ABSTRACT

Title of dissertation: Decision Making Under Uncertainty:
New Models and Applications
Cheng Jie, Doctor of Philosophy, 2018

Dissertation directed by: Professor Michael C. Fu
Robert H. Smith School of Business
and Institute for Systems Research

In the settings of decision-making-under-uncertainty problems, an agent takes an action on the environment and obtains a non-deterministic outcome. Such problem settings arise in various applied research fields such as financial engineering, business analytics and speech recognition. The goal of the research is to design an automated algorithm for an agent to follow in order to find an optimal action according to his/her preferences.

Typically, the criterion for selecting an optimal action/policy is a performance measure, determined jointly by the agent's preference and the random mechanism of the agent's surrounding environment. The random mechanism is reflected through a random variable of the outcomes attained by a given action, and the agent's preference is captured by a transformation on the potential outcomes from the set of possible actions.

Many decision-making-under-uncertainty problems formulate the performance measure objective function and develop optimization schemes on that objective function. Although the idea on the high-level seems straightforward, there are many challenges, both conceptually and computationally, that arise in the process of finding the optimal action.

The thesis studies a special class of performance measure, defined based on Cumulative Prospect Theory (CPT), which has been used as an alternative to expected-utility based performance measure for evaluating human-centric systems. The first part of the thesis designs a simulation-based optimization framework on the CPT-based performance measure. The framework includes a sample-based estimator for the CPT-value and stochastic approximation algorithms for searching the optimal action/policy. We prove that, under reasonable assumptions, the CPT-value estimator is asymptotically consistent and our optimization algorithms are asymptotically converging to the optimal point. The second part of the thesis introduces an abstract dynamic programming framework whose transitional measure is defined through the CPT-value. We also provide sufficient conditions under which the CPT-driven dynamic programming would attain a unique optimal solution. Empirical experiments presented in the last part of thesis illustrate that the CPT-estimator is consistent and that the CPT-based performance measure may lead to an optimal policy very different from those obtained using traditional expected utility.

DECISION MAKING UNDER UNCERTAINTY: NEW MODELS AND APPLICATIONS

by

Cheng Jie

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
2018

Advisory Committee:
Professor Michael C. Fu, Chair/Advisor
Professor Leonid Koralov
Professor Steve I. Marcus, Co-Advisor
Professor Ilya O. Ryzhov
Professor Eric V. Slud

Dedication

To my family.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Professor Michael C. Fu, for giving me an invaluable opportunity to work on challenging and extremely interesting projects over the past four years. He has always made himself available for help and advice. And it has been a pleasure to work with and learn from such an extraordinary individual.

I would also like to thank my co-advisor, Prof Steven I. Marcus. His advises are always brilliant and his guidance is an important part of this thesis. Meanwhile, his dedication and patience have made a lasting impression on me.

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 Eric Slud, I learned priceless knowledge through his statistics course and gained many insights towards my research through our discussions. Professor Leonid Koralov is always there to help me when I struggled in mathematical problems. Professor Ryzhov's course on stochastic optimization has inspired me a lot when learning and researching in that field.

Also, I would like to express many thanks to my co-workers on the papers we finished throughout my four year P.h.D. in the university. Prashanth L.A. is a wonderful person to work with and gave me invaluable help during the two years he spent in Maryland. Kun Lin provided me a lot of guidance on my research through his previous work.

I have wonderful colleagues in the group meeting, Guowei Sun, James Ferlez, Bhaskar Ramasubramanian, Yunchuan Li. I would like to thank them for the many interesting discussions on topics both technical and non-technical.

My classmates in the mathematics department are one of the best groups to get along with, and the time I spent with them made my four years graduate school experience unforgettable.

I owe my deepest thanks to my family who have always stood by me and guided me through my career, and have pulled me through against impossible odds at times. Words cannot express the gratitude I owe them.

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.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CPT	Cumulative Prospect Theory
DP	Dynamic Programming
IPA	Infinitesimal Perturbation Analysis
MDP	Markov Decision Process
MPS	Model-based Parameter Search
SPSA	Simultaneous Perturbation Stochastic Approximation
SPSA-N	Simultaneous Perturbation Stochastic Approximation with Newton method
SSP	Stochastic Short Path
TLC	Traffic Light Control

Chapter 1

Introduction

1.1 Decision making

In decision making problem settings, an agent is an individual acting based on the observations of its surrounding environment. Agents can either be physical or nonphysical entities. Physical entities include humans and robots, and a nonphysical entity is usually a decision support system that is completely implemented in software. Figure 1.1 shows that the interaction between the agent and the world follows an observe-act cycle.

At time t, the agent receives an observation o_t resulting from the action a_t it carries out. Observations are often incomplete or noisy. And the agent's act may have a nondeterministic effect on the environment. We focus on the scenarios in which agents interact intelligently with the environment over time. Given the historical trajectories of observations and knowledge about the environment, the agent will choose a_t that optimally achieves its objectives.

1.2 Example Applications

There are many examples of problems in which accounting for uncertainty is necessary. This section outlines few of them, and the first of which will be revisited in the latter part of this thesis.

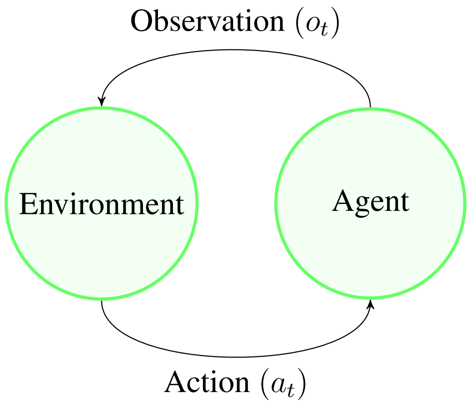


Figure 1.1: Interaction between the environment and the agent

1.2.1 Traffic light control system

An example of a decision support system that plays an important role in city planning is a traffic signal control system (TLCS). A TLCS coordinates individual traffic signals to achieve the objectives of network-wide traffic operations objectives. These systems consist of intersection traffic signals, a communications network to tie them together, and a central computer or network of computers to manage the whole system. Coordination can be implemented through several techniques including time-based and hardwired interconnection methods.

The key purpose of a traffic-signal system is to deliver favorable signal timings to motorists. Designing an efficient automated traffic-signal system appears to be an arduous task, because of the scale of the problem but also the unpredictable nature patterns of drivers' behaviors. A vast amount of research has been devoted to designing a better control system.

1.2.2 Keyword auction

Keyword auction, also known as sponsored search auction, refers to the results from a search engine that generate separate advertisements paid by third parties rather than through the main search algorithm [47]. These advertisements are typically related to the search terms and contain a link with some information, with the hope that the click can convert a valuable action to the advertisers. In sponsored search auctions, there are usually limited numbers of slots as compared to the number of advertisers, so an auction is used to determine how the slots will be assigned.

Sponsored search auction constitutes a big part of search engine marketing. It provides a marketplace where advertisers can bid for advertising opportunities to enhance customers' impressions on their products and related websites, and therefore increase the volume of their sales. From the perspective of an advertiser, designing an effective bidding strategy is critical to their success.

The main challenge of a reward strategy is to choose effective keywords and phrases. There is a myriad of keywords and phrases that can be chosen from [18]. And various forms of uncertainty prevent advertisers from determining their optimal bidding strategies.

Given the economic impact and intricate nature aforementioned, sponsored search auctions have been drawing a lot of attention from both researchers and practitioners.

1.3 Methods for Designing Decision Agents

There are many different methods for designing decision agents. The methods differ in the responsibilities of the designer and the tasks left to automation. We briefly outline a few of the methods in this section, and the thesis will primarily focus on one of them, optimization.

1.3.1 Explicit Programming

In some simple settings, one can explicitly lay out a action for the agent to execute by anticipating all the different scenarios the agent might find itself in the environment [48]. Explicit Programming is the most direct method for designing a decision agent. However, in most of the cases, it is generally impossible for a designer to provide a complete strategy.

1.3.2 Optimization

Another approach is for the designer to specify the space of possible decision strategies and a performance measure to be maximized. Evaluating the performance of a decision strategy generally involves running a batch of simulations with the decision strategy. The optimization algorithm then performs a search in this space for the optimal strategy. If the space of possible strategies is relatively low dimensional and the performance measure does not have many local optima, then various local or global search strategies may be appropriate. Although knowledge of a dynamic model is generally assumed in order to run the simulations, it is not otherwise used to guide the search for the optimal strategy, which can be important in complex problems.

1.3.3 Reinforcement Learning

Reinforcement learning (RL) is an area of machine learning concerning how agents should take actions in an environment so as to maximize some notion of cumulative reward. Under the principles of reinforcement learning, the decision-making strategy is learned while the agent interacts with the world. One of the interesting complexities that arises in reinforcement learning is that the choice of action impacts not only the immediate success of the agent in achieving its objectives but also the agent's ability to learn about the environment and identify the characteristics of the problem that it can exploit. Due to its generality, reinforcement learning problem has been studied in many disciplines, such as control theory, game theory and simulation optimization. A comprehensive introduction to reinforcement learning can be found in [77].

1.4 Representation of Uncertainty

Uncertainty can arise from incomplete observation about the state of the agent's surrounding environment, and given the incomplete information at hand, we are unable to make an assessment with complete accuracy. Uncertainty can also arise from our inability to fully predict future events. For example, a hedge fund manager cannot predict exactly how a stock price would behave even a few hours after a buyout action, and a traffic control system cannot fully predict the behavior of all the drivers on the road.

Uncertainty is reflected in the randomness of the outcome, and such randomness is usually represented by a probability distribution. Probability distributions used to represent uncertainty take forms ranging from the simple and well-known Gaussian distribution to complex graphical probabilistic models. Notice that in certain cases of decision-making-under-uncertainty problems, the underlying probability distributions are unknown to the agent. Under such circumstances, the agent usually needs to apply statistical learning methods, either parametric or nonparametric, to estimate the probability distributions through the samples of the outcome of the action.

A detailed discussion of representing uncertainty can be found in [48].

1.5 Utility Theory

This section briefly introduces the foundation of utility theory and how it forms the basis for decision making under uncertainty.

1.5.1 Utility measure/function

For any decision-making-under-uncertainty problem, the utility measure, usually denoted in this dissertation as u, is a real-valued function applied on the space of all the outcomes such that it represents the preference of the agent on the outcome. To illustrate this point, suppose A_1 and A_2 are two observed outcomes and the agent prefers A_1 over A_2, then $u(A_1) > u(A_2)$. Moreover, the utility function is unique up to *affine transformation*.

1.5.2 Expected utility theory

Given the probability distribution of the outcomes and the utility function, defining an appropriate performance measure is of paramount importance. An inappropriate performance measure may lead to an inefficient policy. A widely used form of performance measure is the expected value of the utilities.

Expected utility theory emerges from assumptions about agents' behavior, and a critical one of them is rationality. Vast amounts of literature have been written to explain why a rational agent always chooses to maximize expected utility. One prominent argument depends on evidence that expected-utility maximization is a profitable policy in the long term. A version of this argument can be found in the work of [30], and it relies on a well-known limit theorem in statistics: the Strong Law of Large Numbers. There are also arguments based on representation theorems, which suggest that certain rational constraints on preference entail that all rational agents maximize expected utility. Among the variations of the arguments based on representation theorems, one influential work is [56], which claims that preferences can be defined over a domain of lotteries.

1.5.3 Irrationality assumption

Despite its popularity, expected utility theory does have some limitations, especially when the agent in the decision-making process is human, since humans are subject to various emotional and cognitive biases when making decisions. As the psychology literature points out, human preferences are inconsistent with expected utilities regardless of what nonlinear forms of utility functions are used [1, 28, 46]. Such an argument is justified by a celebrated experiment called the Allais Paradox [1], which is briefly outlined here:

Allais Paradox

Suppose we have the following two switching gambling policies:

[Policy 1] A gain (number of vehicles that reach destination per unit time) of 1000 w.p. 1. Let this be denoted by $(1000, 1)$.

[Policy 2] $(10000, 0.1; 1000, 0.89; 100, 0.01)$ i.e., gains 10000, 1000 and 100 with respective probabilities 0.1, 0.89 and 0.01.

Humans usually choose Policy 1 over Policy 2. On the other hand, consider the following two policies:

[Policy 3] (100, 0.89; 1000, 0.11)

[Policy 4] (100, 0.9; 10000, 0.1)

Humans usually choose Policy 4 over Policy 3.

We can now argue against using expected utility (EU) as an objective as follows: Let u be the utility function in EU.

$$\begin{aligned}
&\text{Policy 1 is preferred over Policy 2}\\
&\Rightarrow u(1000) > 0.1u(10000) + 0.89u(1000) + 0.01u(100)\\
&\Rightarrow 0.11u(1000) > 0.1u(10000) + 0.01u(100) \qquad (1.1)
\end{aligned}$$

$$\begin{aligned}
&\text{Policy 4 is preferred over Policy 3}\\
&\Rightarrow 0.89u(100) + 0.11u(1000) < 0.9u(100) + 0.1u(10000)\\
&\Rightarrow 0.11u(1000) < 0.1u(10000) + 0.01u(100) \qquad (1.2)
\end{aligned}$$

And we have a contradiction from (1.1) and (1.2).

It has been argued that even human experts may have an inconsistent set of preferences, which can be problematic when designing a decision support system that attempts to maximize expected utility[48] .

1.5.4 Cumulative prospect theory

In terms of modeling humans' preferences, one of the approaches is [46]'s celebrated *prospect theory* (PT). The theory claims that people evaluate potential losses and gains based on certain heuristics, and make choices based on that evaluation. The theory tries to model real-life choices, rather than optimal decisions.

To formulate the model of prospect theory, denote X as the random variable of the outcomes, and let $p_i, i = 1, \ldots, K$ denote the probability of incurring a gain/loss $x_i, i = 1, \ldots, K$. Given a utility function u and weighting function w, the prospect theory (PT) value is defined as $\mathbb{P}(X) = \sum_{i=1}^{K} u(x_i)w(p_i)$.

The idea of modeling humans' choice through prospect theory is to take an utility function that is S-shaped, so that it satisfies the *diminishing sensitivity* property. If we take the weighting function w to be the identity, then one recovers the classic expected utility. A general weight function inflates low probabilities and deflates high probabilities, and this has been shown to be close to the way humans make decisions (see [46], [31] for a justification, in particular via empirical tests using human subjects).

However, *Prospect Theory* gave rise to violations of first-order *stochastic dominance*. Consider the following example from [31]: Suppose there are 20 prospects (outcomes) ranging from -10 to 180, each with probability 0.05. If the weight function is such that $w(0.05) > 0.05$, then it uniformly overweights all *low-probability* prospects and the resulting PT value is higher than the expected value 85. This violates stochastic dominance, since a shift in the probability mass from bad outcomes did not result in a better prospect.

Our thesis is based on *cumulative prospect theory* (CPT), a later, refined variant of prospect theory due to [78]. CPT generalizes expected utility theory in that in addition to having a utility function transforming the outcome, it introduces another function which distorts the cumulative distribution function. As compared to prospect theory, CPT is monotone with respect to stochastic dominance, a property that is thought to be useful and (mostly) consistent with human preferences.

Cumulative prospect theory (CPT) uses a similar measure as PT, except that the weights are a function of cumulative probabilities. First, separate the gains and losses as $x_1 \leq \ldots \leq x_l \leq 0 \leq x_{l+1} \leq \ldots \leq x_K$. Then, the CPT-value is defined as

$$\begin{aligned}\mathbb{C}(X) =& u^-(x_1) \cdot w^-(p_1) + \sum_{i=2}^{l} u^-(x_i) \Big(w^-(\sum_{j=1}^{i} p_j) - w^-(\sum_{j=1}^{i-1} p_j) \Big) \\ &+ \sum_{i=l+1}^{K-1} u^+(x_i) \Big(w^+(\sum_{j=i}^{K} p_j) - w^+(\sum_{j=i+1}^{K} p_j) \Big) + u^+(x_K) \cdot w^+(p_K),\end{aligned}$$

where u^+, u^- are utility functions and w^+, w^- are weight functions corresponding to gains and losses, respectively. The utility functions u^+ and u^- are non-decreasing, while the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. Unlike PT, the CPT-value does not violate stochastic dominance. In the aforementioned example, increasing $w^-(0.05)$ and $w^+(0.05)$ does not impact outcomes other than those on the extreme, i.e., -10 and 180, respectively. For instance, the weight for outcome 100 would be $w^{+}(0.45) - w^{+}(0.40)$. Thus, CPT formalizes the intuitive notion that humans are sensitive to extreme outcomes and relatively insensitive to intermediate ones.

The main assumption supporting CPT (and Prospect Theory) is that people tend to evaluate possible outcomes usually relative to a certain reference point rather than to the raw observation. Moreover, CPT assumes that people have different risk attitudes towards gains and losses and are generally more concerned about potential losses than potential gains (loss aversion). Finally, people tend to overweight extreme, but unlikely events, but underweight “average" events.

1.6 Outline of Thesis

Chap. 2 is devoted to designing a CPT-value based stochastic optimization framework. We lay out a sample-efficient estimation scheme of the CPT-value of a given policy, and we also propose a few nonparametric policy-optimization algorithms concerning the CPT-value based performance measure. Our CPT-value estimator is proved to converge asymptotically under relatively mild conditions. We will also present sample complexity properties of our estimation scheme. All of the policy-optimization algorithms designed in this chapter are theoretically guaranteed to converge to a local optimal point of the corresponding performance function.

Chap. 3 puts forward a dynamic programming structure driven by the CPT-based transitional measure. We prove that the nested-structure of the CPT-dynamic programming has a unique optimal solution under some reasonable assumptions. Moreover, the optimal solution can be found by value and policy iteration algorithms.

