{20}$th percentile would contain the second lowest optimal solution obtained among the 20 trials.

[8] as $\sigma \to \infty$

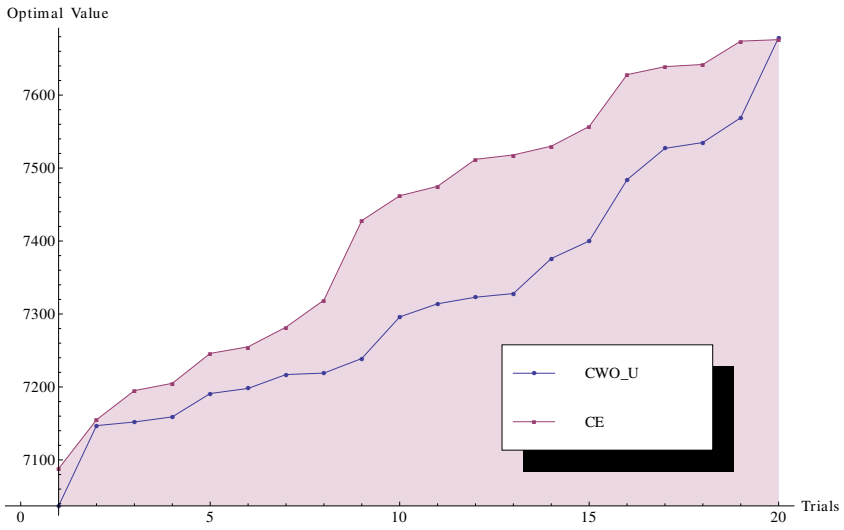


Figure 3.6.2: CE vs. CWO_U Sorted Trial Runs

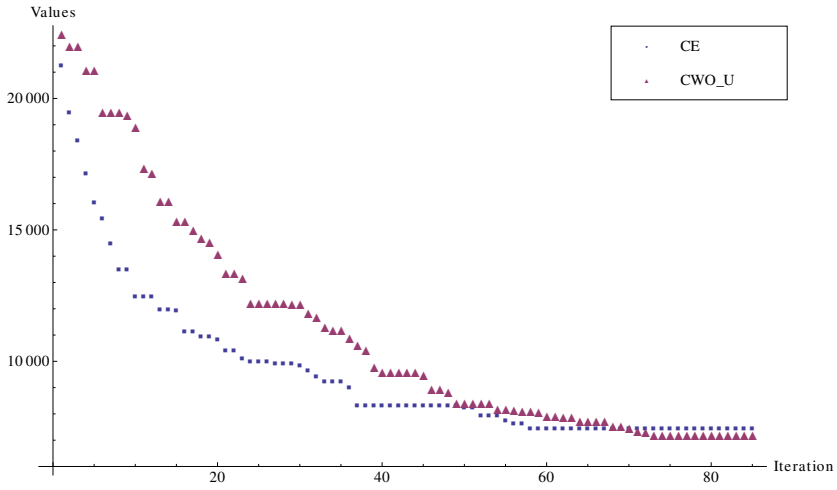


Figure 3.6.3: One trial of CE vs. CWO_U

In Figure 3.6.3, which displays a typical convergence of a single run of CE vs. that of CWO_U, we observe that although CE converges faster at the beginning, CWO_U is able to eventually overtake CE and finishes at a lower value.

3.7 Conclusion

In the first part of this chapter, we proved the convergence of CWO-based algorithms. The proofs provided a rigorous mathematical foundation for the two practical algorithms we proposed in the numerical examples section. These two algorithms are variations of the generic CWO algorithm described in Algorithm 1. The two algorithm variations, CWO_T and CWO_U, differ by how they update their probability density functions over the solution space for each iteration. The first approach, CWO_T, weights the samples according to their outcome values. On the other hand, CWO_U, uniformly weights the samples. We benchmarked the performance of the CWO_T algorithm and summarized the results in Table 3.6.1. Although the numeric values are quite satisfactory, we wanted to see if we could improve these results. This effort led us to the development of the second approach, CWO_U, which we consider as the preferred implementation of the CWO-base algorithm. Perhaps the most surprising fact is that by not taking into account the outcome values of the samples, we are able to achieve better performance results. Even more interesting is the fact that the standard cross-entropy approach is just a limiting case of the CWO_U approach. Comparing the numerical results of CWO_U with those of CE, we believe our algorithm is better at obtaining an optimal solution (see Fig. 3.6.2). Of course, the improvement in performance is at the cost of increasing computational costs.