In Chap. 4, we conduct numerical experiments, which are designed for the following purposes: to we test the asymptotic convergence properties and sample complexity properties of our proposed CPT-estimator, and to investigate the difference between CPT-based decision making and traditional expected-utility-based decision making.

Chapter 2

Stochastic optimization in a cumulative prospect theory framework

2.1 Overview

In this chapter, we bring CPT to a stochastic optimization framework and propose algorithms for both estimation and optimization of CPT-value objectives. We propose an empirical distribution function-based scheme to estimate the CPT-value and then use this scheme in the inner loop of a CPT-value optimization procedure. We propose both gradient-based as well as gradient-free CPT-value optimization algorithms that are based on two well-known simulation optimization ideas: simultaneous perturbation stochastic approximation (SPSA) and model-based parameter search (MPS), respectively. We provide theoretical convergence guarantees for all the proposed algorithms and also illustrate the potential of CPT-based criteria in a traffic signal control application with numerical experiments in Chapter 4. The content of this chapter is based on joint work with L.A. Prashanth et al. ([44], [62]).

2.2 Contribution of the chapter

In this chapter we consider *stochastic optimization problems* where a designer optimizes the system to produce outcomes that are maximally aligned with the preferences of one or possibly multiple humans. As a running example, consider traffic optimization where the goal is to maximize travelers' satisfaction, a challenging problem many may agree is still inadequately addressed today, at least in big cities. In this example, the outcomes ("return") are travel times, or delays. To capture human preferences, the outcomes are mapped to a single numerical quantity.

To the best of our knowledge, we are the first to incorporate CPT into an online stochastic optimization framework. Although on the surface the combination may seem straightforward, in fact there are many challenges that arise from trying to optimize a CPT objective in the stochastic optimization framework, as we will soon see. In this short summary, we outline these challenges as well as our approach to addressing them.

The first challenge stems from the fact that the CPT-value assigned to a random variable is defined through a nonlinear transformation of the cumulative distribution function associated with the underlying random variable (see Section 2.4 for the definitions). In the case of the classic value function, which is an expectation, a simple sample mean can be used for estimation, facilitating the use of temporal difference type algorithms. On the other hand, CPT-value involves a distribution that is distorted using nonlinear weight functions and hence, requires that the *entire* distribution be estimated. Therefore, even the problem of estimating the CPT-value given a random sample is challenging.

In this chapter, we consider a natural quantile-based estimator and analyze its behavior. Under certain technical assumptions, we prove consistency and give sample complexity bounds, the latter based on the Dvoretzky-Kiefer-Wolfowitz (DKW) theorem [82, Chapter 2]. As an example, we show that the sample complexity to estimate the CPT-value for Lipschitz probability distortion weight functions is $O\left(\frac{1}{\epsilon^2}\right)$, for a given accuracy ϵ. This sample complexity coincides with the canonical rate for Monte Carlo-type schemes and is thus unimprovable. Since weight functions that fit well to human preferences are only Hölder continuous, we also consider this case and find that (unsurprisingly) the sample complexity degrades to $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ where $\alpha \in (0,1]$ is the weight function's Hölder exponent.

Our results on estimating CPT-values form the basis of the algorithms that we propose to maximize CPT-values based on interacting either with a real environment or with a simulator. We set up this problem as an instance of policy search, and we consider a smooth parameterization of the CPT-value and propose four algorithms for updating the CPT-value parameter. The first algorithm is a stochastic gradient scheme that uses two-point randomized gradient estimators, borrowed from simultaneous perturbation stochastic approximation (SPSA) [72]. The second algorithm is a modification of the first algorithm in that it uses the two-point randomized perturbation idea to estimate both the gradient and Hessian of the CPT-value at the given point, and it updates the parameter through a Newton-Raphson strategy. The third algorithm provides an unbiased gradient estimator built from infinitesimal perturbation analysis (IPA), which is been comprehensively studied in [39]. And lastly, the chapter provides a non-parametric algorithm for maximizing CPT-value based on Model Reference Adaptive Search (MRAS) [20]. Even though incorporating CPT-based criteria incurs extra sample complexity in estimation as compared to that of the classic sample mean estimator for expected value, the optimization schemes based either on SPSA or model-based parameter search [20] that we propose converge at the same rate as that of their expected value counterparts.

2.3 Related Work

Various risk measures have been proposed in the literature, e.g., mean-variance tradeoff [55], exponential utility [2], value at risk (VaR) and conditional value at risk (CVaR) [65]. A large body of literature involves risk-sensitive optimization in the context of Markov decision processes (MDPs). The stochastic optimization context of this chapter translates to a risk-sensitive reinforcement learning (RL) problem, and it has been observed in earlier works that risk-sensitive RL is generally hard to solve. For instance, in [71], [32] and [54], the authors provide NP-hardness results for finding a globally variance-optimal policy in discounted and average reward MDPs. Solving CVaR constrained MDPs is equally complicated (cf. [15, 59]).

In an abstract MDP setting, a CPT-based risk measure has been proposed in [52]. As compared to [52], *(i)* we do not assume a nested structure for the CPT-value, and this implies the lack of a Bellman equation for our CPT measure; *(ii)* we do not assume model information, i.e., we operate in a more general stochastic optimization setting; *(iii)* we develop both estimation and optimization algorithms with convergence guarantees for the CPT-value function. More recently, the authors in [36] incorporate CPT-based criteria into a multi-armed bandit setting, while employing the estimation scheme that we proposed in the shorter version of this [62].

The rest of the chapter is organized as follows: In Section 2.4, we define the notion of CPT-value for a general random variable. In Section 2.5, we describe the empirical distribution-based scheme for estimating the CPT-value of any random variable, and we also prove the convergence of our estimation scheme. In Section 2.6, we present both gradient-based and non-parametric algorithms for optimizing the CPT-value. We provide proofs of convergence for all the proposed algorithms in Section 2.6 as well. Finally, in Section 2.7 we provide concluding remarks.

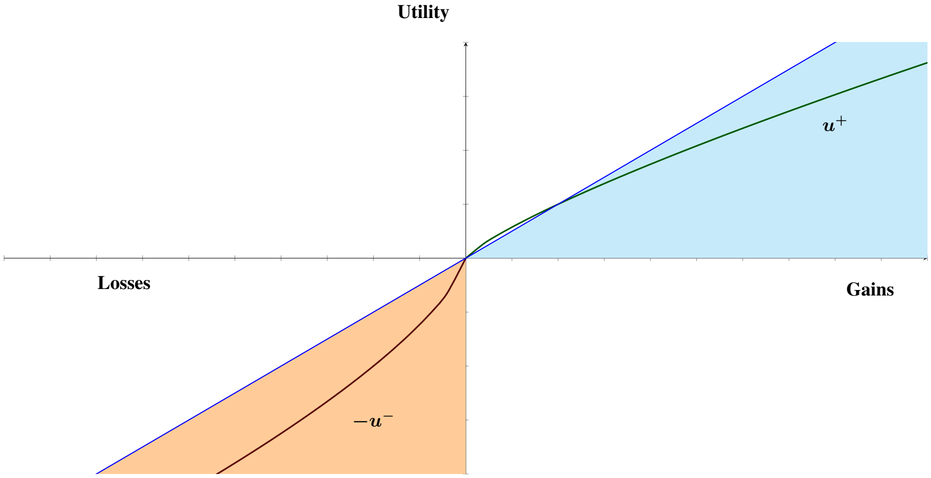


Figure 2.1: An example of a utility function. A reference point on the x axis serves as the point of separating gains and losses. For losses, the disutility $-u^-$ is typically convex and for gains, the utility u^+ is typically concave; both functions are non-decreasing and take the value of zero at the reference point.

2.4 CPT-functional

Without loss of generality, we choose 0 as the reference throughout this chapter. Given random variable X, the functional, denoted by $\mathbb{C}$, depends on function pairs $u = (u^+, u^-)$ and $w = (w^+, w^-)$. As illustrated in Figure 2.1, $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ are continuous, with $u^+(x) = 0$ when $x \leq 0$ and non-decreasing otherwise, and with $u^-(x) = 0$ when $x \geq 0$ and non-increasing otherwise. The functions $w^+, w^- : [0,1] \to [0,1]$, as shown in Figure 2.2, are continuous, non-decreasing and satisfy $w^+(0) = w^-(0) = 0$ and

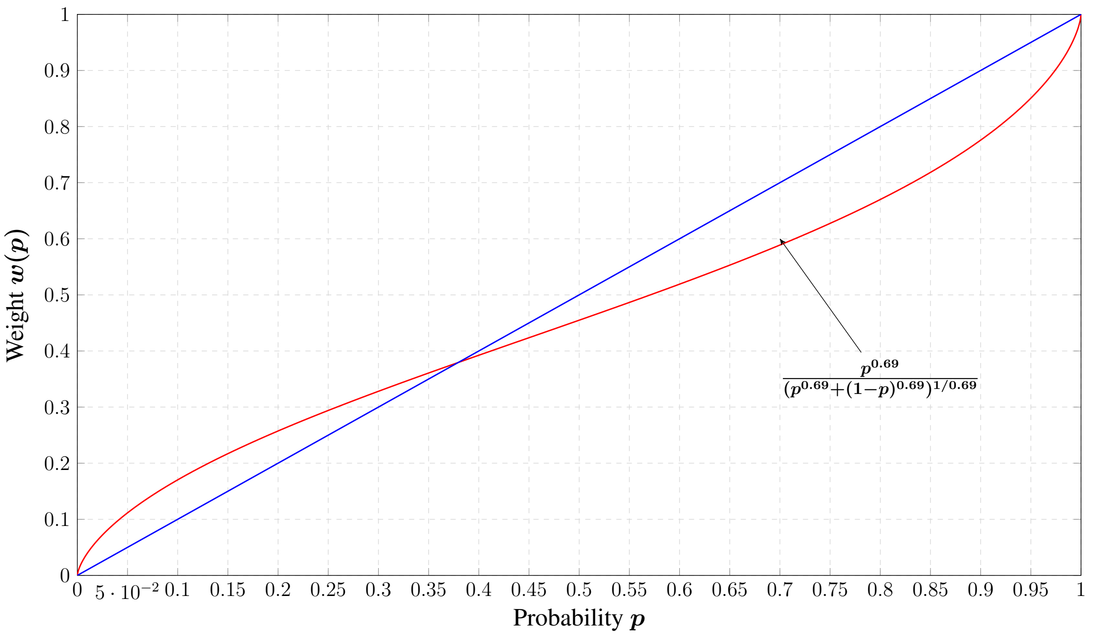


Figure 2.2: An example of a weight function. A typical CPT weight function inflates small, and deflates large probabilities, capturing the tendency of humans when facing with decisions of uncertain outcomes.

$w^+(1) = w^-(1) = 1$.

According to [79], if X is a continuous random variable, its CPT-functional is defined as

$$\begin{aligned} \mathbb{C}(X) = &\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \\ &- \int_0^\infty w^-\left(\mathbb{P}\left(u^-(X) > z\right)\right) dz \,. \end{aligned} \tag{2.1}$$

Consider the case when w^+ and w^- are identity functions, $u^+(x) = x$ for $x \geq 0$ and 0 otherwise, and $u^-(x) = -x$ for $x \leq 0$ and 0 otherwise. Then, letting $(a)^+ = \max(a, 0)$, $(a)^- = \max(-a, 0)$, we have $\mathbb{C}(X) = \int_0^\infty \mathbb{P}\left(X > z\right) dz - \int_0^\infty \mathbb{P}\left(-X > z\right) dz = \mathbb{E}\left[(X)^+\right] - \mathbb{E}\left[(X)^-\right]$, showing the connection to expectations.

In the definition, u^+ and u^- are utility functions corresponding to gains ($X \geq 0$)

and losses ($X \leq 0$), respectively, where zero is chosen the "reference point" to separate gains and losses. Handling losses and gains separately is a salient feature of CPT, and this addresses the tendency of humans to play safe with gains and take risks with losses. To illustrate this tendency, consider a scenario where one can either earn \$$500$ with probability (w.p.) 1 or earn \$$1000$ w.p. 0.5 and nothing otherwise. The human tendency is to choose the former option of a certain gain. If we flip the situation, i.e., a certain loss of \$$500$ or a loss of \$$1000$ w.p. 0.5, then humans choose the latter option. This distinction of playing safe with gains and taking risks with losses is captured by a concave gain-utility u^+ and a convex disutility $-u^-$, as illustrated in Figure 2.1.

The functions w^+, w^-, called the weight functions, capture the idea that humans deflate high-probabilities and inflate low-probabilities. For example, humans usually choose a stock that gives a large reward, e.g., one million dollars w.p. $1/10^6$ over one that gives \$1 w.p. 1 and the reverse when signs are flipped. Thus the value seen by a human subject is non-linear in the underlying probabilities – an observation backed by strong empirical evidence [4]. As illustrated with $w = w^+ = w^-$ in Fig 2.2, the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. [78] recommends $w(p) = \frac{p^\eta}{(p^\eta + (1-p)^\eta)^{1/\eta}}$, while [63] recommends $w(p) = \exp(-(-\ln p)^\eta)$, with $0 < \eta < 1$. In both cases, the weight function has an inverted-s shape.

Illustrative application example: Stochastic Shortest Path

To present a running example where a CPT-functional can be applied, we consider a stochastic shortest path (SSP) problem with states $\mathcal{S} = \{0, \ldots, \}$, where 0 is a special reward-free absorbing state. Each state $s \in \mathcal{S}$ is associated with an action space $\mathcal{A}(s)$ that an agent can choose from. Once an action $a \in \mathcal{A}(s)$ is taken, the pair (s, a) will result a reward $r(s, a)$ and a transitional probability distribution $\mathbb{P}(\cdot|s, a)$ over the state space $\mathcal{S}$. The probability distribution $\mathbb{P}(\cdot|s, a)$ thereby governs the move of the agent for the next step. A randomized policy π is a function that maps any state $s \in \mathcal{S}$ onto a probability distribution over the actions $\mathcal{A}(s)$ in state s. As is standard in policy gradient algorithms, we parameterize π and assume it is continuously differentiable in its parameter $\theta \in \mathbb{R}^d$. An *episode* is a simulated sample path of the shortest path problem based on policy θ and transitional distributions $\mathbb{P}(\cdot|s, a), \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. It starts in state $s_0 \in \mathcal{S}$, visits $\{s_1, \ldots, s_{\tau-1}\}$ before ending in the absorbing state 0, where τ is the first passage time to state 0. Let $D^\theta(s_0)$ be a random variable (r.v) that denote the total reward from an episode start with s_0, defined by

$$D^\theta(s_0) = \sum_{m=0}^{\tau-1} r(s_m, a_m),$$

where the actions a_m are chosen using policy θ and $r(s_m, a_m)$ is the single-stage reward in state $s_m \in \mathcal{S}$ when action $a_m \in \mathcal{A}(s_m)$ is chosen.

Instead of the traditional RL objective for an SSP of maximizing the expected value $\mathbb{E}(D^\theta(s_0))$, we adopt the CPT approach and aim to solve the following problem:

$$\max_{\theta \in \Theta} \mathbb{C}(D^\theta(s_0)),$$

where Θ is the set of admissible policies that are *proper*[1] and the CPT-value function $\mathbb{C}(D^\theta(s_0))$ is defined as

$$\begin{aligned}\mathbb{C}(D^\theta(s_0)) = &\int_0^\infty w^+(P(u^+(D^\theta(s_0))) > z)dz \\ &- \int_0^\infty w^-(P(u^-(D^\theta(s_0))) > z)dz.\end{aligned} \tag{2.2}$$

Remark 1. *(Generalization) As noted earlier, the CPT-value is a generalization of mathematical expectation. It is also possible to get* (2.1) *to coincide with risk measures (e.g. VaR and CVaR) by appropriate choice of weight functions.*

Remark 2. *(Sensitivity) Traditional EU-based approaches are sensitive to modeling errors as illustrated in the following example: Suppose stock $\mathcal{A}$ gains* $\$10000$ *w.p* 0.001 *and loses nothing w.p.* 0.999*, while stock $\mathcal{B}$ surely gains* 11*. With the classic value function objective, it is optimal to invest in stock $\mathcal{B}$ as it returns* 11*, while $\mathcal{A}$ returns* 10 *in expectation (assuming utility function to be the identity map). Now, if the gain probability for stock $\mathcal{A}$ was* 0.002*, then it is no longer optimal to invest in stock $\mathcal{B}$ and investing in stock A is optimal. Notice that a very slight change in the underlying probabilities resulted in a big difference in the investment strategy and a similar observation carries over to a multi-stage scenario (see the house buying example in Chapter 4). A randomized policy that* 50% *in stock $\mathcal{A}$ and the rest in a risk-free asset is less sensitive to the error in under-estimating the loss probability.*

Using CPT makes sense because it inflates low probabilities and thus can account for modeling errors, especially considering that model information is unavailable in prac-

[1] A policy θ is proper if 0 is recurrent and all other states are transient for the Markov chain underlying θ. It is standard to assume that policies are proper in an SSP setting - cf. [7].

tice. Note also that in MDPs with expected utility objective, there exists a deterministic policy that is optimal. However, with CPT-value objective, the optimal policy is not necessarily deterministic - See also the organ transplant example on pp. 75-81 of [52].

2.5 CPT-value estimation

For a given random variable X, we devise a scheme for estimating the CPT-value $\mathbb{C}(X)$ given only samples from the distribution of X. Meanwhile, we show that, under a set of reasonable assumptions on the random variable X and probability weighting functions, our estimator (presented next) converges almost surely. Before diving into the details of CPT-value estimation, let us discuss the conditions necessary for the CPT-value to be well-defined. Observe that the first integral in (2.1), i.e., $\int_0^{+\infty} w^+ \left(\mathbb{P}\left(u^+(X) > z\right)\right) dz$ may diverge even if the first moment of random variable $u^+(X)$ is finite. For example, suppose U has the tail distribution function $\mathbb{P}\left(U > z\right) = \frac{1}{z^2}, z \in [1, +\infty)$, and $w^+(z)$ takes the form $w(z) = z^{\frac{1}{3}}$. Then, the first integral in (2.1), i.e., $\int_1^{+\infty} z^{-\frac{2}{3}}\, dz$ does not even exist. A similar argument applies to the second integral in (2.1). To overcome the integrability issues, we assume that the weight functions w^+, w^- satisfy one of the following assumptions for continuous valued r.v.s:

Assumption 1. *The weight functions $w^{\pm}$ are Hölder continuous with common order α and constant H, i.e., $\sup_{x\neq y} \frac{|w^{\pm}(x)-w^{\pm}(y)|}{|x-y|^{\alpha}} \leq H, \forall x, y \in [0,1]$. Further, there exists $\gamma \leq \alpha$ such that (s.t.) $\int_0^{+\infty} \mathbb{P}^{\gamma}(u^+(X) > z)dz < +\infty$ and $\int_0^{+\infty} \mathbb{P}^{\gamma}(u^-(X) > z)dz < +\infty$, where $\mathbb{P}^{\gamma}(\cdot) = (\mathbb{P}(\cdot))^{\gamma}$.*

Assumption 2. *The weight functions w^+, w^- are Lipschitz with common constant L, and*

$u^+(X)$ and $u^-(X)$ both have bounded first moments.

The first property of the CPT-value is claimed and proved in the following theorem:

Theorem 1. *Under assumption 1 or 2, the CPT-value $\mathbb{C}(X)$ as defined by* (2.1) *is finite.*

Proof. Hölder continuity of w^+ and $w^+(0) = 0$ imply that

$$\int_0^\infty w^+ \left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \leq H \int_0^\infty \mathbb{P}^\alpha \left(u^+(X) > z\right) dz$$
$$\leq H \int_0^\infty \mathbb{P}^\gamma \left(u^+(X) > z\right) dz < \infty.$$

The second inequality is valid since $\mathbb{P}\left(u^+(X) > z\right) \leq 1.$ The claim follows for the first integral in (2.1), and the finiteness of the second integral in (2.1) can be argued in an analogous fashion. □

Assumption 2, even though it implies assumption 1, is a useful special case because it does away with additional assumptions required to establish asymptotic consistency under assumption 1. For the theoretical results, we also require the following assumption on the utility functions:

Assumption 3. *The utility functions u^+ and $-u^-$ are continuous and non-decreasing on their support $\mathbb{R}^+$ and $\mathbb{R}^-$, respectively.*

Finally, we also analyze the setting where X is a discrete valued r.v. Such a setting is common in practice and carries the additional advantage that, under a local Lipschitz assumption on the distribution of X, one gets better sample complexity as compared to those under assumptions 1 and 2.

2.5.1 CPT-value estimation using quantiles

Let ξ_k^+ and ξ_k^- denote the kth quantiles of the r.v.s $u^+(X)$ and $u^-(X)$, respectively. Then, it can be seen that (see Theorem 1 in Section 2.5.2)

$$\begin{aligned}&\lim_{n\to\infty} \sum_{i=1}^{n} \xi^+_{\frac{i}{n}} \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\\ &= \int_0^{+\infty} w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz. \end{aligned} \tag{2.3}$$

A similar claim holds with $u^-(X)$, ξ_k^-, w^- in place of $u^+(X)$, ξ_α^+, w^+, respectively.

However, we do not know the distribution of $u^+(X)$ or $u^-(X)$ and hence, we next present a procedure that uses order statistics for estimating quantiles and this in turn assists estimation of the CPT-value along the lines of (2.3). The estimation scheme is presented in Algorithm 1.

Algorithm 1 CPT-value estimation

```
Input: sample $X_1, \dots, X_n$ from the distribution of $X$.
Arrange the samples in ascending order and label them as follows: $X_{[1]}, X_{[2]}, \dots, X_{[n]}$.
Let
    $\overline{\mathbb{C}}_n^+ := \sum_{i=1}^{n} u^+(X_{[i]}) \left( w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right),$
    $\overline{\mathbb{C}}_n^- := \sum_{i=1}^{n} u^-(X_{[i]}) \left( w^-\left(\frac{i}{n}\right) - w^-\left(\frac{i-1}{n}\right)\right).$
Return $\overline{\mathbb{C}}_n = \overline{\mathbb{C}}_n^+ - \overline{\mathbb{C}}_n^-$.
```

Consider the special case when $w^+(p) = w^-(p) = p$ and both u^+ and $(-u^-)$, when restricted to the positive (respectively, negative) half line, are the identity functions.

In this case, the CPT-value estimator $\overline{\mathbb{C}}_n$ coincides with the sample mean estimator for regular expectation.

Notice that the CPT estimator $\overline{\mathbb{C}}_n$ in Algorithm 1 can be written equivalently as follows:

$$\overline{\mathbb{C}}_n = \int_0^{\infty} w^+ \left(1 - \hat{F}_n^+(x)\right) dx - \int_0^{\infty} w^- \left(1 - \hat{F}_n^-(x)\right) dx. \tag{2.4}$$

The above relation holds because

$$\begin{aligned}
&\sum_{i=1}^{n} u^+\left(X_{[i]}\right)\left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\\
&= \sum_{i=1}^{n-1} w^+\left(\frac{n-i}{n}\right)\left(u^+\left(X_{[i+1]}\right) - u^+\left(X_{[i]}\right)\right) + u^+(X_{[1]})\\
&= \int_0^{\infty} w^+\left(1 - \hat{F}_n^+(x)\right) dx, \text{ and}\\
&\sum_{i=1}^{n} u^-\left(X_{[i]}\right)\left(w^-\left(\frac{i}{n}\right) - w^-\left(\frac{i-1}{n}\right)\right)\\
&= \int_0^{\infty} w^-\left(1 - \hat{F}_n^-(x)\right) dx,
\end{aligned}$$

where $\hat{F}_n^+(x)$ and $\hat{F}_n^-(x)$ are the empirical distributions of $u^+(X)$ and $u^-(X)$, respectively.

2.5.2 Results for Hölder and Lipschitz continuous weights

Theorem 2. ***(Asymptotic consistency)*** *Let $w^{\pm}$ satisfy the Hölder continuous assumption 1 with $\alpha > \frac{1}{2}$ and let assumption 3 hold. If $F^+(\cdot)$ and $F^-(\cdot)$, the respective distribution functions of $u^+(X)$ and $u^-(X)$, satisfy the property that there exist constants L^+ and L^- such that*

$$|F^+(x) - F^+(y)| \geq L^+|x-y|, \ \forall x, y \in \mathcal{U}^+ \subset \mathbb{R}$$

and

$$|F^-(x) - F^-(y)| \geq L^-|x-y|, \ \forall x, y \in \mathcal{U}^- \subset \mathbb{R},$$

with $\mathcal{U}^+$ and $\mathcal{U}^-$ the connected and compact support of $u^+(X)$ and $u^-(X)$, and if the random variables $u^+(X), u^-(X)$ satisfy

$$\lim_{n\to\infty} \frac{u^+(X_{[n]})}{n^\alpha} \to 0 \textit{ and } \lim_{n\to\infty} \frac{u^-(X_{[n]})}{n^\alpha} \to 0 \textit{ a.s.},$$

where α is the Hölder exponent for $w^\pm$ defined in assumption 1, then we have

$$\overline{\mathbb{C}}_n \to \mathbb{C}(X) \textit{ a.s. as } n \to \infty, \tag{2.5}$$

where $\overline{\mathbb{C}}_n$ is as defined in Algorithm 1 and $\mathbb{C}(X)$ as in (2.1).

We now state and prove a lemma that will be used in the proof of Theorem 2.

Lemma 1. *Let $\xi^+_{\frac{i}{n}}$ and $\xi^-_{\frac{i}{n}}$ denote the $\frac{i}{n}$th quantile of $u^+(X)$ and $u^-(X)$, respectively. If assumption 1 holds, then we have*

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right)\right) \\
&\quad = \int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz < \infty, \qquad (2.6)\\
&\lim_{n\to\infty} \sum_{i=1}^{n-1} \xi^-_{\frac{i}{n}} \left(w^-\left(\frac{i}{n}\right) - w^-\left(\frac{i-1}{n}\right)\right) \\
&\quad = \int_0^\infty w^-\left(\mathbb{P}\left(u^-(X) > z\right)\right) dz < \infty. \qquad (2.7)
\end{aligned}$$

Proof. We will focus on proving equation (2.6). For all $z \in (0, +\infty)$, the following convergence claim holds w.p.1:

$$\sum_{i=1}^{n-1} w^+\left(\frac{i}{n}\right) I_{\left[\xi^+_{\frac{n-i-1}{n}}, \xi^+_{\frac{n-i}{n}}\right]}(z) \xrightarrow{n\to\infty} w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right). \tag{2.8}$$

To infer the above claim, observe that since $u^+(X)$ ranges in $(0, +\infty), \forall z$, there exists i such that $z \in [\xi^+_{\frac{n-i-1}{n}}, \xi^+_{\frac{n-i}{n}}]$, which implies that

$$w^+\left(\mathbb{P}\left(u^+\left(X\right) \geq z\right)\right) \in \left[w^+\left(\frac{i}{n}\right), w^+\left(\frac{i+1}{n}\right)\right].$$

Hence, we have

$$\left|\sum_{j=1}^{n-1} w^+\left(\frac{j}{n}\right) I_{\left[\xi^+_{\frac{n-j-1}{n}}, \xi^+_{\frac{n-j}{n}}\right]}(z) - w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right)\right|$$
$$\leq \left|w^+\left(\frac{i}{n}\right) - w^+\left(\frac{i+1}{n}\right)\right|$$

Since w^+ is Hölder continuous, we have

$$\left|w^+\left(\frac{i}{n}\right) - w^+\left(\frac{i+1}{n}\right)\right| \xrightarrow{n\to\infty} 0,$$

and the claim in (2.8) follows.

Further, for all $z \in [0, \infty)$,

$$\sum_{j=1}^{n-1} w^+\left(\frac{j}{n}\right) I_{\left[\xi^+_{\frac{n-j-1}{n}}, \xi^+_{\frac{n-j}{n}}\right]}(z) \leq w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right). \tag{2.9}$$

The integral of the LHS of (2.8) can be simplified as follows:

$$\int_0^\infty \sum_{j=0}^{n} w^+\left(\frac{j}{n}\right) I_{\left[\xi^+_{\frac{n-j-1}{n}}, \xi^+_{\frac{n-j}{n}}\right]}(z)dz$$
$$= \sum_{j=0}^{n-1} w^+\left(\frac{j}{n}\right)\left(\xi^+_{\frac{n-j}{n}} - \xi^+_{\frac{n-j-1}{n}}\right)$$
$$= \sum_{j=0}^{n-1} \xi^+_{\frac{j}{n}}\left(w^+\left(\frac{n-j}{n}\right) - w^+\left(\frac{n-j-1}{n}\right)\right). \tag{2.10}$$

Now, the main claim in (2.6) can be inferred from (2.8), (2.9) and (2.10) in conjunction with the dominated convergence theorem.

The second part of (2.6) follows in a similar fashion. □

Before proving Theorem 2, we need following result of Hoeffding ([40]).

Lemma 2. ***Hoeffding***. *Let* $U_1, \ldots, U_n$ *be independent random variables satisfying* $P(a \leq U_i \leq b) = 1, \forall i$. *Then, for* $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} U_i - \sum_{i=1}^{n} E\left(U_i\right) \geq nt\right) \leq e^{-2nt^2/(b-a)^2}.$$

Proof. (***Theorem 2***)

Without loss of generality, assume that w^+ and w^- are both Hölder continuous with common order α and common constant $H = 1$. We prove the claim for the first integral in the CPT-value estimator $\overline{\mathbb{C}}_n$ in Algorithm 1, i.e., we show that

$$\begin{aligned} &\lim_{n\to\infty} \sum_{i=1}^{n} u^+\left(X_{[i]}\right)\left(w^+\left(\frac{n-i+1}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right) \\ &= \int_0^\infty w^+\left(P\left(u^+(X) > z\right)\right) dz \text{ a.s.} \end{aligned} \tag{2.11}$$

The main part of the proof is focused on finding an upper bound for the probability

$$\begin{aligned} \mathbb{P}\Bigg(&\Bigg|\sum_{i=1}^{n-1} u^+\left(X_{[i]}\right)\left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right)\right) \\ &- \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}}\left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right)\right)\Bigg| > \epsilon\Bigg). \end{aligned}$$

Observe the fact that

$$\begin{aligned} &\sum_{i=1}^{n} u^+\left(X_{[i]}\right)\left(w^+\left(\frac{n-i+1}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right) \\ &- \sum_{i=1}^{n-1} u^+\left(X_{[i]}\right)\left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right)\right) \\ &= \sum_{i=1}^{n} \left(u^+\left(X_{[i]}\right) - u^+\left(X_{[i-1]}\right)\right) w^+\left(\frac{n+1-i}{n}\right) \\ &- \sum_{i=1}^{n} \left(u^+\left(X_{[i]}\right) - u^+\left(X_{[i-1]}\right)\right) w^+\left(\frac{n-i}{n}\right) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left(u^{+}\left(X_{[i]}\right) - u^{+}\left(X_{[i-1]}\right)\right) \\
&\qquad \times \left(w^{+}\left(\frac{n+1-i}{n}\right) - w^{+}\left(\frac{n-i}{n}\right)\right) \\
&\leq u^{+}\left(X_{[n]}\right) \times \frac{1}{n^{\alpha}},
\end{aligned}
$$

by defining $u^{+}(X_{[0]}) = 0$. Notice that the term $\frac{u^{+}\left(X_{[n]}\right)}{n^{\alpha}}$ converges to 0 under the statement of the theorem. Hence, for the asymptotic convergence of estimator, thanks to Lemma 1, it suffices to show that

$$
\begin{aligned}
\lim_{n\to\infty} &\left| \sum_{i=1}^{n-1} u^{+}\left(X_{[i]}\right) \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right. \\
&\left. - \sum_{i=1}^{n-1} \xi^{+}_{\frac{i}{n}} \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right| = 0 \ \ a.s.
\end{aligned}
$$

Observe that, for any given $\epsilon > 0$, we have

$$
\begin{aligned}
&\mathbb{P}\left(\left| \sum_{i=1}^{n-1} u^{+}\left(X_{[i]}\right) \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right.\right. \\
&\qquad \left.\left. - \sum_{i=1}^{n-1} \xi^{+}_{\frac{i}{n}} \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right| > \epsilon \right) \\
&\leq \mathbb{P}\left(\bigcup_{i=1}^{n-1} \left\{ \left| u^{+}\left(X_{[i]}\right) \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right.\right.\right. \\
&\qquad \left.\left.\left. - \xi^{+}_{\frac{i}{n}} \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right| > \frac{\epsilon}{n-1} \right\} \right) \\
&\leq \sum_{i=1}^{n-1} \mathbb{P}\left(\left| u^{+}\left(X_{[i]}\right) \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right.\right. \\
&\qquad \left.\left. - \xi^{+}_{\frac{i}{n}} \left(w^{+}\left(\frac{n+1-i}{n}\right) - w^{+}\left(\frac{n-i}{n}\right)\right) \right| > \frac{\epsilon}{n-1} \right) \\
&\leq \sum_{i=1}^{n-1} \mathbb{P}\left(\left| \left(u^{+}\left(X_{[i]}\right) - \xi^{+}_{\frac{i}{n}}\right) \right.\right. \\
&\qquad \left.\left. \times \left(w^{+}\left(\frac{n-i}{n}\right) - w^{+}\left(\frac{n-i-1}{n}\right)\right) \right| > \frac{\epsilon}{n-1} \right) \\
&\leq \sum_{i=1}^{n-1} \mathbb{P}\left(\left| \left(u^{+}\left(X_{[i]}\right) - \xi^{+}_{\frac{i}{n}}\right) \left(\frac{1}{n}\right)^{\alpha} \right| > \frac{\epsilon}{n-1} \right) \qquad (2.12)
\end{aligned}
$$

$$\leq \sum_{i=1}^{n-1} \mathbb{P}\left(\left|\left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}}\right)\left(\frac{1}{n}\right)^{\alpha}\right| > \frac{\epsilon}{n}\right)$$
$$\leq \sum_{i=1}^{n-1} \mathbb{P}\left(\left|\left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}}\right)\right| > \frac{\epsilon}{n^{1-\alpha}}\right). \tag{2.13}$$

In the above, (2.12) follows from the fact that w^+ is Hölder with constant 1.