Chapter 4

Contributions and Future Work

4.1 Contributions

A new family of performance criteria has been proposed in the first part of this thesis. These performance criteria are inspired by cumulative prospect theory, which has substantial empirical support. They include the performance criteria used in classical risk-sensitive control problems (e.g., expected utility). We proved the class of non-convex risk-sensitive control problems is still solvable via dynamic programming. We investigated both finite-horizon and infinite-horizon cases, and offered numerical examples to demonstrate the applications of our approach.

The second part of this thesis presented a novel approach for solving stochastic global optimization problems. This new approach, cumulative weighting optimization, is also inspired by cumulative prospect theory. We proved the convergence to an optimal solution of the cumulative weighting optimization algorithms given a mild assumption on the initial condition. In addition, the algorithms in CWO include the well-known cross-entropy optimization algorithm. Since we have proved the convergence for all CWO algorithms, we naturally have obtained a convergence proof for the cross-entropy algorithm, which to the best of our knowledge, has not been done before. In addition, we presented the numerical analysis of our algorithms, where we compared the performance of two weight updating schemes. We

also compared the performance of our algorithms against that of the cross-entropy.

4.2 Future Work

In the future, we would like to develop a new game theory framework which captures risk-sensitivity. The effects of incorporating the CPT-based distortions could provide a novel perspective into the well-established field of game theory. Take, for example, the classic prisoner's dilemma game. The payoff matrix for the players is given in Table 4.2.1. John von Neumann and Oskar Morgenstern [84] showed that

		Player 2	
	Decision	Cooperate	Defect
Player 1	Cooperate	R=3 R=3	S=0 T=5
	Defect	T=5 S=0	P=1 P=1

R:Reward S:Sucker T:Temptation P:Penalty

Table 4.2.1: Classic Prisoner's Dilemma Problem

a mixed strategy Nash equilibrium exists for any zero-sum game with a finite set of actions. Although the prisoner's dilemma game is not a zero-sum game, analyzing how it reacts to mixed strategies is still important. The analysis of the effects of mixed strategies on the prisoner's dilemma starts with generating the "risk-neutral" (i.e., non-distorted) discretized mixed strategy reward table for player 1. In the matrix below, each element is the expected reward value given the probabilities of player 1 and player 2 cooperate. The i-th row and j-th column entry is calculated as the expected reward with the probability $\frac{i-1}{5}$ that player 1 cooperating and the probability $\frac{j-1}{5}$ that player 2 cooperating, where $i, j \in \{1, 2, \ldots, 6\}$.

$$\begin{pmatrix} 1. & 1.8 & 2.6 & 3.4 & 4.2 & 5. \\ 0.8 & 1.56 & 2.32 & 3.08 & 3.84 & 4.6 \\ 0.6 & 1.32 & 2.04 & 2.76 & 3.48 & 4.2 \\ 0.4 & 1.08 & 1.76 & 2.44 & 3.12 & 3.8 \\ 0.2 & 0.84 & 1.48 & 2.12 & 2.76 & 3.4 \\ 0. & 0.6 & 1.2 & 1.8 & 2.4 & 3. \end{pmatrix}$$

By introducing probability weighting distortion into the prisoner's dilemma game, we alter the analogous risk-sensitive expected reward matrix:

$$\begin{pmatrix} 1. & 1.00021 & 1.35889 & 3.28597 & 4.72747 & 5. \\ 0.931868 & 0.962743 & 1.19336 & 2.63784 & 4.20448 & 4.86374 \\ 0.571493 & 0.758082 & 1.06917 & 2.19005 & 3.35859 & 4.14299 \\ 0.0897219 & 0.354482 & 0.850323 & 2.03271 & 2.87765 & 3.17944 \\ 0.0000523055 & 0.0423175 & 0.533821 & 1.90096 & 2.82637 & 3.0001 \\ 0. & 0.000156917 & 0.269166 & 1.71448 & 2.7956 & 3. \end{pmatrix}.$$

For the matrix displayed above, the weighting function used is

$$w(x) := \exp\left[-3.0\left(-\log(x)\right)^{2.5}\right], \tag{4.2.1}$$

and the risk adjusted expected reward is calculated as:

$$\mathcal{E}(p_1, p_2) := 5\left(w\left(\left(1-p_1\right)p_2\right)\right) + 3\left(w\left(\left(1-p_1\right)p_2 + p_1p_2\right) - w\left(\left(1-p_1\right)p_2\right)\right)$$

$$+1\left(w\left(p_1p_2 + \left(1-p_1\right)p_2 + \left(1-p_1\right)\left(1-p_2\right)\right) - w\left(p_1p_2 + \left(1-p_1\right)p_2\right)\right),$$

where p_1 and p_2 are the probabilities that player 1 and player 2 cooperate, respectively.