Now we find an upper bound for the probability of a single term in the sum above, i.e.,

$$\mathbb{P}\left(\left|u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}}\right| > \frac{\epsilon}{n^{(1-\alpha)}}\right) = \mathbb{P}\left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}\right)$$
$$+ \mathbb{P}\left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{n^{(1-\alpha)}}\right).$$

We focus on the first term above, and

$$\text{let } W_j = I_{\left(u^+(X_j) > \xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)}, j = 1, \ldots, n.$$

Using the fact that a probability distribution function is non-decreasing, we obtain

$$\mathbb{P}\left(u^+(X_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}\right) = \mathbb{P}\left(\sum_{j=1}^{n} W_j > n - i\right)$$
$$= \mathbb{P}\left(\sum_{j=1}^{n} W_j > n\left(1 - \frac{i}{n}\right)\right)$$
$$= \mathbb{P}\left(\sum_{j=1}^{n} W_j - n\left[1 - F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)\right]\right.$$
$$\left. > n\left[F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right) - \frac{i}{n}\right]\right).$$

Using the fact that $\mathbb{E}W_j = 1 - F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)$ in conjunction with Hoeffding's inequality, we obtain

$$\mathbb{P}\left(\sum_{i=1}^{n} W_j - n\left[1 - F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)\right]\right.$$

$$> n\left[F^{+}\left(\xi^{+}_{\frac{i}{n}}+\frac{\epsilon}{n^{(1-\alpha)}}\right)-\frac{i}{n}\right]\right)\leq e^{-2n(\delta_i^{'})^2},$$

where $\delta_i^{'} = F^{+}\left(\xi^{+}_{\frac{i}{n}}+\frac{\epsilon}{n^{(1-\alpha)}}\right)-\frac{i}{n}$. According to the conditions imposed on F^{+}, we have that $\delta_i^{'} \geq \frac{L^{+}\epsilon}{n^{(1-\alpha)}}$. Hence, we obtain

$$\begin{aligned}\mathbb{P}\left(u^{+}(X_{[i]})-\xi^{+}_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}\right) &\leq e^{-2n(\frac{L^{+}\epsilon}{n^{(1-\alpha)}})^2}\\ &= e^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}.\end{aligned} \tag{2.14}$$

In a similar fashion, one can show that

$$\mathbb{P}\left(u^{+}(X_{[i]})-\xi^{+}_{\frac{i}{n}} < -\frac{\epsilon}{n^{(1-\alpha)}}\right) \leq e^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}. \tag{2.15}$$

Combining (2.14) and (2.15), we obtain

$$\mathbb{P}\left(\left|u^{+}(X_{[i]})-\xi^{+}_{\frac{i}{n}}\right| < -\frac{\epsilon}{n^{(1-\alpha)}}\right) \leq 2e^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}.$$

Plugging the above in (2.13), we obtain

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^{n-1}u^{+}\left(X_{[i]}\right)\left(w^{+}\left(\frac{n-i}{n}\right)-w^{+}\left(\frac{n-i-1}{n}\right)\right)\right.\right.\\ &\qquad\left.\left.-\sum_{i=1}^{n-1}\xi^{+}_{\frac{i}{n}}\left(w^{+}\left(\frac{n-i}{n}\right)-w^{+}\left(\frac{n-i-1}{n}\right)\right)\right|>\epsilon\right)\\ &\leq 2(n-1)e^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}\leq 2ne^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}.\end{aligned} \tag{2.16}$$

Notice that $\sum_{n=1}^{\infty}2ne^{-2n^{2\alpha-1}(L^{+}\epsilon)^2} < \infty$ with $\alpha > \frac{1}{2}$ since the sequence $2ne^{-2n^{2\alpha-1}(L^{+}\epsilon)^2}$ will decrease faster than the sequence $\frac{1}{n^k}$ provided $k > 1$.

By applying the Borel-Cantelli lemma, $\forall\epsilon > 0$, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\sum_{i=1}^{n-1}u^{+}\left(X_{[i]}\right)\left(w^{+}\left(\frac{n-i}{n}\right)-w^{+}\left(\frac{n-i-1}{n}\right)\right)\right.\right.\\ &\qquad\left.\left.-\sum_{i=1}^{n-1}\xi^{+}_{\frac{i}{n}}\left(w^{+}\left(\frac{n-i}{n}\right)-w^{+}\left(\frac{n-i-1}{n}\right)\right)\right|>\epsilon, i.o.\right)\end{aligned}$$

$$= 0,$$

which implies (2.11).

The proof of $\overline{\mathbb{C}}_n^- \to \mathbb{C}^-(X)$ follows in a similar manner as above by replacing $u^+(X_{[i]})$ by $u^-(X_{[n-i]})$, after observing that u^- is decreasing, which in turn implies that $u^-(X_{[n-i]})$ is an estimate of the quantile $\xi^-_{\frac{i}{n}}$. □

Under an additional assumption on the utility functions, our next result shows that $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ number of samples are sufficient to get a high-probability estimate of the CPT-value that is ϵ-accurate. Before the result is presented, we recall the class of sub-Gaussian distributions:

Definition 1. ***(Sub-Gaussian distribution)*** *Formally, the probability distribution of a random variable X is called sub-Gaussian if there are positive constants C, v such that for every $t > 0$,*

$$\mathbb{P}\left(|X| > t\right) \leq Ce^{-vt^2}.$$

Theorem 3. ***(Sample complexity.)*** *If assumptions 1 and 3 hold, and also that the utilities $u^+(X)$ and $u^-(X)$ are bounded by a constant M. Then, $\forall \epsilon > 0$, we have*

$$\mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \geq \epsilon\right) \leq 2e^{-2n\left(\frac{\epsilon}{HM}\right)^{\frac{2}{\alpha}}}. \tag{2.17}$$

Instead, if the utilities functions $u^+(X)$ and $u^-(X)$ are sub-Gaussian defined in Definition 1 with respective constant $C = t = 1$, then $\forall \epsilon > 0$ and $n \geq \left(\frac{1}{\alpha}\ln 4H - \ln \alpha\epsilon\right)^{\frac{\alpha+2}{2}}$, we have

$$\mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \geq \epsilon\right) \leq 2ne^{-n^{\frac{\alpha}{2+\alpha}}} + 2e^{-n^{\frac{\alpha}{2+\alpha}}\left(\frac{\epsilon}{2H}\right)^{\frac{2}{\alpha}}}. \tag{2.18}$$

Corollary 1. *If assumptions 1 and 3 hold, and if utilities $u^+(X)$ and $u^-(X)$ are bounded by M, then*

$$\mathbb{E}\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \leq \frac{(8HM)\,\Gamma\left(\alpha/2\right)}{n^{\alpha/2}}.$$

Instead, if the utilities are sub-Gaussian with respective constant $C = t = 1$, then

$$\mathbb{E}\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \leq \frac{4\Gamma\left(2\right)}{n^{\frac{2\alpha}{\alpha+2}}} + \frac{\Gamma\left(\alpha\right)2^{\alpha}\left(2H\right)^2}{n^{\frac{\alpha^2}{2+\alpha}}}.$$

For proving Theorem 3, we require the DKW inequality for empirical distributions, which is reviewed in the appendix B.

Lemma 3. ***(DKW inequality)***

Let F denote the cdf of r.v. U and $\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I_{[U_i \leq u]}$ denote the empirical distribution of U, with $U_1, \ldots, U_n$ sampled from F. Then, for any $\epsilon > 0$, we have

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

The DKW-inequality provides a finite-sample bound on the distance between the empirical distribution and the true distribution. With the DKW inequality, we layout the proof of Theorem 3 below:

Proof. (***Theorem 3***)

To prove (2.18) where the r.v.s $u^+(X)$ and $u^-(X)$ are sub-Gaussian with constants $t = C = 1$, we only need to address the w^+ part, and the w^- part follows in a similar fashion. Observe that for all $c > 0$, we have

$$\begin{aligned}\mathbb{P}\left(|\mathbb{C}_n - \mathbb{C}(X)| > \epsilon\right) \leq \mathbb{P}\left(u^+\left(X_{[n]}\right) \geq n^c\right)& \\ + \mathbb{P}\left(\{|\mathbb{C}_n - \mathbb{C}(X)| > \epsilon\} \bigcap \{u^+\left(X_{[n]}\right) < n^c\}\right)&.\end{aligned}$$

On the event $\{u^+\left(X_{[n]}\right) < n^c\}$, we have

$$
\begin{aligned}
&\left|\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^\infty w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| \\
=&\left|\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| \\
=&\left|\int_0^{n^c} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| \\
+&\left|\int_{n^c}^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt\right|,
\end{aligned}
$$

since $1 - \hat{F}_n^+(t) = 0, \forall t \geq n^c$. Notice that

$$
\begin{aligned}
&\int_{n^c}^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt \\
&\leq H \int_{n^c}^\infty \frac{t}{n^c} e^{-\alpha t^2} dt \\
&= \frac{H}{n^c}\frac{2}{\alpha} e^{-\alpha (n^c)^2} \\
&\leq \frac{\epsilon}{2} \quad \text{for } n \geq \left(\frac{1}{\alpha}\ln\frac{4H}{\alpha\epsilon}\right)^{\frac{1}{2c}},
\end{aligned}
$$

where the first inequality is obtained by Hölder continuous property from assumption 1 and sub-Gaussian property of the distribution $u^+(X)$. An easy observation indicates that if $\left|\int_{n^c}^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt\right| \leq \frac{\epsilon}{2}$, then a necessary condition for

$$
\left|\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| > \epsilon
$$

is

$$
\left|\int_0^{n^c} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| > \frac{\epsilon}{2}.
$$

Thus, we obtain that, for $n \geq \left(\frac{1}{\alpha}\ln\frac{4H}{\alpha\epsilon}\right)^{\frac{1}{2c}}$,

$$
\mathbb{P}\left(\{|\mathbb{C}_n - \mathbb{C}(X)| > \epsilon\} \bigcap \{u^+\left(X_{[n]}\right) < n^c\}\right) \leq
$$

$$\mathbb{P}\left(\left|\int_0^{n^c} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| > \frac{\epsilon}{2}\right).$$

Now, plugging in the DKW inequality, we have

$$\begin{aligned}
&\mathbb{P}\left(\left|\int_0^{n^c} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{n^c} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| > \frac{\epsilon}{2}\right) \\
&\le \mathbb{P}\left(Hn^c \sup_{t\in\mathbb{R}} \left|\mathbb{P}\left(u^+\left(X\right) < t\right) - \hat{F}_n^+\left(t\right)\right|^\alpha > \frac{\epsilon}{2}\right) \\
&\le 2e^{-2n\left(\frac{\epsilon}{2Hn^c}\right)^{\frac{2}{\alpha}}} = 2e^{-2n^{1-\frac{2c}{\alpha}}\left(\frac{\epsilon}{2H}\right)^{\frac{2}{\alpha}}}. \qquad (2.19)
\end{aligned}$$

Meanwhile, by sub-Gaussianity, we infer that

$$\begin{aligned}
&\mathbb{P}\left(u^+\left(X_{[n]}\right) > n^c\right) = 1 - \mathbb{P}\left(u^+\left(X_{[n]}\right) \le n^c\right) \\
&= 1 - \left(\mathbb{P}\left(X_i \le n^c\right)\right)^n \le 1 - \left(1 - e^{-2n^c}\right)^n \\
&\le 1 - \left(1 - ne^{-2n^c}\right) = ne^{-2n^c},
\end{aligned}$$

where the last inequality is obtained by a Taylor series approximation. As a result,

$$\mathbb{P}\left(\left|\mathbb{C}_n - \mathbb{C}(X)\right| > \epsilon\right) \le ne^{-2n^c} + 2e^{-2n^{1-\frac{2c}{\alpha}}\left(\frac{\epsilon}{2H}\right)^{\frac{2}{\alpha}}}.$$

The right side of the above inequality will be optimized with $c = 1 - \frac{2c}{\alpha}$, i.e., for $c = \frac{\alpha}{2+\alpha}$. The claim in (2.18) follows.

To prove (2.17) under the condition that utilities functions are bounded by M, notice that

$$\begin{aligned}
&\left|\int_0^{\infty} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{\infty} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| \\
&= \left|\int_0^{M} w^+\left(\mathbb{P}\left(u^+(X) > t\right)\right) dt - \int_0^{M} w^+\left(1 - \hat{F}_n^+(t)\right) dt\right| \\
&\le HM \sup_{x\in\mathbb{R}} \left|\mathbb{P}\left(u^+(X) < t\right) - \hat{F}_n^+(t)\right|^\alpha.
\end{aligned}$$

The bound in (2.17) can be inferred by replacing n^c by M and $\frac{\epsilon}{2}$ by ϵ in inequality (2.19).

□

Proof. (***Corollary 1***)

When the utilities are bounded by M, integrating the high-probability bound (2.17) in Theorem 3, we obtain

$$\begin{aligned}&\mathbb{E}\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \leq \int_0^\infty \mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \geq \epsilon\right) d\epsilon \\ &\leq 4\int_0^\infty \exp\left(-2n\left(\epsilon/HM\right)^{2/\alpha}\right) d\epsilon \leq \frac{8HM\Gamma\left(\alpha/2\right)}{n^{\alpha/2}}. \qquad (2.20)\end{aligned}$$

For the sub-Gaussian case, notice that if we truncate $u^+\left(X_{[n]}\right)$ by $n^c\sqrt{\epsilon}$ instead of n^c and repeat the steps used in the proof of Theorem 3, we obtain

$$\mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \geq \epsilon\right) \leq ne^{-2n^c\epsilon^{\frac{1}{2}}} + 2e^{-2n^{1-\frac{2c}{\alpha}}\left(\frac{\epsilon^{\frac{1}{2}}}{2H}\right)^{\frac{2}{\alpha}}}.$$

Setting $c = \frac{\alpha}{2+\alpha}$, we obtain the following:

$$\mathbb{E}\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \leq \frac{4\Gamma\left(2\right)}{n^{\frac{2\alpha}{\alpha+2}}} + \frac{\Gamma\left(\alpha\right)2^\alpha\left(2H\right)^2}{n^{\frac{\alpha^2}{2+\alpha}}}.$$

□

Lipschitz continuous weights

Setting $\alpha = 1$, one can obtain the asymptotic consistency claim in Theorem 2 for Lipschitz weight functions. However, this result is under an assumption which requires a lower bound on the difference quotient of the distribution functions of $u^+(X)$ and $u^-(X)$. Using a different proof technique and assumption 2 in place of assumption 1, we can obtain a result similar to Theorem 2 without additional assumption on the distribution functions of $u^\pm$. The following claim makes this precise.

Theorem 4. *(**Asymptotic consistency**) If assumptions 2 and 3 hold, then we have*

$$\overline{\mathbb{C}}_n \to \mathbb{C}(X) \textit{ a.s. as } n \to \infty.$$

In addition, if we assume that the utilities $u^+(X)$ and $u^-(X)$ are bounded above by $M < \infty$ w.p. 1, then we have $\forall \epsilon > 0, \delta > 0$,

$$\mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \le \epsilon\right) > 1 - \delta\,, \forall n \ge \ln\left(\frac{1}{\delta}\right) \cdot \frac{4L^2M^2}{\epsilon^2}.$$

Proof. (***Theorem 4***)

We first prove the asymptotic convergence claim for the first integral (2.4) in the CPT-value estimator in Algorithm 1, i.e., we show

$$\int_0^\infty w^+\left(1 - \hat{F}_n^+(x)\right) dx \to \int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > x\right)\right) dx. \tag{2.21}$$

Since w^+ is Lipschitz continuous with constant L, we have almost surely that $w^+\left(1 - \hat{F}_n(x)\right) \le L\left(1 - \hat{F}_n(x)\right)$, for all n and $w^+\left(\mathbb{P}\left(u^+(X) > x\right)\right) \le L\left(\mathbb{P}\left(u^+(X) > x\right)\right)$, since $w^+(0) = 0$.

We have

$$\begin{aligned} &\int_0^\infty \left(\mathbb{P}\left(u^+(X) > x\right)\right) dx = \mathbb{E}\left[u^+(X)\right], \text{ and} \\ &\int_0^\infty \left(1 - \hat{F}_n^+(x)\right) dx = \int_0^\infty \int_x^\infty d\hat{F}_n(t)\, dx. \end{aligned} \tag{2.22}$$

Since $\hat{F}_n^+(x)$ has bounded support on $\mathbb{R}$ $\forall n$, the integral in (2.22) is finite. Applying Fubini's theorem to the RHS of (2.22), we obtain

$$\int_0^\infty \int_x^\infty d\hat{F}_n(t)\, dx = \int_0^\infty t d\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n u^+\left(X_{[i]}\right),$$

where $u^+\left(X_{[i]}\right), i = 1, \ldots, n$ are the order statistics, i.e., $u^+\left(X_{[1]}\right) \le \ldots \le u^+\left(X_{[n]}\right)$.

Notice that

$$\frac{1}{n}\sum_{i=1}^{n} u^{+}\left(X_{[i]}\right) = \frac{1}{n}\sum_{i=1}^{n} u^{+}\left(X_i\right) \xrightarrow{a.s} \mathbb{E}\left[u^{+}\left(X\right)\right].$$

From the foregoing,

$$\lim_{n\rightarrow\infty}\int_0^{\infty} L\left(1-\hat{F}_n\left(x\right)\right)dx \xrightarrow{a.s} \int_0^{\infty} L\left(\mathbb{P}\left(u^{+}\left(X\right)>x\right)\right)dx.$$

The claim in (2.21) now follows by invoking the generalized dominated convergence theorem by setting $f_n = w^{+}(1-\hat{F}_n^{+}(x))$ and $g_n = L(1-\hat{F}_n(x))$, and noticing that $L(1-\hat{F}_n(x)) \xrightarrow{a.s.} L(\mathbb{P}\left(u^{+}(X) > x\right))$ uniformly over x. The latter fact is implied by the Glivenko-Cantelli theorem (see appendix B).

Following similar arguments, it is easy to show that

$$\int_0^{\infty} w^{-}\left(1-\hat{F}_n^{-}(x)\right)dx \rightarrow \int_0^{\infty} w^{-}\left(\mathbb{P}\left(u^{-}(X)>x\right)\right)dx.$$

The final claim regarding the almost sure convergence of $\overline{\mathbb{C}}_n$ to $\mathbb{C}(X)$ now follows. □

2.5.3 Lower bound for estimation error

Setting $\alpha = 1$ in Theorem 3, we observe that one can achieve the canonical Monte Carlo rate for Lipschitz continuous weights. Choosing the weights to be the identity function, we observe that the sample complexity cannot be improved. On the other hand, for Hölder continuous weights, we incur a sample complexity of order $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ for accuracy $\epsilon > 0$ and this is generally worse than the canonical Monte Carlo rate of $O\left(\frac{1}{\epsilon^2}\right)$, for $\alpha < 1$. An interesting question here is if the sample complexity from Theorem 3 be improved upon, say to $O(1/\epsilon^2)$ for achieving ϵ accuracy? The next result shows that the best achievable sample complexity, in the minimax sense, is $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ over the class of Hölder-continuous weight functions.