The rate that the expected reward decreases down each column for the risk-neutral matrix is constant, whereas the same rate for the risk-sensitive matrix slows down as we traverse toward the bottom of each column. The particular function (Eq. 4.2.1) used represents risk-aversion, hence we see from this example how risk-sensitivity changes player 1's expected payoff.

Aside from having Nash equilibria, the prisoner's dilemma game could also have ϵ-equilibria. An ϵ-equilibrium is formally defined below.

Definition 28. Let $G = \left(N, A = A_1 \times \cdots \times A_N, r : A \to \mathcal{R}^N\right)$ be a N-player game with action sets A_i for each player i and a reward function r. The space of probability distributions over A_i is denoted by $\mathcal{P}(A_i)$. Let $r_i(\pi)$ denote the payoff to player i when the policy $\pi = \{\pi_1 \times \cdots \times \pi_N\}$ is played, where $\pi_i \in \mathcal{P}(A_i)$. A policy π is an ϵ-Nash Equilibrium for the game G if

$$r_i(\pi) \geq r_i(\pi_i', \pi_{-i}) - \epsilon, \quad \forall \pi_i' \in \mathcal{P}(A_i), \ i \in \{1, \ldots, N\},$$

where π_{-i} denotes all the mixed strategies except the i-th policy.

When $\epsilon = 0$, an ϵ-equilibrium is exactly the well-known Nash equilibrium. By picking an $\epsilon > 0$, two additional equilibria are found in the matrix above. More specifically, if player 1 is risk-averse, with an appropriate ϵ, the player could always cooperate, which is not a Nash equilibrium. This example is only one of the many ways in which, by introducing risk-sensitivity into the game, we alter the standard

conclusion. It is interesting to note that the same effect cannot be achieved by using a utility function.

In addition to game theory, we will apply the CPT-based risk-sensitive measures to classic control problems and study the structure of the optimal policies obtained. Systems that are human-centric might find it beneficial to be controlled by an optimal controller derived using our risk-sensitive performance measure. In the future, we will investigate the effects of CPT-based risk-sensitive measures for iterated games.

Appendix A

Prospect Theory

A.1 St. Petersburg Paradox

The example below is from [86].

Example 13. [St. Petersburg paradox]. Consider the following game. A fair coin will be flipped until the first heads shows up. If heads shows up at the k-th flip, then you receive $\$2^k$. Thus, immediate heads gives only \$2, and after each tails the amount doubles. After 19 tails you are sure to be a millionaire. Think for yourself how much it would be worth to you to play this game. The expected value of the game is

$$\frac{1}{2} \times 2 + \frac{1}{4} \times 4 + \frac{1}{8} \times 8 + \frac{1}{16} \times 16 + \cdots = 1 + 1 + 1 + 1 + \cdots = \infty.$$

Thus if you maximize expected value, then this game is worth more to you than any amount of money. In reality, people pay considerably less to participate in the game, something like \$5.

A.2 Axiomatization of Expected Utility:

We need a little notation before we can write down the axioms. Consider two outcomes L_1 and L_2 with known probabilities. We use the notation $L_1 \succcurlyeq L_2$ to

mean that the decision maker prefers L_1 over L_2 or considers them to be equally preferred. We assume that the utility here is continuous and strictly increasing.

Axiom 1.

1. *Completeness: For any two outcomes L_1 and L_2, either $L_1 \succcurlyeq L_2$ or $L_2 \succcurlyeq L_2$ or both.*

2. *Transitivity: For any three outcomes L_1, L_2, and L_3, if $L_1 \succcurlyeq L_2$, and $L_2 \succcurlyeq L_3$, then $L_1 \succcurlyeq L_3$.*

3. *Continuity: For any three outcomes $L_1 \succcurlyeq L_2 \succcurlyeq L_3$, there exist $\alpha, \beta \in (0,1)$ such that $\alpha L_1 + (1-\alpha) L_3 \succcurlyeq L_2$, and $L_2 \succcurlyeq \beta L_1 + (1-\beta) L_3$.*

4. *Substitution (Independence) Savage[80]: For any L_1, L_2 and L_3, and any $\alpha \in (0,1)$, $L_1 \succcurlyeq L_2$ if and only if $\alpha L_1 + (1-\alpha) L_3 \succcurlyeq \alpha L_2 + (1-\alpha) L_3$.*

Furthermore, von Neumann and Morgenstern [85] proved the following:

$$L_1 \succcurlyeq L_2 \text{ if and only if } \sum_{i=1}^{n} p_i^1 u\left(x_i^1\right) \geq \sum_{i=1}^{n} p_j^2 u\left(x_j^2\right),$$

where p_i^1 is the probability for the i-th outcome of L_1, x_i^1 is the value for the i-th outcome of L_1 and u is the utility function.