The asymptotic convergence property of CPT-value estimation presented above works for any α Hölder continuous weighting function. In this section, however, we will narrow down our concentration on a smaller class of weighting function which "only" satisfies α Hölder continuous property, i.e., the weighting function $w(\cdot)$ such that

$$h \cdot |x-y|^{\alpha} \leq |w(x) - w(y)| \leq H \cdot |x-y|^{\alpha}, \forall x, y \in [0,1], \tag{2.23}$$

where h and H are two given constants s.t. $H > h$. An example of such a $w(\cdot)$ for $\alpha = 1/2$, as suggested in the proof of Theorem 6 in [36], is: $w(p) = \frac{1}{2} - \frac{1}{\sqrt{2}}\sqrt{\frac{1}{2} - p}$ for $p \in [0, 1/2]$, and $w(p) = \frac{1}{2} + \frac{1}{\sqrt{2}}\sqrt{p - \frac{1}{2}}$ for $p \in (1/2, 1]$.

Before presenting the lower bound result of CPT-value estimator, we will give a brief overview on Minimax Lower bound in the section below:

Minimax Lower Bound

In the field of statistical decision theory, a widespread measure of assessing the quality of an estimator is its minimax risk over a set of different cases. It is believed that the performance of an estimator should be evaluated through not only how well it will do at just one fixed circumstance, but also at a series of cases sharing some common features.

Let us establish the minimax framework in our CPT-estimation scheme as follows: Let $\mathcal{P}$ be a nonempty set of distributions. Let $\mathbb{C}(P)$ denote the CPT-value of a r.v. with distribution $P \in \mathcal{P}$ and $\overline{C}_n : \mathbb{R}^n \to \mathbb{R}$ denote an estimator. The minimax error $\mathcal{R}_n(\mathcal{P})$ is defined by

$$\mathcal{R}_n(\mathcal{P}) := \inf_{\overline{C}_n} \sup_{P \in \mathcal{P}} \mathbb{E}_{X_{1:n} \sim P^{\otimes n}} \left| \overline{\mathbb{C}}_n(X_{1:n}) - \mathbb{C}(P) \right| \tag{2.24}$$

It is however typically impossible (especially in nonparametric problems) to determine the minimax risk exactly. Consequently, one attempts to obtain lower bounds on the minimax risk of $\mathcal{R}_n(\mathcal{P})$.

The following Theorem 5 present the lower bound on the minimax error of 2.24:

Theorem 5. *(**Lower bound**) Given a weighting function $w(\cdot)$ which satisfies* (2.23)*, the minimax error of the associated CPT-value satisfies*

$$\mathcal{R}_n(\mathcal{P}) \geq \frac{1}{C_\alpha (n)^{\frac{\alpha}{2}}}, \textit{ for all } n \geq 1$$

for a set of distributions $\mathcal{P}$ supported within the interval $[0, 1]$, where C_α only related to α.

We use Le Cam's method [85] to establish the lower bound, and before proving theorem 5, we will recall a few types of distances of probability measures.

Definition 2. *(**Probability distances**) Given two probability measures P and Q applied on a set $\mathcal{X}$, which we assume to have densities p and q with respect to a base measure μ, the total variance distance is defined as*

$$\|P - Q\|_{\mathrm{TV}} := \sup_{A \subset \mathcal{X}} |P(A) - Q(A)| = \frac{1}{2} \int_{\mathcal{X}} |p(x) - q(x)| \, dx. \tag{2.25}$$

Furthermore, we recall the Hellinger distance, which takes the form

$$d_{hel}(P, Q)^2 := \int_{\mathcal{X}} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx. \tag{2.26}$$

The KL-divergence is denoted as $D_{kl}(P||Q) := \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)} dx$.

The following lemma relates the total variation distance to each of the other two distances, the proof of which can be found in [24].

Lemma 4. *The total variation distance satisfies the following relationships:*

(a) For the Hellinger distance,

$$\frac{1}{2} d_{hel}(P,Q)^2 \leq \|P-Q\|_{\mathrm{TV}}^2 \leq d_{hel}(P,Q)\sqrt{1-d_{hel}(P,Q)^2/4}. \tag{2.27}$$

(b) Pinsker's inequality: for any distributions P, Q,

$$\|P-Q\|_{\mathrm{TV}}^2 \leq \frac{1}{2} D_{kl}(P||Q). \tag{2.28}$$

Remark 3. *Both KL-divergence and Hellinger distance are very easy to manipulate on product distributions. Specifically, consider the product distributions* $P = P_1 \times \ldots \times P_n$ *and* $Q = Q_1 \times \ldots \times Q_n$*. Then the KL-divergence satisfies the decoupling equality*

$$D_{kl}\left(P||Q\right) = \sum_{i=1}^{n} D_{kl}\left(P_i||Q_i\right),$$

while the Hellinger distance satisfies

$$d_{hel}\left(P,Q\right)^2 = 2 - 2\prod_{i=1}^{n}\left(1 - \frac{1}{2} d_{hel}\left(P_i,Q_i\right)^2\right)$$

Proof. (***Theorem 5***) Without loss of generality, we assume the h from (2.23) equals 1.

Let X_v, $v \in \{-1,+1\}$ denote a Bernoulli r.v. with underlying distribution P_v, $v \in \{+1,-1\}$ defined by

$$P_v(X=1) = \frac{1+v\delta^{\frac{1}{\alpha}}}{2} \text{ and } P_v(X=0) = \frac{1-v\delta^{\frac{1}{\alpha}}}{2},$$

where $\delta \in [0, 2^{-\alpha}]$ is left to be chosen later. Apparently, $\{P_v\}_{v\in\{\pm 1\}} \subset \mathcal{P}$.

Setting $u^+(x) = x, x \geq 0$, $w^+ = w^- = w$, we have

$$\mathbb{C}(P_v) = w(1 + v\delta^{\frac{1}{\alpha}}), \quad v \in \{+1, -1\}.$$

Since that w also satisfies the following condition: $|w(p) - w(\tilde{p})| \geq |p - \tilde{p}|^\alpha$ for $p, \tilde{p} \in (0,1)$, if we let $p = 1 + \delta^{\frac{1}{\alpha}}$ and $\tilde{p} = 1 - \delta^{\frac{1}{\alpha}}$, we have

$$|\mathbb{C}(P_{+1}) - \mathbb{C}(P_{-1})| = |w(p) - w(\tilde{p})| \geq |p - \tilde{p}|^\alpha = \delta\,.$$

By Le Cam's method [85], the minimax error then satisfies

$$\begin{aligned}\mathcal{R}_n(\mathcal{P}) &\geq \frac{\delta}{2}\left(1 - \left\|P^n_{+1} - P^n_{-1}\right\|_{\mathrm{TV}}\right)\\ &\geq \frac{\delta}{2}\Big(1 - \left(\tfrac{1}{2}\, D_{\mathrm{kl}}\left(P^n_{+1} \| P^n_{-1}\right)\right)^{\frac{1}{2}}\Big), \qquad (2.29)\end{aligned}$$

where $P^n_v := \otimes^n P_v$ is the joint distribution of n samples from P_v, $\|\|_{\mathrm{TV}}$ is the total variation distance and (2.29) follows from Pinsker's inequality. We bound the KL-divergences as follows:

$$\begin{aligned}&D_{\mathrm{kl}}\left(P^n_{-} \| P^n_{+}\right) = nD_{\mathrm{kl}}\left(P_+ \| P_-\right)\\ &= \frac{n}{2}\left((1 - \delta^{\frac{1}{\alpha}}) \log \frac{1 - \delta^{\frac{1}{\alpha}}}{1 + \delta^{\frac{1}{\alpha}}} + (1 + \delta^{\frac{1}{\alpha}}) \log \frac{1 + \delta^{\frac{1}{\alpha}}}{1 - \delta^{\frac{1}{\alpha}}}\right)\\ &= n\delta^{\frac{1}{\alpha}} \log \frac{1 + \delta^{\frac{1}{\alpha}}}{1 - \delta^{\frac{1}{\alpha}}} \leq 3n\delta^{\frac{2}{\alpha}}\,,\end{aligned}$$

where the first equality uses chain rule of KL-divergences, the second follows by the definition of KL-divergences between two Bernoulli distributions, and the final inequality follows by using the fact that for $x \in [0, 1/2]$, $x \log \frac{1+x}{1-x} \leq 3x^2$.

Plugging the bound on KL-divergences into (2.29), we obtain

$$\mathcal{R}_n(\mathcal{P}) \geq \frac{\delta}{2}\left(1 - \sqrt{\frac{3n}{2}}\delta^{\frac{1}{\alpha}}\right) = \frac{1}{4(6n)^{\frac{\alpha}{2}}}, \qquad (2.30)$$

for $\delta = \frac{1}{(6n)^{\frac{\alpha}{2}}}$. Noting that $\delta \in [0, 2^{-\alpha}]$ for any $n \geq 1$ finishes the proof.

An alternative proof can be established by invoking the relationship between KL-distance and *Hellinger* distance. To be more precise, notice that

$$\|P_1^n - P_2^n\|_{\mathrm{TV}} \leq d_{hel}\left(P_1^n, P_2^n\right) = \sqrt{2 - 2\left(1 - d_{hel}\left(P_1, P_2\right)^2\right)^n},$$

and based on the definition of P_1 and P_2, we will obtain

$$d_{hel}\left(P_1, P_2\right)^2 = \left(\sqrt{\frac{1+\delta^{\frac{1}{\alpha}}}{2}} - \sqrt{\frac{1-\delta^{\frac{1}{\alpha}}}{2}}\right)^2 = 1 - \sqrt{1 - \delta^{\frac{2}{\alpha}}} = \frac{1}{2}\delta^{\frac{2}{\alpha}} + o\left(\delta^{\frac{2}{\alpha}}\right).$$

Meanwhile, note that $(1 - \delta^{\frac{2}{\alpha}}) = e^{-\delta^{\frac{2}{\alpha}}} + o(\delta^{\frac{2}{\alpha}})$, we have, up to lower order terms in δ, that $\|P_1^n, P_2^n\|_{\mathrm{TV}} \leq \sqrt{2 - 2\exp\left(-\delta^{\frac{2}{\alpha}} n/2\right)}$. Choosing $\delta^{\frac{2}{\alpha}} = 1/(4n)$, we have $\sqrt{2 - 2\exp\left(-\delta^{\frac{2}{\alpha}} n/2\right)} \leq 1/2$, thus giving the lower bound

$$\mathcal{R}_n(\mathcal{P}) \geq \frac{1}{2}\delta^{\frac{2}{\alpha}}\left(1 - \frac{1}{2}\right) = \frac{1}{16n^{\frac{\alpha}{2}}}. \tag{2.31}$$

Both equations (2.30) and (2.31) indicate that the estimation error $\mathcal{R}_n(\mathcal{P})$ cannot be improved beyond $(n^{-1})^{\frac{\alpha}{2}}$. Meanwhile, if we let $C_\alpha = \min\{4(6)^{\frac{\alpha}{2}}, 16\}$, we prove the lower-bound property stated in Theorem 5. □

2.5.4 Locally Lipschitz weights and discrete-valued X

Here we assume that X is a discrete valued r.v. with finite support. Let $p_i, i = 1, \ldots, K$, denote the probability of incurring a gain/loss $x_i, i = 1, \ldots, K$, where $x_1 \leq \ldots \leq x_l \leq 0 \leq x_{l+1} \leq \ldots \leq x_K$ and let

$$F_k = \sum_{i=1}^{k} p_i \text{ if } k \leq l \text{ and } \sum_{i=k}^{K} p_i \text{ if } k > l. \tag{2.32}$$

In this setting, the first integral, say $\mathbb{C}^+(X)$, in the definition of CPT-value (2.1) can be simplified as follows:

$$\begin{aligned}
\mathbb{C}^+(X) = & \int_0^{u^+(x_{l+1})} w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \\
& + \sum_{k=l+1}^{K-1} \int_{u^+(x_k)}^{u^+(x_{k+1})} w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \\
& + \int_{u^+(x_K)}^{\infty} w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \\
= & \, w^+(F_{l+1})u^+(x_{l+1}) + \sum_{i=l+2}^{K} w^+(F_i)(u^+(x_i) - u^+(x_{i-1})) \\
= & \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(F_i) - w^+(F_{i+1})\Big) + u^+(x_K)w^+(p_K).
\end{aligned}$$

The second integral in (2.1) can be simplified in a similar fashion, and we obtain the following form for the overall CPT-value of a discrete-valued X:

$$\begin{aligned}
\mathbb{C}(X) = & \Big(\sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(F_i) - w^+(F_{i+1})\Big) + u^+(x_K)w^+(p_K)\Big) \\
& - \Big((u^-(x_1))w^-(p_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(F_i) - w^-(F_{i-1})\Big)\Big).
\end{aligned}$$

Estimation scheme

Let $X_1, \ldots, X_n$ be n samples from the distribution of X. Define $\hat{p}_k := \frac{1}{n}\sum_{i=1}^n I_{\{X_i = x_k\}}$ and

$$\hat{F}_k = \sum_{i=1}^{k} \hat{p}_k \text{ if } k \leq l \text{ and } \sum_{i=k}^{K} \hat{p}_k \text{ if } k > l. \tag{2.33}$$

Then, we estimate $\mathbb{C}(X)$ as follows:

$$\overline{\mathbb{C}}_n = \Big(\sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\hat{F}_i) - w^+(\hat{F}_{i+1})\Big) + u^+(x_K)w^+(\hat{p}_K)\Big)$$

$$- \Big(u^-(x_1)w^-(\hat{p}_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\hat{F}_i) - w^-(\hat{F}_{i-1})\Big)\Big).$$

Assumption 4. *The weight functions w^+ and w^- are locally Lipschitz continuous, i.e., for any $k = 1, \dots, K$, there exist $L_k < \infty$ and $\rho_k > 0$, such that, for $k = 1, \dots, l$,*

$$|w^-(F_k) - w^-(p)| \le L_k |F_k - p|, \quad \forall p \in (F_k - \rho_k, F_k + \rho_k),$$

and for $k = 1 + 1, \dots, K$,

$$|w^+(F_k) - w^+(p)| \le L_k |F_k - p|, \quad \forall p \in (F_k - \rho_k, F_k + \rho_k).$$

Theorem 6. *Let $L = \max_{k=1,\dots,K} L_k$ and $\rho = \min\{\rho_k\}$, where L_k and ρ_k are as defined in assumption 4, and let $M = \max\{u^-(x_k), k = 1, \dots, l\} \bigcup \{u^+(x_k), k = l+1, \dots, K\}$. If assumption 4 holds, then, $\forall \epsilon > 0, \delta > 0$, we have*

$$\mathbb{P}\left(\left|\overline{\mathbb{C}}_n - \mathbb{C}(X)\right| \le \epsilon\right) > 1 - \delta, \forall n \ge \frac{1}{\kappa} \ln\left(\frac{1}{\delta}\right) \ln\left(\frac{4K}{M}\right),$$

where $\kappa = \min(\rho^2, \epsilon^2/(KLM)^2)$.

In comparison to Theorems 3 and 4, observe that the sample complexity for discrete X scales with the local Lipschitz constant L even if the weight functions may not be lipschitz globally. Further, the local Lipschitz constant L can be much smaller than the global Lipschitz constant of the weight functions.

Before proving theorem 6, assume $w^+ = w^- = w$ without loss of generality, and let

$$\hat{F}_k = \begin{cases} \sum_{i=1}^{k} \hat{p}_k & \text{if } k \le l \\ \\ \sum_{i=k}^{K} \hat{p}_k & \text{if } k > l. \end{cases} \tag{2.34}$$

The following theorem gives the rate at which $\hat{F}_k$ converges to F_k.