Appendix B

Multifunctions and Selectors

Our main source of reference for this section is [48].

Let X and A be (nonempty) Borel spaces.

Definition 29. A *multifunction* (also known as a correspondence or set-valued mapping) ψ from X to A is a function such that $\psi(x)$ is a nonempty subset of A for all $x \in X$. (A single-valued mapping $\psi : X \to A$ is of course an example of a multifunction.) The graph of the multifunction ψ is the subset of $X \times A$ defined as

$$Gr(\psi) := \{(x, a) \,|x \in X,\ a \in \psi(x)\}.$$

A multifunction could have one of the properties described below. For every subset B of A, let $\psi^{-1}[B] := \{x \in X | \psi(x) \cap B \neq \emptyset\}$.

Definition 30. A multifunction ψ from X to A is said to be

(a) *Borel measurable* if $\psi^{-1}[G]$ is a Borel subset of X for every open set $G \subset A$;

(b) upper semi-continuous ($u.s.c$) if $\psi^{-1}[F]$ is closed in X for every closed set $F \subset A$;

(c) lower semi-continuous ($l.s.c$)if $\psi^{-1}[G]$ is open in X for every open set $G \subset A$;

(d) continuous if it is both $u.s.c$ and $l.s.c$.

A multifunction ψ is said to be *closed-valued (resp. compact-valued) if* $\psi(x)$ *is a closed (resp. compact) set for all* $x \in X$. *The multifunction is said to be closed if its graph is closed.*

Proposition 5. *Let* ψ *be a compact-valued multifunction from* X *to* A. *Then the following statements are equivalent:*

(a) ψ *is Borel-measurable;*

(b) $\psi^{-1}[F]$ *is a Borel subset of* X *for every closed set* $F \subset A$;

(c) $Gr(\psi)$ *is a Borel subset of* $X \times A$;

(d) ψ *is a measurable function from* X *to the space of nonempty compact subsets of* A *topologized by the Hausdorff metric.*

Proof. See [50] and [81]. □

Throughout the remainder of this appendix, ψ is a given Borel-measurable multifunction from X to A, and $\mathbb{F}$ denotes the set of (single-valued) measurable multifunction from X to A, and $\mathbb{F}$ denotes the set of (single-valued) measurable functions $f : X \to A$ with $f(x) \in \psi(x)$ for all $x \in X$. A function $f \in \mathbb{F}$ is called a selector (or measurable selector or choice or decision function) for the multifunction ψ. Moreover, $v : Gr(\psi) \to \mathbb{R}$ is a given measurable function and

$$v^*(x) := \inf_{\psi(x)} v(x, a), \ x \in X.$$

If $v(x, \cdot)$ attains its minimum at some point in $\psi(x)$, we write "min" instead of "inf."

Proposition 6. *Suppose that ψ is compact-valued.*

(a) If $v(x,\cdot)$ is l.s.c. on $\psi(x)$ for every $x \in X$, then exists a selector $f^* \in \mathbb{F}$ such that

$$v(x, f^*(x)) = v^*(x) = \min_{\psi(x)} v(x,a), \ \forall x \in X$$

and $v^*$is measurable. Similarly, if $v(x,\cdot)$ is u.s.c. on $\psi(x)$ for every $x \in X$, then exists a selector $f^* \in \mathbb{F}$ such that

$$v(x, f^*(x)) = v^*(x) = \max_{\psi(x)} v(x,a), \ \forall x \in X$$

and v^* is measurable.

(b) If ψ is u.s.c and v is l.s.c and bounded below on $Gr(\psi)$, then there exists $f^* \in \mathbb{F}$ for which

$$v(x, f^*(x)) = v^*(x) = \min_{\psi(x)} v(x,a), \ \forall x \in X$$

holds, and v^* is l.s.c. and bounded below on X.

Appendix C

Spaces of Probability Measures

The content of this section can be found in the appendices of [15] and [13].

C.1 Polish Spaces

A Polish space, $\mathcal{X}$, is a topological space which is separable and admits a complete metrization. Examples of such spaces are: separable Banach spaces, compact metric spaces, the space $D[0,1]$ of cadlag path from $[0,1]$ to $\mathcal{R}$ with Skorohod topology. Let $\mathcal{X}$ be a Polish space and $d(\cdot,\cdot)$ a complete metric on it.

Definition 31. A probability measure $\mathbb{P}$ on $\mathcal{X}$ is said to be *tight* if for each $\epsilon > 0$, there exists a compact set $K_\epsilon \subset \mathcal{X}$ with $\mathbb{P}(K_\epsilon) \geq 1 - \epsilon$. Analogously, a family $\mathbb{P}_\alpha$, $\alpha \in \mathcal{I}$, ($\mathcal{I}$ being an index set) is said to be tight if the above holds for all $\mathbb{P}_\alpha$ uniformly in α, i.e., the set K_ϵ above can be chosen to be the same for all α.

C.2 The Prohorov Topology

Let $C_b(\mathcal{X})$, $\mathcal{P}_\mathcal{X}$ denote respectively the space of bounded continuous real valued functions on $\mathcal{X}$, and the space of probability measures on $\mathcal{X}$. Endow $C_b(\mathcal{X})$ with the supremum norm $\|\cdot\|$. $\mathcal{P}_\mathcal{X}$ will be given the topology for with a local base at

$\mathbb{P} \in \mathcal{P}_{\mathcal{X}}$ is given by the sets of the type

$$\left\{ \mathbb{Q} \in \mathcal{P}_{\mathcal{X}} | \left| \int f_i d\mathbb{Q} - \int f_i d\mathbb{P} \right| < \epsilon_i,\ 1 \leq i \leq k \right\}$$

for some $k \geq 1,\ \epsilon_i > 0$ and $f_i \in C_b(\mathcal{X})$ for $1 \leq i \leq k$.

It is easily seen that this topology is Hausdorff and is coarser than the one induced by the total variation norm. It is called the Prohorov topology or the topology of weak convergence. Some other possible choices for the local basis at $\mathbb{P} \in \mathcal{P}_{\mathcal{X}}$ are given below.

$$\left\{ \mathbb{Q} \in \mathcal{P}_{\mathcal{X}} | \mathbb{Q}(F_i) < \mathbb{P}(F_i) + \epsilon_i,\ 1 \leq i \leq k \right\},\ F_i \subset \mathcal{X} \text{ closed}$$

$$\left\{ \mathbb{Q} \in \mathcal{P}_{\mathcal{X}} | \mathbb{Q}(G_i) > \mathbb{P}(G_i) - \epsilon_i,\ 1 \leq i \leq k \right\},\ G_i \subset S \text{ open}$$

$$\left\{ \mathbb{Q} \in \mathcal{P}_{\mathcal{X}} | \left| \mathbb{Q}(A_i) - \mathbb{P}(A_i) \right| < \epsilon_i,\ 1 \leq i \leq k \right\},\ A_i \subset S$$

satisfy $\mathbb{P}(\partial A_i) = 0$ where ∂A_i is the boundary of A_i,

$$\left\{ \mathbb{Q} \in \mathcal{P}_{\mathcal{X}} | \left| \int f_i d\mathbb{P} - \int f_i d\mathbb{Q} \right| < \epsilon_i,\ 1 \leq i \leq k \right\},$$

f_i are bounded and uniformly continuous with respect to the metric d. Here, $\epsilon > 0,\ k \leq 1$.

C.3 Compactness in $\mathcal{P}_{\mathcal{X}}$

Theorem 20. *A subset $\mathcal{L} \subset \mathcal{P}_{\mathcal{X}}$ is relatively compact if and only if it is tight.*

C.4 Metrics on $\mathcal{P}_{\mathcal{X}}$

Definition 32. For any $\epsilon > 0$ and any Borel set $A \subset \mathcal{X}$, let $A^{\epsilon} = \{x \in \mathcal{X} | d(x, A) < \epsilon\}$.

For $\mu,\ \nu \in \mathcal{P}_{\mathcal{X}}$, we define the metric

$$d(\mu, \nu) = \inf_{\epsilon} \{\epsilon > 0 | \mu(A) \leq \nu(A^{\epsilon}) + \epsilon,\ \nu(A) \leq \mu(A^{\epsilon}) + \epsilon \text{ for all Borel subset A of } S\}.$$

Theorem 21. *$d(\cdot, \cdot)$ defines a metric on $\mathcal{P}_{\mathcal{X}}$ consistent with the Prohorov topology.*

References

[1] O. Alagoz, L. M. Maillart, A. J. Schaefer, and M. S. Roberts, *The optimal timing of living-donor liver transplantation*, Management Science, 50 (2004), pp. 1420–1430.

[2] M. Allais, *Le comportement de lHomme rationnel devant le risque: Critique des postulats et axiomes de lÉcole americaine*, Econometrica, 21 (1953), pp. 503–546. ArticleType: research-article / Full publication date: Oct., 1953 / Copyright © 1953 The Econometric Society.