Theorem 7. *Let F_k and $\hat{F}_k$ be as defined in* (2.32) *and* (2.33) *respectively. Then, for every $\epsilon > 0$,*

$$P(|\hat{F}_k - F_k| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Proof. We focus on the case when $k > l$, while the case of $k \leq l$ is proved in a similar fashion.

$$\begin{aligned}
&\mathbb{P}\left(\left|\hat{F}_k - F_k\right| > \epsilon\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \geq x_k\}} - \frac{1}{n}\sum_{i=1}^{n} E(I_{\{X_i \geq x_k\}})\right| > \epsilon\right) \\
&= \mathbb{P}\left(\left|\sum_{i=1}^{n} I_{\{X_i \geq x_k\}} - \sum_{i=1}^{n} E(I_{\{X_i \geq x_k\}})\right| > n\epsilon\right) && (2.35) \\
&\leq 2e^{-2n\epsilon^2}, && (2.36)
\end{aligned}$$

where the last inequality above follows by an application of the Hoeffding inequality after observing that X_i are independent of each other and for each i, the corresponding r.v. in (2.35) is an indicator that is bounded above by 1. □

Theorem 8. *Under the conditions of Theorem 6, we have*

$$\begin{aligned}
&\mathbb{P}\left(\left|\sum_{i=1}^{K} u_k w(\hat{F}_k) - \sum_{i=1}^{K} u_k w(F_k)\right| > \epsilon\right) \\
&\leq K\left(e^{-2n\rho^2} + e^{-2n\epsilon^2/(KLM)^2}\right), \text{ where} \\
&u_k = u^-(x_k) \text{ if } k \leq l \text{ and } u^+(x_k) \text{ if } k > l. && (2.37)
\end{aligned}$$

Proof. Observe that

$$\mathbb{P}\left(\left|\sum_{k=1}^{K} u_k w(\hat{F}_k) - \sum_{k=1}^{K} u_k w(F_k)\right| > \epsilon\right)$$

$$\leq \mathbb{P}\left(\bigcup_{k=1}^{K} \left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right)$$
$$\leq \sum_{k=1}^{K} \mathbb{P}\left(\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right) . \tag{2.38}$$

For each $k = 1, \ldots.K$, the function w is locally Lipschitz on $[p_k - \rho, p_k + \rho)$ with common constant L. Therefore, for each k, we can decompose the corresponding probability in (2.38) as follows:

$$\begin{aligned}
&\mathbb{P}\left(\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right) \\
&= \mathbb{P}\left(\left\{\left| F_k - \hat{F}_k \right| > \rho\right\} \bigcap \left\{\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right\}\right) \\
&+ \mathbb{P}\left(\left\{\left| F_k - \hat{F}_k \right| \leq \rho\right\} \bigcap \left\{\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right\}\right) \\
&\leq \mathbb{P}\left(\left| F_k - \hat{F}_k \right| > \rho\right) \\
&+ \mathbb{P}\left(\left\{\left| F_k - \hat{F}_k \right| \leq \rho\right\} \bigcap \left\{\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right\}\right) .
\end{aligned} \tag{2.39}$$

Using the fact that w is L-Lipschitz together with Theorem 7, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left\{\left| F_k - \hat{F}_k \right| \leq \rho\right\} \bigcap \left\{\left| u_k w(\hat{F}_k) - u_k w(F_k) \right| > \frac{\epsilon}{K}\right\}\right) \\
&\leq \mathbb{P}\left(u_k L \left| F_k - \hat{F}_k \right| > \frac{\epsilon}{K}\right) \\
&\leq e^{-2n\epsilon/(KLu_k)^2} \leq e^{-2n\epsilon/(KLM)^2}, \forall k.
\end{aligned} \tag{2.40}$$

Using Theorem 7, we obtain

$$\mathbb{P}\left(\left| F_k - \hat{F}_k \right| > \rho\right) \leq e^{-2n\rho^2}, \forall k. \tag{2.41}$$

Using (2.40) and (2.41) in (2.39), we obtain

$$\mathbb{P}\left(\left| \sum_{k=1}^{K} u_k w(\hat{F}_k) - \sum_{k=1}^{K} u_k w(F_k) \right| > \epsilon\right)$$

$$\leq \sum_{k=1}^{K} \mathbb{P}\left(\left|u_k w(\hat{F_k}) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right)$$

$$\leq K\left(e^{-2n\rho^2} + e^{-2n\epsilon^2/(KLM)^2}\right).$$

The claim follows. □

Proof. (***Theorem 6***)

With u_k as defined in (2.37), we need to prove that, $\forall n \geq \frac{1}{\kappa}\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{4K}{M}\right)$, the following high-probability bound holds

$$\begin{aligned}\mathbb{P}\Bigg(\Bigg|&\sum_{i=1}^{K} u_k\left(w\left(\hat{F}_k\right) - w\left(\hat{F}_{k+1}\right)\right)\\ &-\sum_{i=1}^{K} u_k\left(w\left(F_k\right) - w\left(F_{k+1}\right)\right)\Bigg| \leq \epsilon\Bigg) > 1-\delta.\end{aligned} \tag{2.42}$$

Recall that w is locally Lipschitz continuous with constants $L_1, ...L_K$ at the points $F_1, ...F_K$. From a parallel argument to that in the proof of Theorem 8, it is easy to infer that

$$\begin{aligned}\mathbb{P}\left(\left|\sum_{i=1}^{K} u_k w(\hat{F}_{k+1}) - \sum_{i=1}^{K} u_k w(F_{k+1})\right| > \epsilon\right)&\\ \leq K\left(e^{-2n\rho^2} + e^{-2n\epsilon^2/(KLM)^2}\right)&.\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{P}\Bigg(\Bigg|&\sum_{i=1}^{K} u_k\left(w\left(\hat{F}_k\right) - w\left(\hat{F}_{k+1}\right)\right)\\ &-\sum_{i=1}^{K} u_k\left(w\left(F_k\right) - w\left(F_{k+1}\right)\right)\Bigg| > \epsilon\Bigg)\\ \leq \mathbb{P}&\left(\left|\sum_{i=1}^{K} u_k\left(w\left(\hat{F}_k\right)\right) - \sum_{i=1}^{K} u_k\left(w\left(F_k\right)\right)\right| > \epsilon/2\right)\\ + \mathbb{P}&\left(\left|\sum_{i=1}^{K} u_k\left(w\left(\hat{F}_{k+1}\right)\right) - \sum_{i=1}^{K} u_k\left(w\left(F_{k+1}\right)\right)\right| > \epsilon/2\right)\end{aligned}$$

$$\leq 2K(e^{-2n\rho^2} + e^{-2n\epsilon^2/(KLM)^2}).$$

The claim in (2.42) now follows. □

A variant of Corollary 1 can be obtained by integrating the high-probability bound in Theorem 6; we omit the details here.

2.6 Gradient-based stochastic optimization on CPT-value

2.6.1 Optimization objective

Suppose the r.v. X in (2.1) is a function of a d-dimensional parameter θ. In this section we consider the problem

$$\text{Find } \theta^* = \operatorname*{arg\,max}_{\theta \in \Theta} \mathbb{C}(X^\theta), \tag{2.43}$$

where Θ is a compact and convex subset of $\mathbb{R}^d$. The above problem encompasses policy optimization in an MDP that can be discounted or average or stochastic shortest path and/or partially observed. The difference here is that we apply the CPT-functional to the return of a policy, instead of using the expected return.

2.6.2 Gradient algorithm using SPSA (CPT-SPSA)

Gradient estimation through finite-difference

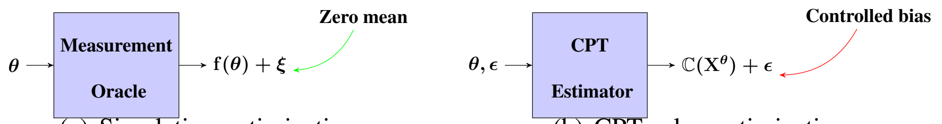
Given that we operate in a learning setting and only have asymptotically unbiased estimates of the CPT-value from Algorithm 1, we require a simulation scheme to estimate $\nabla\mathbb{C}(X^\theta)$. Simultaneous perturbation methods are a general class of stochastic gradient schemes that optimize a function given only noisy sample values - see [11] for a textbook introduction. SPSA is a well-known scheme that estimates the gradient using two sample values In our context, at any iteration n of CPT-SPSA, with parameter θ_n, the gradient $\nabla \mathbb{C}(X^{\theta_n})$ is estimated as follows: For any $i = 1, \ldots, d$,

$$\widehat{\nabla}_i \mathbb{C}(X^{\theta}) = \frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n \Delta_n^i}, \tag{2.44}$$

where δ_n is a positive scalar that satisfies assumption 5 below, $\Delta_n = \left(\Delta_n^1, \ldots, \Delta_n^d\right)^\mathsf{T}$, where $\{\Delta_n^i, i = 1, \ldots, d\}$, $n = 1, 2, \ldots$ are i.i.d. symmetric, ±1-valued Bernoulli r.v.s, independent of $\theta_0, \ldots, \theta_n$ and $\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}$ (resp. $\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}$) denotes the CPT-value estimate that uses m_n samples of the r.v. $X^{\theta_n+\delta_n\Delta_n}$ (resp. $\overline{X}^{\theta_n-\delta_n\Delta_n}$). From the asymptotic mean square analysis that we present later, it is optimal to set $\delta_n = \delta_0/n^{0.16}$. The (asymptotic) unbiasedness of the gradient estimate is proven in Lemma 5.

This idea of using two-point feedback for estimating the gradient has been employed in various settings. Machine learning applications include bandit/stochastic convex optimization - cf. [37], [26]. However, the idea applies to non-convex functions as well - cf. [74], [11].

Remark 4. *A key feature of the gradient estimator in* (2.44) *is that the numerator is fixed for all i. This feature conforms with the main principle of SPSA, which is perturbing all directions simultaneously and estimating the gradient by only two samples, independent of the dimension d of θ.*



(a) Simulation optimization (b) CPT-value optimization

Figure 2.3: Illustration of difference between classic simulation optimization and CPT-value optimization settings

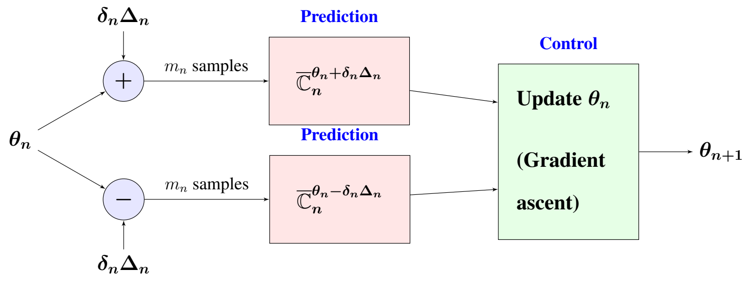


Figure 2.4: Overall flow of CPT-SPSA

Update rule

We incrementally update the parameter θ in the ascent direction as follows:

$$\theta_{n+1} = \Pi\left(\theta_n + \gamma_n \widehat{\nabla}\mathbb{C}(X^{\theta_n})\right), \tag{2.45}$$

where γ_n is a step-size chosen to satisfy assumption 5 below and $\Pi = (\Pi_1, \ldots, \Pi_d)$ is a projection operator that ensures that the update (2.45) stays bounded within the compact and convex set Θ. To be more precise, $\forall \theta^\star \in \mathbb{R}^d,\ \Pi(\theta^\star) = arg\min_{\theta\in\Theta} ||\theta - \theta^\star||_2$. The detailed CPT-SPSA algorithm is illustrated in algorithm 2 and Figures 2.3 and 2.4.

On the number of samples m_n per iteration

Recall that the CPT-value estimation scheme is asymptotically unbiased, i.e., providing samples with parameter θ_n at instant n, we obtain its CPT-value estimate as $\mathbb{C}(X^{\theta_n}) + \psi_n^\theta$, with ψ_n^θ denoting the error in estimation. The estimation error can be controlled by increasing the number of samples m_n in each iteration of CPT-SPSA. This is unlike many simulation optimization settings where one only sees function evaluations with zero mean noise and there is no question of deciding on m_n to control the estimation error as we have in our setting.

Algorithm 2 Structure of CPT-SPSA algorithm.

```
**Input:** initial parameter $\theta_0 \in \Theta$ where $\Theta$ is a compact and convex subset of $\mathbb{R}^d$, perturbation constants $\delta_n > 0$, sample sizes $\{m_n\}$, step-sizes $\{\gamma_n\}$, operator $\Pi : \mathbb{R}^d \rightarrow \Theta$.
**for** $n = 0, 1, 2, \ldots$ **do**
    Generate $\{\Delta_n^i, i = 1, \ldots, d\}$ using symmetric, ±1-valued Bernoulli distribution, independent of $\{\Delta_m, m = 0, 1, \ldots, n-1\}$.
    *CPT-value Estimation (Trajectory 1)*
        Simulate $m_n$ samples using $(\theta_n + \delta_n\Delta_n)$.
        Obtain CPT-value estimates $\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}$ and $\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}$ from Algorithm 1 using $m_n$ samples.
    *CPT-value Estimation (Trajectory 2)*
        Simulate $m_n$ samples using $(\theta_n - \delta_n\Delta_n)$.
        Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}$ from Algorithm 1 using $m_n$ samples.
    *Gradient Ascent*
        Update $\theta_n$ using (2.45).
**end for**
```

To motivate the choice for m_n, we first rewrite the update rule (2.45) as follows:

$$\theta^i_{n+1} = \Pi_i\left(\theta^i_n + \gamma_n\Big(\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta^i_n}\Big) + \kappa_n\right),$$

where $\kappa_n = \frac{(\psi_n^{\theta_n+\delta_n\Delta_n} - \psi_n^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta^i_n}$. Let $\zeta_n = \sum_{l=0}^{n} \gamma_l\kappa_l$. Then, a critical requirement that allows us to ignore the estimation error term ζ_n is the following condition (see Lemma 1 in Chapter 2 of [14]):

$$\sup_{l\geq 0}\left(\zeta_{n+l} - \zeta_n\right) \to 0 \text{ as } n \to \infty.$$

While Theorems 2–3 show that the estimation error ψ^θ is bounded above, to establish convergence of the CPT-SPSA, we increase the number of samples m_n so that the bias vanishes asymptotically. The assumption below provides a condition on the increase rate of m_n.

Assumption 5. *The step-sizes γ_n and the perturbation constants δ_n are positive $\forall n$ and satisfy*

$$\gamma_n, \delta_n \to 0, \;\; \frac{1}{m_n^{\alpha/2}\delta_n} \to 0,$$
$$\sum_n \gamma_n = \infty \;\; \textit{and} \;\; \sum_n \frac{\gamma_n^2}{\delta_n^2} < \infty.$$

While the conditions on γ_n and δ_n are standard for SPSA-based algorithms, the condition on m_n is motivated by the earlier discussion. A simple choice that satisfies the above conditions is $\gamma_n = a_0/n$, $m_n = m_0 n^\nu$ and $\delta_n = \delta_0/n^\gamma$, for some $\nu, \gamma > 0$ with $\gamma > \nu\alpha/2$.

Assumption 6. *The CPT-value $\mathbb{C}(X^\theta)$ is a continuously differentiable function of θ, with bounded third derivative.*

In a typical RL setting involving finite state action spaces, a sufficient condition for ensuring assumption 6 holds is to assume that the policy is continuously differentiable in θ.

Convergence result for CPT-SPSA

We use the ordinary differential equation (ODE) method for establishing asymptotic convergence of CPT-SPSA. Consider the ODE:

$$\dot{\theta}^i_t = \check{\Pi}_i \left(-\nabla_i \mathbb{C}(X^{\theta^i_t}) \right), \text{ for } i = 1, \ldots, d, \tag{2.46}$$

where $\check{\Pi}_i(f(\theta)) := \lim\limits_{\vartheta \downarrow 0} \frac{\Pi_i(\theta + \vartheta f(\theta)) - \theta}{\vartheta}$, for any continuous $f(\cdot)$. Let $\mathcal{K} \subset \{\theta^* \mid \check{\Pi}_i \left(\nabla_i \mathbb{C}(X^{\theta^*})\right) = 0, \forall i = 1, \ldots, d\}$ denote the set of asymptotically stable equilibrium points of the ODE (2.46). That $\mathcal{K} \neq \phi$ can be inferred by using the fact that $\mathbb{C}(X^\theta)$ itself serves as a Lyapunov function for (2.46).

The main convergence result is stated below.

Theorem 9. *If assumptions 1, 3, 5, and 6 hold, then, $\mathcal{K} \neq \phi$ and for θ_n governed by* (2.45)*, we have*

$$\theta_n \to \mathcal{K} \text{ a.s. as } n \to \infty.$$

To prove the main result in Theorem 9, we first show, in the following lemma, that the gradient estimate using SPSA is only an order $O(\delta_n^2)$ term away from the true gradient. The proof differs from the corresponding claim for regular SPSA (see Lemma 1 in [72]), since we have a non-zero bias in the function evaluations, while regular SPSA doesn't. Following this lemma, we complete the proof of Theorem 9 by invoking the well-known Kushner-Clark lemma [49].

Lemma 5. *Let $\mathcal{F}_n = \sigma(\theta_m, m \leq n)$, $n \geq 1$. Then, for any $i = 1, \ldots, d$, we have almost surely,*

$$\left| \mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \right| \mathcal{F}_n \right] - \nabla_i \mathbb{C}(X^{\theta_n}) \Bigg| \xrightarrow{n\to\infty} 0.$$

Proof. Notice that

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] \tag{2.47}$$

$$= \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] + \mathbb{E}\left[\kappa_n \mid \mathcal{F}_n\right], \tag{2.48}$$

where $\kappa_n = \left(\dfrac{\psi^{\theta_n+\delta_n\Delta} - \psi^{\theta_n-\delta_n\Delta}}{2\delta_n\Delta_n^i} \right)$ is the estimation error arising out of the empirical distribution based CPT-value estimation scheme. From Corollary 1 and the fact that $\frac{1}{m_n^{\alpha/2}\delta_n} \to 0$ by assumption (5), we have that

$$\mathbb{E}\kappa_n \to 0 \text{ a.s. as } n \to \infty.$$

Thus,

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right]$$

$$\xrightarrow{n\to\infty} \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right]. \tag{2.49}$$

We now analyze the RHS of (2.49). By using suitable Taylor's series expansions,

$$\mathbb{C}(X^{\theta_n \pm \delta_n\Delta_n}) = \mathbb{C}(X^{\theta_n}) \pm \delta_n \Delta_n^{\mathsf{T}} \nabla \mathbb{C}(X^{\theta_n})$$
$$+ \frac{\delta^2}{2} \Delta_n^{\mathsf{T}} \nabla^2 \mathbb{C}(X^{\theta_n}) \Delta_n + \frac{\delta_n^3}{6} \nabla^3 \mathbb{C}(X^{\tilde{\theta}_n^{\pm}})(\Delta_n \otimes \Delta_n \otimes \Delta_n),$$

where $\otimes$ denotes the Kronecker product and $\tilde{\theta}_n^{+}$ (resp. $\tilde{\theta}_n^{-}$) lie on the line segment connecting θ_n and $(\theta_n + \delta_n\Delta_n)$ (resp. $(\theta_n - \delta_n\Delta_n)$). Using assumption 5 and arguments similar to those used in the proof of Lemma 1 in [72], the fourth order term in each of the Taylor's series expansions above can be shown to be $O(\delta_n^3)$.