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

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

[5] R. Bellman, *On the theory of dynamic programming*, Proceedings of the National Academy of Sciences of the United States of America, 38 (1952), pp. 716–719. PMID: 16589166 PMCID: PMC1063639.

[6] ——, *Applied Dynamic Programming*, Princeton University Press, Princeton, 1957.

[7] D. Bernoulli, *Exposition of a new theory on the measurement of risk*, Econometrica, 22 (1954), pp. 23–36. ArticleType: research-article / Full publication date: Jan., 1954 / Copyright © 1954 The Econometric Society.

[8] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, Athena Scientific, 2nd ed., nov 2000.

[9] D. P. Bertsekas, *Dynamic Programming and Optimal Control: Approximate dynamic programming*, Athena Scientific, 2012.

[10] D. P. Bertsekas and S. E. Shreve, *Stochastic Optimal Control: The Discrete Time Case*, Academic Press, Nov. 1978.

[11] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*, Athena Scientific, 1996.

[12] N. P. Bhatia and G. P. Szegö, *Stability Theory of Dynamical Systems*, Springer, 1970.

[13] P. Billingsley, *Convergence of Probability Measures*, Wiley-Interscience, 2 ed., July 1999.

[14] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, Journal of Political Economy, 81 (1973), pp. 637–654. ArticleType: research-article / Full publication date: May - Jun., 1973 / Copyright ©1973 The University of Chicago Press.

[15] ——, *Topics in Controlled Markov Chains*, CRC Press, Apr. 1991.

[16] ——, *Probability Theory: An Advanced Course*, Springer, Oct. 1995.

[17] M. Bouakiz and M. J. Sobel, *Inventory control with an exponential utility criterion*, Operations Research, 40 (1992), pp. 603–608. ArticleType: research-article / Full publication date: May - Jun., 1992 / Copyright ©1992 INFORMS.

[18] O. Çavuş and A. Ruszczyński, *Risk-averse control of undiscounted transient markov models*, arXiv preprint arXiv:1203.5437, (2012).

[19] A. Chateauneuf and P. Wakker, *An axiomatization of cumulative prospect theory for decision under risk*, Journal of Risk and Uncertainty, 18 (1999), pp. 137–145.

[20] P. Cheridito, F. Delbaen, and M. Kupper, *Coherent and convex monetary risk measures for bounded cádlág processes*, Stochastic Processes and their Applications, 112 (2004), pp. 1–22.

[21] ——, *Coherent and convex monetary risk measures for unbounded cádlág processes*, Finance and Stochastics, 10 (2006), pp. 427–448.

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

[23] K.-J. Chung and M. J. Sobel, *Discounted MDP's: distribution functions and exponential utility maximization*, SIAM journal on control and optimization, 25 (1987), pp. 49–62.

[24] S. P. Coraluppi and S. I. Marcus, *Risk-sensitive and minimax control of discrete-time, finite-state markov decision processes*, Automatica, 33 (1999), pp. 301–309.

[25] ——, *Mixed risk-neutral/minimax control of discrete-time, finite-state markov decision processes*, IEEE Transactions on Automatic Control, 45 (2000), pp. 528 –532.

[26] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, *A tutorial on the cross-entropy method*, Annals of Operations Research, 134 (2005), pp. 19–67.

[27] B. de Finetti, *Logical foundations and measurement of subjective probability*, Acta Psychologica, 34 (1970), pp. 129–145.

[28] F. Delbaen and E. T. Hochschule, *Coherent risk measures on general probability spaces*, in In essays in honour of Dieter Sondermann, Springer, 2002, pp. 1–37.

[29] E. Diecidue, U. Schmidt, and H. Zank, *Parametric weighting functions*, Journal of Economic Theory, 144 (2009), pp. 1102–1118.

[30] P. Dupuis, M. R. James, and I. Petersen, *Robust properties of risk-sensitive control*, Mathematics of Control, Signals, and Systems (MCSS), 13 (2000), pp. 318–332.

[31] A. Eichhorn and W. Römisch, *Polyhedral risk measures in stochastic programming*, SIAM Journal of Optimization, 16 (2005), pp. 69–95.

[32] E. A. Feinberg and A. Shwartz, *Handbook of Markov Decision Processes: Methods and Applications*, Springer, 2002.

[33] E. Fernández-Gaucherand and S. I. Marcus, *Risk-sensitive optimal control of hidden markov models: structural results*, IEEE Transactions on Automatic Control, 42 (1997), pp. 1418 –1422.

[34] P. C. Fishburn, *Utility theory for decision making*, tech. rep., DTIC Document, June 1970.