From the above, it is easy to see that

$$\begin{aligned}
&\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} - \nabla_i\mathbb{C}(X^{\theta_n}) \\
&= \underbrace{\sum_{j=1, j\neq i}^{N} \frac{\Delta_n^j}{\Delta_n^i}\nabla_j\mathbb{C}(X^{\theta_n})}_{(I)} + O(\delta_n^2).
\end{aligned}$$

Therefore, we simplify the RHS of (2.49) as follows:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] \\
&= \nabla_i\mathbb{C}(X^{\theta_n}) + \mathbb{E}\left[\sum_{j=1, j\neq i}^{N} \frac{\Delta_n^j}{\Delta_n^i}\right]\nabla_j\mathbb{C}(X^{\theta_n}) + O(\delta_n^2) \\
&= \nabla_i\mathbb{C}(X^{\theta_n}) + O(\delta_n^2). \qquad (2.50)
\end{aligned}$$

The first equality above follows from the fact that Δ_n is distributed according to a d-dimensional vector of symmetric, ±1-valued Bernoulli r.v.s and is independent of $\mathcal{F}_n$. The second inequality follows by observing that Δ_n^i is independent of Δ_n^j, for any $i, j = 1, \ldots, d$, $j \neq i$.

The claim follows by using the fact that $\delta_n \to 0$ as $n \to \infty$.

□

Proof. (***Theorem 9***)

Recall that $\mathcal{F}_n = \sigma(\theta_m, m \leq n; \Delta_m, m < n)$, $n \geq 1$. We first rewrite the update rule (2.45) as follows: For $i = 1, \ldots, d$,

$$\theta_{n+1}^i = \Pi_i\left(\theta_n^i + \gamma_n(\nabla_i\mathbb{C}(X^{\theta_n}) + \beta_n + \xi_n)\right), \qquad (2.51)$$

where

$$\beta_n = \mathbb{E}\left(\frac{(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right) - \nabla_i \mathbb{C}(X^{\theta_n}),$$

$$\xi_n = \left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right) - \mathbb{E}\left(\frac{(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right).$$

In the above, β_n is the bias in the gradient estimate due to SPSA and $\{\xi_n\}$ is a martingale difference sequence.

To prove the main claim, we list and verify assumptions (B1)-(B5), which are necessary to invoke Theorem 5.3.1 on pp. 191-196 of [49].

(B1): $\nabla\mathbb{C}(\cdot)$ is a continuous $\mathbb{R}^d$-valued function: holds by assumption in our setting.

(B2): The sequence $\{\beta_n, n \geq 0\}$ is a bounded random sequence with $\beta_n \to 0$ a.s. as $n \to \infty$: follows from Lemma 5.

(B3): The step-sizes $\gamma_n, n \geq 0$ satisfy $\gamma_n \to 0$ as $n \to \infty$ and $\sum_n \gamma_n = \infty$: holds by assumption 5.

(B4): $\{\xi_n, n \geq 0\}$ is a sequence such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\sup_{m\geq n}\left\|\sum_{k=n}^{m}\gamma_k\xi_k\right\| \geq \epsilon\right) = 0. \tag{2.52}$$

We verify this assumption using arguments similar to those used in [72] for SPSA. We first recall Doob's martingale inequality (see (2.1.7) on pp. 27 of [49]):

$$\mathbb{P}\left(\sup_{l\geq 0}\|W_l\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\lim_{l\to\infty}\mathbb{E}\|W_l\|^2. \tag{2.53}$$

Notice that

$$\mathbb{E}\|\xi_n\|^2 \leq \mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2 \tag{2.54}$$

$$\leq \left(\left[\mathbb{E} \left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \right)^2 \right]^{\frac{1}{2}} + \left[\mathbb{E} \left(\frac{\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \right)^2 \right]^{\frac{1}{2}} \right)^2 \tag{2.55}$$

$$\begin{aligned} \leq & \frac{1}{4\delta_n^2} \left[\mathbb{E} \left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}} \right) \right]^{\frac{1}{1+\alpha_1}} \left(\left[\mathbb{E} \left[(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}) \right]^{2+2\alpha_2} \right]^{\frac{1}{1+\alpha_2}} \right. \\ & \qquad \left. + \left[\mathbb{E} \left[(\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}) \right]^{2+2\alpha_2} \right]^{\frac{1}{1+\alpha_2}} \right) \end{aligned} \tag{2.56}$$

$$\leq \frac{\left[\mathbb{E} \left[\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} \right]^{2+2\alpha_2} \right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E} \left[\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n} \right]^{2+2\alpha_2} \right]^{\frac{1}{1+\alpha_2}}}{4\delta_n^2} \tag{2.57}$$

$$\leq \frac{C}{\delta_n^2}, \text{ for some } C < \infty. \tag{2.58}$$

The inequality in (2.54) uses the fact that, for any random variable X, $\mathbb{E} \left\| X - E[X \mid \mathcal{F}_n] \right\|^2 \leq \mathbb{E}X^2$. The inequality in (2.55) follows by the fact that $\mathbb{E}(X+Y)^2 \leq \left((\mathbb{E}X^2)^{1/2} + (\mathbb{E}Y^2)^{1/2} \right)^2$. The inequality in (2.57) uses Hölder's inequality, with $\alpha_1, \alpha_2 > 0$ satisfying $\frac{1}{1+\alpha_1} + \frac{1}{1+\alpha_2} = 1$ and the fact that $\mathbb{E} \left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}} \right) = 1$ as Δ_n^i is a symmetric, ±1-valued Bernoulli r.v. The inequality in (2.58) follows by using the fact that , for any θ, the CPT-value estimate $\bar{\mathbb{C}}(X^\theta) = \mathbb{C}(X^\theta) + \epsilon^\theta$ is bounded a.s. It is because we consider only *proper* policies (which implies that the total cost X^θ is bounded for any policy θ) and finally, by assumption 1, the weight functions are Hölder - these together imply $\mathbb{C}(X^\theta)$ is bounded a.s. for any parameter θ and the estimation error is bounded by Corollary 1. Thus, $\mathbb{E} \left\| \xi_n \right\|^2 \leq \frac{C}{\delta_n^2}$ for some $C < \infty$.

Applying Doob's martingale inequality to the martingale difference $W_l := \sum_{n=0}^{l-1} \gamma_n \xi_n$, $l \geq 1$, we obtain

$$\mathbb{P} \left(\sup_{l \geq k} \left\| \sum_{n=k}^{l} \gamma_n \xi_n \right\| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \mathbb{E} \left\| \sum_{n=k}^{\infty} \gamma_n \xi_n \right\|^2 \leq \frac{1}{\epsilon^2} \sum_{n=k}^{\infty} \gamma_n^2 \mathbb{E} \left\| \xi_n \right\|^2 \leq \frac{dC}{\epsilon^2} \sum_{n=k}^{\infty} \frac{\gamma_n^2}{\delta_n^2},$$

and (2.52) follows by taking limits above and using assumption 5.

(B5): There exists a compact subset $\mathcal{K}$ which is the set of asymptotically stable equilibrium points for the ODE (2.46): To verify this assumption, observe that $\mathbb{C}(X^\theta)$ serves as a strict Lyapunov function for the ODE (2.46), since

$$\frac{d\mathbb{C}(X^\theta)}{dt} = \nabla\mathbb{C}(X^\theta)\dot{\theta} = \nabla\mathbb{C}(X^\theta)\check{\Pi}\bigl(-\nabla\mathbb{C}(X^\theta) \leq 0,$$

with strict inequality outside the set $\mathcal{K}' = \{\theta \mid \check{\Pi}_i\bigl(-\nabla\mathbb{C}(X^\theta)\bigr) = 0, \forall i = 1, \ldots, d\}$. Hence, the set $\mathcal{K}'$ serves as the asymptotically stable attractor for the ODE (2.46). The claim follows from the Kushner-Clark lemma. □

2.6.3 Newton algorithm using SPSA (CPT-SPSA-N)

Need for second-order methods

While stochastic gradient methods are useful in maximizing the CPT-value given biased estimates, they are sensitive to the choice of the step-size sequence $\{\gamma_n\}$. In particular, for a step-size choice $\gamma_n = a_0/n$, if a_0 is not chosen to be greater than $1/\bigl(3\lambda_{min}(\nabla^2\mathbb{C}(X^{\theta^*}))\bigr)$, then the optimum rate of convergence is not achieved, where $\lambda_{\min}$ denotes the minimum eigenvalue, and $\theta^* \in \mathcal{K}$ (see Theorem 9). A standard approach to overcome this step-size dependency is to use iterate averaging, suggested independently by Polyak [58] and Ruppert [66]. The idea is to use larger step-sizes $\gamma_n = 1/n^\varsigma$, where $\varsigma \in (1/2, 1)$, for the update iteration (2.45) and average the iterates in the end, i.e., $\bar{\theta}_{n+1} = \frac{1}{n}\sum_{m=1}^{n}\theta_m$. However, it is well known that iterate averaging is optimal only in an asymptotic sense, while finite-time bounds show that the initial condition is not forgotten sub-exponentially fast (see Theorem 2.2 in [29]). Thus, it is optimal to average iterates only after a sufficient number of iterations have passed, which implies that the iterates are already close to the optimum and the updates can be stopped. An alternative approach is to employ step-sizes of the form $\gamma_n = (a_0/n)M_n$, where M_n converges to $\left(\nabla^2\mathbb{C}(X^{\theta^*})\right)^{-1}$, i.e., the inverse of the Hessian of the CPT-value at the optimum θ^*. Such a scheme gets rid of the step-size dependency (one can set $a_0 = 1$) and still obtains optimal convergence rates. This is the motivation behind having a second-order optimization scheme.

Gradient and Hessian estimation

We estimate the Hessian of the CPT-value function using the scheme suggested by [12]. As in the first-order method, we use ±1 Bernoulli random variables to simultaneously perturb all the coordinates. However, in this case, we require three system trajectories with corresponding parameters $\theta_n + \delta_n(\Delta_n + \widehat{\Delta}_n)$, $\theta_n - \delta_n(\Delta_n + \widehat{\Delta}_n)$ and θ_n, where $\{\Delta_n^i, \widehat{\Delta}_n^i, i = 1, \ldots, d\}$ are i.i.d. ± 1 Bernoulli and independent of $\theta_0, \ldots, \theta_{n-1}$. Using the CPT-value estimates for the aforementioned parameters, we estimate the Hessian and the gradient of the CPT-value function as follows: For $i, j = 1, \ldots, d$, set

$$\widehat{\nabla}_i\mathbb{C}(X_n^{\theta_n}) = \frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}}{2\delta_n\Delta_n^i},$$

$$\widehat{H}_n^{i,j} = \frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)} - 2\overline{\mathbb{C}}_n^{\theta_n}}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}.$$

Notice that the above estimates require three samples, while the second-order SPSA algorithm proposed first in [73] required four. Both the gradient estimate $\widehat{\nabla}\mathbb{C}(X_n^{\theta_n}) = [\widehat{\nabla}_i\mathbb{C}(X_n^{\theta_n})], i = 1, \ldots, d$, and the Hessian estimate $\widehat{H_n} = [\widehat{H}_n^{i,j}], i, j = 1, \ldots, d$, can be shown to be an $O(\delta_n^2)$ term away from the true gradient $\nabla\mathbb{C}(X_n^\theta)$ and Hessian $\nabla^2\mathbb{C}(X_n^\theta)$, respectively (see Lemmas 6–7 below).

Update rule

We update the parameter incrementally using a Newton decrement as follows: For $i = 1, \ldots, d$,

$$\theta^i_{n+1} = \Gamma_i\left(\theta^i_n + \gamma_n \sum_{j=1}^{d} M^{i,j}_n \widehat{\nabla}_j \mathbb{C}(X^\theta_n)\right), \tag{2.59}$$

$$\overline{H}_n = (1-\xi_n)\overline{H}_{n-1} + \xi_n \widehat{H}_n, \tag{2.60}$$

where ξ_n is a step-size sequence that satisfies $\sum_n \xi_n = \infty, \sum_n \xi_n^2 < \infty$ and $\frac{\gamma_n}{\xi_n} \to 0$ as $n \to \infty$. These conditions on ξ_n ensure that the updates to $\overline{H}_n$ proceed on a timescale that is faster than that of θ_n in (2.59) - see Chapter 6 of [14]. Further, Γ is a projection operator as in CPT-SPSA-G and $M_n = [M^{i,j}_n] = \Upsilon(\overline{H}_n)^{-1}$. Notice that we invert $\overline{H}_n$ in each iteration, and to ensure that this inversion is feasible (so that the θ-recursion descends), we project $\overline{H}_n$ onto the set of positive definite matrices using the operator Υ. The operator has to be such that asymptotically $\Upsilon(\overline{H}_n)$ should be the same as $\overline{H}_n$ (since the latter would converge to the true Hessian), while ensuring inversion is feasible in the initial iterations. The assumption below makes these requirements precise.

Assumption 7. *For any* $\{A_n\}$ *and* $\{B_n\}$, $\lim\limits_{n\to\infty} \|A_n - B_n\| = 0 \Rightarrow \lim\limits_{n\to\infty} \parallel \Upsilon(A_n) - \Upsilon(B_n) \parallel = 0$. *Further, for any* $\{C_n\}$ *with* $\sup\limits_n \parallel C_n \parallel < \infty$, $\sup\limits_n \left(\parallel \Upsilon(C_n) \parallel + \parallel \{\Upsilon(C_n)\}^{-1} \parallel\right) < \infty$.

A simple way to define $\Upsilon(\overline{H}_n)$ is to first perform an eigenvalue decomposition of $\overline{H}_n$, followed by projecting all the eigenvalues onto the positive side (see [35] for a similar operator). A simple way to ensure the above is to have $\Upsilon(\cdot)$ as a diagonal matrix and then add a positive scalar δ_n to the diagonal elements so as to ensure invertibility - see [35],

[73] for a similar operator. Such choice satisfies requirement (ii) in Theorem 10 presented below. Algorithm 3 presents the pseudocode, and the main convergence result is stated below.

Theorem 10. *Let assumptions 1, 3, 5, 6 and 7 hold, and consider the ODE:*

$$\dot{\theta}^i_t = \check{\Pi}_i\left(-\Upsilon(\nabla^2\mathbb{C}(X^{\theta_t}))^{-1}\nabla\mathbb{C}(X^{\theta^i_t})\right), \textit{ for } i = 1, \ldots, d,$$

where $\bar{\Pi}_i$ *is as defined in Theorem 9. Let* $\mathcal{K} = \{\theta \in \Theta \mid \nabla\mathbb{C}(X^{\theta^i})\check{\Pi}_i\left(-\Upsilon(\nabla^2\mathbb{C}(X^{\theta}))^{-1}\nabla\mathbb{C}(X^{\theta^i})\right) = 0, \forall i = 1, \ldots, d\}$*. Then, for* θ_n *governed by* (2.59)*, we have*

$$\theta_n \to \mathcal{K} \quad \textit{a.s. as } n \to \infty.$$

Proofs for CPT-SPSA-N

To simplify notation, we will use X^+ (resp. X^-) to denote $X^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}$ (resp. $X^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}$) in the proofs below.

Before proving Theorem 10, we bound the bias in the SPSA-based estimate of the Hessian in the following lemma.

Lemma 6. *For any* $i, j = 1, \ldots, d$*,*

$$\left|\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)} - 2\overline{\mathbb{C}}_n^{\theta_n}}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\middle| \mathcal{F}_n\right]\right.$$
$$\left. -\nabla^2_{i,j}\mathbb{C}(X^{\theta_n})\right| \xrightarrow{n\to\infty} 0 \;\; a.s.$$

Algorithm 3 Structure of CPT-SPSA-N algorithm.

```
Input: initial parameter $\theta_0 \in \Theta$ where $\Theta$ is a compact and convex subset of $\mathbb{R}^d$, perturbation constants $\delta_n > 0$, sample sizes $\{m_n\}$, step-sizes $\{\gamma_n, \xi_n\}$, operator $\Pi : \mathbb{R}^d \to \Theta$.
for $n = 0, 1, 2, \dots$ do
    Generate $\{\Delta_n^i, \widehat{\Delta}_n^i, i = 1, \dots, d\}$ using $\pm 1$ Bernoulli distribution, independent of $\{\Delta_m, \widehat{\Delta}_m, m = 0, 1, \dots, n-1\}$.
    CPT-value Estimation (Trajectory 1)
        Simulate $m_n$ samples using parameter $(\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n))$.
        Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)}$.
    CPT-value Estimation (Trajectory 2)
        Simulate $m_n$ samples using parameter $(\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n))$.
        Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}$.
    CPT-value Estimation (Trajectory 3)
        Simulate $m_n$ samples using parameter $\theta_n$.
        Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n}$ using Algorithm 1.
    Newton step
        Gradient estimate $\widehat{\nabla}_i \mathbb{C}(X_n^\theta) = \dfrac{\overline{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \widehat{\Delta}_n)} - \overline{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \widehat{\Delta}_n)}}{2\delta_n \Delta_n^i}$
        Hessian estimate $\widehat{H}_n = \dfrac{\overline{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \widehat{\Delta}_n)} - 2\widehat{\nabla}_i \mathbb{C}(X_n^\theta)}{\delta_n^2 \Delta_n^i \widehat{\Delta}_n^j}$
        Update the parameter and Hessian according to (2.59)–(2.60).
end for
```

Proof. **Lemma 6** As in the proof of Lemma 5, we can ignore the bias from the CPT-value estimation scheme and conclude that

$$\begin{aligned}&\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}+\overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}-2\overline{\mathbb{C}}_n^{\theta_n}}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\mid\mathcal{F}_n\right]\\&\xrightarrow{n\to\infty}\mathbb{E}\left[\frac{\mathbb{C}(X^+)+\mathbb{C}(X^-)-2\mathbb{C}(X^{\theta_n})}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\mid\mathcal{F}_n\right].\end{aligned}\tag{2.61}$$

Now, the RHS of (2.61) approximates the true gradient with only an $O(\delta_n^2)$ error; this can be inferred using arguments similar to those used in the proof of Proposition 4.2 of [12]. We provide the proof here for the sake of completeness. Using a Taylor series expansion as in Lemma 5, we obtain

$$\begin{aligned}&\frac{\mathbb{C}(X^+)+\mathbb{C}(X^-)-2\mathbb{C}(X^{\theta_n})}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\\&=\frac{(\Delta_n+\hat{\Delta}_n)^\intercal\nabla^2\mathbb{C}(X^{\theta_n})(\Delta_n+\hat{\Delta}_n)}{\triangle_i(n)\hat{\triangle}_j(n)}+O(\delta_n^2)\\&=\sum_{l=1}^{d}\sum_{m=1}^{d}\frac{\Delta_n^l\nabla_{l,m}^2\mathbb{C}(X^{\theta_n})\Delta_n^m}{\Delta_n^i\hat{\Delta}_n^j}\\&+2\sum_{l=1}^{d}\sum_{m=1}^{d}\frac{\Delta_n^l\nabla_{l,m}^2\mathbb{C}(X^{\theta_n})\hat{\Delta}_n^m}{\Delta_n^i\hat{\Delta}_n^j}\\&+\sum_{l=1}^{d}\sum_{m=1}^{d}\frac{\hat{\Delta}_n^l\nabla_{l,m}^2\mathbb{C}(X^{\theta_n})\hat{\Delta}_n^m}{\Delta_n^i\hat{\Delta}_n^j}+O(\delta_n^2).\end{aligned}$$

Taking conditional expectation, we observe that the first and last term above become zero, while the second term becomes $\nabla_{ij}^2\mathbb{C}(X^{\theta_n})$. The claim follows by using the fact that $\delta_n\to 0$ as $n\to\infty$. □

Lemma 7. *For any* $i=1,\ldots,d$,

$$\left|\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n)}-\overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_i\mathbb{C}(X^{\theta_n})\right|\to 0\ \ a.s.\ \textit{as}\ n\to\infty.$$

Proof. As in the proof of Lemma 5, we can ignore the bias from the CPT-value estimation scheme and conclude that

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] \xrightarrow{n\to\infty} \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right].$$

The rest of the proof amounts to showing that the RHS of the above approximates the true gradient with an $O(\delta_n^2)$ correcting term; this can be done in a similar manner as the proof of Lemma 5. Follows by using completely parallel arguments to that in Lemma 5. □

The following lemma establishes that the Hessian recursion (2.59) converges to the true Hessian, for any policy θ.