[35] ——, *The axioms of subjective probability*, Statistical Science, 1 (1986), pp. 335–345. Mathematical Reviews number (MathSciNet): MR858514.

[36] W. H. Fleming, *Optimal long term growth rate of expected utility of wealth*, The Annals of Applied Probability, 9 (1999), pp. 871–903. Mathematical Reviews number (MathSciNet): MR1722286; Zentralblatt MATH identifier: 0962.91036.

[37] W. H. Fleming and S. J. Sheu, *Risk-sensitive control and an optimal investment model*, Mathematical Finance, 10 (2000), pp. 197–213.

[38] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer, 2nd ed., Nov. 2005.

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

[40] H. Föllmer and A. Schied, *Convex measures of risk and trading constraints*, Finance and Stochastics, 6 (2002), pp. 429–447.

[41] H. Föllmer, A. Schied, and T. Lyons, *Stochastic finance. an introduction in discrete time*, The Mathematical Intelligencer, 26 (2004), pp. 67–68.

[42] T. D. Frank, *Nonlinear Fokker-Planck Equations - Fundamentals and Applications*, Springer-Verlag, 2005.

[43] M. Frittelli and G. Scandolo, *Risk measures and capital requirments for processes*, Mathematical Finance, 16 (2006), pp. 589–612.

[44] U. G. Haussmann, *Some examples of optimal stochastic controls or: The stochastic maximum principle at work*, SIAM Review, 23 (1981), pp. 292–307. ArticleType: research-article / Full publication date: Jul., 1981 / Copyright © 1981 Society for Industrial and Applied Mathematics.

[45] X. D. He and X. Y. Zhou, *Portfolio choice via quantiles*, Mathematical Finance, 21 (2011), pp. 203–231.

[46] D. Hernández-Hernández and S. I. Marcus, *Risk sensitive control of markov processes in countable state space*, Systems & Control Letters, 29 (1996), pp. 147–155.

[47] ——, *Existence of risk-sensitive optimal stationary policies for controlled markov processes*, Applied Mathematics & Optimization, 40 (1999), pp. 273–285.

[48] O. Hernández-Lerma and J. B. Lasserre, *Discrete-Time Markov Control Processes: Basic Optimality Criteria*, Springer, 1 ed., Dec. 1996.

[49] O. Hernández-Lerma and J.-B. Lasserre, *Further topics on discrete-time Markov control processes*, Springer, June 1999.

[50] C. J. Himmelberg, T. Parthasarathy, and F. S. VanVleck, *Optimal plans for dynamic programming problems*, Mathematics of Operations Research, 1 (1976), pp. 390–394.

[51] J. Hofbauer, J. Oechssler, and F. Riedel, *Brown-von Neumann-Nash dynamics: The continuous strategy case*, Games and Economic Behavior, 65 (2009), pp. 406–429.

[52] R. A. Howard and J. E. Matheson, *Risk-sensitive markov decision processes*, Management Science, 18 (1972), pp. 356–369. ArticleType: research-article / Issue Title: Theory Series / Full publication date: Mar., 1972 / Copyright ©1972 INFORMS.

[53] J. Hu, M. C. Fu, and S. I. Marcus, *A model reference adaptive search method for global optimization*, Operations Research, 55 (2007), pp. 549–568.

[54] J. Hu, Y. Wang, E. Zhou, M. C. Fu, and S. I. Marcus, *A survey of some model-based methods for global optimization*, in Optimization, Control, and Applications of Stochastic Systems, D. Hernández-Hernández and J. A. Minjárez-Sosa, eds., Birkhäuser Boston, Boston, 2012, pp. 157–179.

[55] H. W. James and E. J. Collins, *An analysis of transient markov decision processes*, Journal of Applied Probability, 43 (2006), pp. 603–621. Mathematical Reviews number (MathSciNet): MR2274787; Zentralblatt MATH identifier: 1145.90099.

[56] S. C. Jaquette, *Markov decision processes with a new optimality criterion: Discrete time*, The Annals of Statistics, 1 (1973), pp. 496–505. Mathematical Reviews number (MathSciNet): MR378839; Zentralblatt MATH identifier: 0259.90054.

[57] S. C. Jaquette, *A utility criterion for markov decision processes*, Management Science, 23 (1976), pp. 43–49. ArticleType: research-article / Full publication date: Sep., 1976 / Copyright ©1976 INFORMS.

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

[59] D. Kahneman and A. Tversky, *Prospect theory: an analysis of decision under risk*, National Emergency Training Center, 1979.

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

[61] V. N. Kolokoltsov, *Nonlinear Markov Processes and Kinetic Equations*, Cambridge University Press, July 2010.

[62] ——, *Markov Processes, Semigroups and Generators*, Walter de Gruyter, Mar. 2011.

[63] H. J. Kushner, *Introduction to Stochastic Control*, Holt, Rinehart and Winston, 1971.

[64] J. Leitner, *A short note on second-order stochastic dominance preserving coherent risk measures*, Mathematical Finance, 15 (2005), pp. 649–651.

[65] J. Oechssler and F. Riedel, *Evolutionary dynamics on infinite strategy spaces*, Economic Theory, 17 (2001), pp. 141–162.

[66] S. Onay and A. Öncüler, *Intertemporal choice under timing risk: An experimental approach*, Journal of Risk and Uncertainty, 34 (2007), pp. 99–121.

[67] S. Peng, *A general stochastic maximum principle for optimal control problems*, SIAM Journal on Control and Optimization, 28 (1990), p. 966–979. ACM ID: 84050.

[68] H. Pham, *Continuous-time Stochastic Control and Optimization with Financial Applications*, Springer, 1 ed., July 2009.

[69] S. R. Pliska, *Dynamic Programming and Its Applications*, Academic Press, 1978, ch. On the transient case for Markov decision chains with general state spaces, pp. 335–349.

[70] S. R. Pliska, *A stochastic calculus model of continuous trading: Optimal portfolios*, Mathematics of Operations Research, 11 (1986), pp. 371–382. Ar-

ticleType: research-article / Full publication date: May, 1986 / Copyright © 1986 INFORMS.

[71] T. Post, M. J. van den Assem, G. Baltussen, and R. H. Thaler, *Deal or no deal? decision making under risk in a large-payoff game show*, American Economic Review, 98 (2008), pp. 38–71.

[72] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*, John Wiley & Sons, Sept. 2007.

[73] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, Inc., New York, NY, USA, 1st ed., 1994.

[74] F. Riedel, *Dynamic coherent risk measures*, Stochastic Processes and their Applications, 112 (2004), pp. 185–200.

[75] R. Y. Rubinstein, *Combinatorial optimization, cross-entropy, ants and rare events*, in Stochastic Optimization: Algorithms and Applications, P. M. P. S. Uryasev, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2001, pp. 304–358.

[76] A. Ruszczyński, *Risk-averse dynamic programming for markov decision processes*, Mathematical Programming, 125 (2010), pp. 235–261.

[77] A. Ruszczyński and A. Shapiro, *Conditional risk mappings*, Mathematics of Operations Research, 31 (2006), pp. 544–561. ArticleType: research-article / Full publication date: Aug., 2006 / Copyright ©2006 INFORMS.

[78] ——, *Optimization of convex risk functions*, Mathematics of Operations Research, 31 (2006), pp. 433–452. ArticleType: research-article / Full publication date: Aug., 2006 / Copyright ©2006 INFORMS.

[79] ——, *Optimization of risk measures*, in Probabilistic and Randomized Methods for Design under Uncertainty, G. Calafiore and F. Dabbene, eds., Springer London, 2006, pp. 119–157.

[80] L. J. Savage, *The Foundation of Statistics*, Courier Dover Publications, 1972.

[81] M. Schäl, *Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal*, Probability Theory and Related Fields, 32 (1975), pp. 179–196.

[82] C. Starmer, *Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk*, Journal of Economic Literature, 38 (2000), pp. 332–382. ArticleType: research-article / Full publication date: Jun., 2000 / Copyright © 2000 American Economic Association.

[83] A. Tversky and D. Kahneman, *Advances in prospect theory: Cumulative representation of uncertainty*, Journal of Risk and Uncertainty, 5 (1992), pp. 297–323.

[84] J. von Neumann and O. Morgenstern, *The Theory of Games and Economic Behavior*, The Theory of Games and Economic Behavior, Princeton University Press, 1947.

[85] J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior (Commemorative Edition)*, Princeton University Press, Mar. 2007.

[86] P. P. Wakker, *Prospect Theory: For Risk and Ambiguity*, Cambridge University Press, July 2010.

[87] Y. Wang, *Simulation-Based Methods for Stochastic Control and Global Optimization*, PhD thesis, University of Maryland - College Park, 2011.

[88] Y. Wang, M. C. Fu, and S. I. Marcus, *Model-based evolutionary optimization*, in Proceedings of the Winter Simulation Conference, WSC '10, Winter Simulation Conference, 2010, pp. 1199—1210.

[89] J. Yong and X. Y. Zhou, *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer, 1 ed., June 1999.

[90] E. Zeidler, *Nonlinear Functional Analysis and Its Applications: Part 2 B: Nonlinear Monotone Operators*, Springer, Dec. 1989.