Lemma 8. *For any* $i, j = 1, \ldots, d$*,*

$$\left\| H_n^{i,j} - \nabla^2_{i,j}\mathbb{C}(X^{\theta_n}) \right\| \to 0 \ a.s. \textit{ and}$$

$$\left\| \Upsilon(\overline{H}_n)^{-1} - \Upsilon(\nabla^2_{i,j}\mathbb{C}(X^{\theta_n}))^{-1} \right\| \to 0 \ a.s.$$

Proof. Follows in a similar manner as in the proofs of Lemmas 7.10 and 7.11 of [11]. □

Proof. ***(Theorem 10)*** The proof follows in a similar manner as the proof of Theorem 7.1 in [11]; we provide a sketch below for the sake of completeness.

We first rewrite the recursion (2.59) as follows: For $i = 1, \ldots, d$

$$\theta_{n+1}^i = \Pi_i\Bigg(\theta_n^i + \gamma_n \sum_{j=1}^{d} \bar{M}^{i,j}(\theta_n)\nabla_j\mathbb{C}(X_n^\theta) + \gamma_n\zeta_n \\ + \chi_{n+1} - \chi_n\Bigg), \tag{2.62}$$

where

$$\bar{M}^{i,j}(\theta) = \Upsilon(\nabla^2\mathbb{C}(X^\theta))^{-1},$$

$$\chi_n = \sum_{m=0}^{n-1} \gamma_m \sum_{k=1}^{d} \bar{M}_{i,k}(\theta_m) \left(\frac{\mathbb{C}(X^-) - \mathbb{C}(X^+)}{2\delta_m \Delta_m^k} \right.$$

$$\left. - E\left[\frac{\mathbb{C}(X^-) - \mathbb{C}(X^+)}{2\delta_m \Delta_m^k} \mid \mathcal{F}_m \right] \right) \text{ and}$$

$$\zeta_n = \mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n)} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}}{2\delta_n \Delta_n^i} \middle| \mathcal{F}_n \right] - \nabla_i \mathbb{C}(X^{\theta_n}).$$

In lieu of Lemmas 6–8, it is easy to conclude that $\zeta_n \to 0$ as $n \to \infty$, χ_n is a martingale difference sequence and that $\chi_{n+1} - \chi_n \to 0$ as $n \to \infty$. Thus, it is easy to see that (2.62) is a discretization of the ODE:

$$\dot{\theta}_t^i = \check{\Pi}_i \left(-\nabla \mathbb{C}(X^{\theta_t^i}) \Upsilon(\nabla^2 \mathbb{C}(X^{\theta_t}))^{-1} \nabla \mathbb{C}(X^{\theta_t^i}) \right). \qquad (2.63)$$

Since $\mathbb{C}(X^\theta)$ serves as a Lyapunov function for the ODE (2.63), it is easy to see that the set $\mathcal{K}= \{\theta \mid \nabla \mathbb{C}(X^{\theta^i}) \check{\Pi}_i \left(-\Upsilon(\nabla^2 \mathbb{C}(X^\theta))^{-1} \nabla \mathbb{C}(X^{\theta^i}) \right) = 0, \forall i = 1, \ldots, d\}$ is an asymptotically stable attractor set for the ODE (2.63). The claim now follows from the Kushner-Clark lemma. □

2.6.4 Gradient algorithm using infinitesimal perturbation analysis (CPT-IPA)

In this section, we develop a gradient based optimization algorithm using perturbation analysis. The spirit of perturbation analysis is to derive an estimator of the gradient of $\mathbb{E}\left(F\left(X^\theta\right)\right)$ through the sample path-wise derivative $\frac{\partial X^\theta}{\partial \theta_i}$ for each $i = 1, ..., d$; see [33] for a detailed review. We design a gradient-based optimization algorithm which directly estimates $\nabla \mathbb{C}\left(X^\theta\right)$ without additional simulations. A parallel work to this section can be found in [17].

A running example of IPA derivative

Assume θ is a one-dimensional scalar, and let $X^\theta \sim \exp(\theta)$, an exponential random variable with mean θ and p.d.f given by

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} I\{x > 0\},$$

where $I\{\cdot\}$ denotes the indicator function. The random variable is usually constructed by $X^\theta = -\theta \ln U$, where $U \sim U(0,1)$, a uniform distribution on $[0,1]$. Differentiating with respect to θ, we get

$$\frac{dX^\theta}{d\theta} = -\ln U = \frac{X^\theta}{\theta}.$$

Moreover, when a utility function u is applied in X^θ, the sample path derivative $\frac{du(X^\theta)}{d\theta}$ is simply $u'(X^\theta)\frac{dX^\theta}{d\theta}$.

Another expression for the CPT measure

Without loss of generality, we assume that X^θ has support only on $\mathbb{R}^+$, and denote u and w as u^+ and w^+ for simplicity. The CPT-measure on X^θ takes the form

$$\mathbb{C}\left(X^\theta\right) = \int_0^{+\infty} w\left(P\left(u\left(X^\theta\right) > z\right)\right) dz. \tag{2.64}$$

Throughout this section we assume that both the probability weighting function w and utility function u are strictly increasing and continuously differentiable in the interior of their domain, and also $u(0) = 0$. Let $G_\theta(z) = P\left(u\left(X^\theta\right) > z\right)$ be the survival function of $u\left(X^\theta\right)$. And assume $\mathbb{C}\left(X^\theta\right)$ exists and is finite. Through integration by part, the CPT-measure in (2.64) can then be formulated as

$$\mathbb{C}\left(X^\theta\right) = w\left(G_\theta(z)\right) z|_0^\infty - \int_0^{+\infty} z d\left(w\left(G_\theta(z)\right)\right)$$

$$
\begin{aligned}
&= \int_0^{+\infty} zd\left(-w\left(G_\theta\left(z\right)\right)\right) \\
&= \int_0^{+\infty} z\frac{dw(z)}{dz}|_{z=G_\theta(z)} f_\theta\left(z\right)dz \\
&= \mathbb{E}[u(X^\theta)\frac{\mathrm{d}w(z)}{\mathrm{d}z}|_{z=G_\theta(u(X^\theta))}], \qquad (2.65)
\end{aligned}
$$

where $f_\theta(\cdot)$ is the density function of $u(x^\theta)$, and $\lim_{z\to\infty} w\left(G_\theta\left(z\right)\right)z = 0$ is inferred from the integrability condition of the CPT-value (2.64).

First-order derivative

Since the CPT-measure involves the probability weighting function $w\left(\cdot\right)$, the utility function u and the distribution function $F_\theta\left(z\right) = P\left(u\left(X^\theta\right) \leq z\right)$, we need to impose some conditions to ensure the exchangeability of differentiation and expectation. [51] studied rigorously to address this issue in perturbation analysis for dynamic systems. Generally, one can move the derivative inside the expectation if a certain Lipschitz condition holds. To be more precise, we stated the two assumptions below:

Assumption 8. *$G\left(\theta, x\right)$ is continuously differentiable w.r.t. θ_i with $i = 1, ...d$ and x, and $\mathbb{E}[\nabla_i\{u\left(X^\theta\right)\}|u(X^\theta) = u]$ is continuous in u.*

Assumption 9. *For any θ, consider Δ_i d-dimensional vector with the ith component being Δ and the rests being 0, $\nabla_i u\left(X^\theta\right) = \lim_{\Delta\to 0}\frac{u\left(X^{\theta+\Delta_i}\right)-u\left(X^\theta\right)}{\Delta}$ exists a.s. and there exists a random variable G s.t.,*

$$
\left|u\left(X^{\theta+\Delta_i}\right)\frac{dw\left(z\right)}{dz}|_{z=G_{\theta+\Delta_i}\left(X^{\theta+\Delta_i}\right)} - u\left(X^\theta\right)\frac{dw\left(z\right)}{dz}|_{z=G_\theta\left(X^\theta\right)}\right| \leq G|\Delta|, a.s.
$$

for $|\Delta|$ small enough.

The following theorem gives an explicit form of the estimator of $\nabla_i \mathbb{C}(X^\theta)$.

Theorem 11. *Under assumptions 8 and 9, we have, $\forall i$,*

$$\nabla_i \mathbb{C}\left(X^\theta\right) = \mathbb{E}[\nabla_i u\left(X^\theta\right) \frac{dw(z)}{dz}|_{z=G_\theta\left(X^\theta\right)}].$$

Proof. From the reformulation of $\mathbb{C}\left(X^\theta\right)$ in (2.65) and Assumption 9, we will have, through the dominated convergence theorem,

$$\begin{aligned}\nabla_i \mathbb{C}\left(X^\theta\right) &= \mathbb{E}[\nabla_i\{u\left(X^\theta\right) \frac{dw(z)}{dz}|_{z=G_\theta\left(X^\theta\right)}\}] \\ &= \mathbb{E}[\nabla_i u\left(X^\theta\right) \frac{dw(z)}{dz}|_{z=G_\theta\left(X^\theta\right)}] + \mathbb{E}[u\left(X^\theta\right) \frac{d^2w(z)}{dz^2}|_{z=G_\theta\left(X^\theta\right)} \nabla_i G_\theta\left(X^\theta\right)].\end{aligned}$$

Also, notice that

$$\begin{aligned}-\nabla_i G_\theta\left(X^\theta\right) &= \nabla_i F\left(\theta, u\left(X^\theta\right)\right) \\ &= \frac{\partial}{\partial \theta_i} F\left(\theta, u(X^\theta)\right) + f_\theta\left(u(X^\theta)\right) \nabla_i u\left(X^\theta\right),\end{aligned}$$

where $f_\theta(\cdot)$ is the density function of $u\left(X^\theta\right)$. Theorem 1 in Hong [41] exhibits the relationship between the pathwise derivative of $F(\theta, u)$ and the density function $f_\theta\left(u\right)$: For any realization $u^\star$ of $u\left(X^\theta\right)$, we have

$$\nabla_i F\left(\theta, u^\star\right) = -f_\theta\left(u^\star\right) \mathbb{E}[\nabla_i u\left(X^\theta\right) | u\left(X^\theta\right) = u^\star]$$

Therefore, we have

$$\nabla_i\{G_\theta\left(u\left(X^\theta\right)\right)|_{u\left(X^\theta\right)=u^\star}\} = f_\theta\left(u^\star\right)\{\mathbb{E}[\nabla_i u\left(X^\theta\right) | u\left(X^\theta\right) = u^\star] - \nabla_i u\left(X^\theta\right)|_{u\left(X^\theta\right)=u^\star}\}.$$

Therefore, we have

$$\mathbb{E}[\nabla_i G_\theta\left(u\left(X^\theta\right)\right) | X^\theta]$$

$$= f_\theta\left(u\left(X^\theta\right)\right)\left[\nabla_i u\left(X^\theta\right) - \nabla_i u\left(X^\theta\right)\right] = 0.$$

As a result, we have

$$\mathbb{E}[X^\theta \frac{d^2 w\left(z\right)}{dz}|_{z=G_\theta\left(X^\theta\right)} \nabla_i G_\theta\left(X^\theta\right)] = \mathbb{E}[\mathbb{E}[X^\theta \frac{d^2 w\left(z\right)}{dz}|_{z=G_\theta\left(X^\theta\right)} \nabla_i G_\theta\left(X^\theta\right)|X^\theta]] = 0,$$

thus the statement of the theorem is proved. □

Theorem 11 suggests an unbiased estimator of the gradient $\nabla\mathbb{C}\left(X^\theta\right)$. When available, direct (unbiased) estimators of the gradient will have several advantages over finite difference gradient estimators provided in Sections 2.6.2 and 2.6.3: Direct gradient estimators eliminate the need to determine appropriate values for the finite difference sequences, which influence the accuracy of the estimator. Moreover, direct gradient estimators are computationally efficient in that they generally only require a single run simulation.

However, in many of the simulation optimization or machine learning settings, the distribution of the underline random variables are unknown, and that keeps us from deriving a direct estimator of $\nabla\mathbb{C}(X^\theta)$ through the statement of Theorem 11. One approach to overcome this limit is to approximate $G_\theta(X^\theta)$ by its empirical distribution counterpart.

The algorithm and proof of CPT-IPA

On the high-level, the CPT-IPA optimization algorithm would include the following gradient estimation steps (at round n of updating θ):

Step 1 (Samples from r.v. X^{θ_n}**):** Generate m_n samples $\{\theta_n^1, \ldots, \theta_n^{N_n}\}$ from random variables X^{θ_n}.

Step 2 (Obtain the ordered sample): Obtain the ordered sample $X^{\theta}_{[1]}, X^{\theta}_{[2]}, \ldots, X^{\theta}_{[m_n]}$.

Step 3 (Estimate the gradient) : Obtain the derivative estimate through

$$\widehat{\nabla}_i \mathbb{C}\left(X^{\theta}\right) = \frac{1}{m_n} \sum_{j=1}^{m_n} \nabla_i u\left(X^{\theta}_{[j]}\right) \frac{dw\left(z\right)}{dz}|_{z=\frac{m_n-j-1}{m_n}}. \tag{2.66}$$

The pseudocode of the CPT-IPA optimization scheme is presented in algorithm 4, whereby we choose the parameter updating step-size $\gamma_n = \frac{a_0}{n}$, the same as in assumption 5. And the following theorem shows that, under proper choice of m_n, algorithm 4 will converge to the point that $\nabla \mathbb{C}(X^{\theta}) = 0$, where $\theta \in \Theta$.

Algorithm 4 CPT optimization via IPA gradient estimation.

```
**Input:** Initial parameter $\theta_0 \in \Theta$ where $\Theta$ is a compact and convex subset of $\mathbb{R}$.
**for** $n = 0, 1, 2, \ldots$ **do**
    samples $X_1, \ldots, X_{m_n}$ from the distribution of $X^{\theta}$.
    Obtain the ordered sample $X^{\theta}_{[1]}, X^{\theta}_{[2]}, \ldots, X^{\theta}_{[m_n]}$.
    **for** $i = 0, 1, 2, \ldots, d$ **do**
        Acquire the sample path derivative of each ordered sample $\nabla_i u\left(X^{\theta}_{[j]}\right)$
        Obtain the derivative estimator through
            $\hat{\nabla}_i \mathbb{C}\left(X^{\theta}\right) = \frac{1}{m_n} \sum_{j=1}^{m_n} \nabla_i \{u\left(X^{\theta}_{[j]}\right)\} \frac{dw\left(z\right)}{dz}|_{z=\frac{m_n-j-1}{m_n}}$
    **end for**
    Update $\theta_n$ using $\theta_{n+1} = \Pi\left(\theta_n + \gamma_n \widehat{\nabla} \mathbb{C}(X^{\theta_n})\right)$, where the operator $\Pi$ is defined in (2.45).
**end for**
```

Theorem 12. *If assumptions 6-9 are satisfied, the sequence m_n is choosed properly, and the updating step-size γ_n is given by assumption 5, then $\mathcal{K} \neq \emptyset$ and θ_n generated by algorithm 4 will converge to $\mathcal{K}$ $a.s.$, with $\mathcal{K}$ being the limit trajectory of the ODE defined in* (2.46).

Before proving theorem 12, we need the following well-known lemma about the limit of two given sequences' average, which can also be found in (A30) of [13]:

Lemma 9. ***Cesaro's Lemma*** *If $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$ are two bounded sequences of real numbers such that* $\lim x_n = \lim y_n$ *as $n \to \infty$, then*

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} x_k}{n} = \frac{\sum_{k=1}^{n} y_k}{n}.$$

Proof. ***Theorem 12:*** Without loss of generality, we will prove the theorem under the case where θ has dimension 1. We first rewrite the recursion of algorithm 4 as

$$\theta_{n+1} = \Pi\left(\theta_n + \gamma_n Y_n\right), \tag{2.67}$$

where Π is the projection operator onto the compact support Θ, and Y_n is defined as $\frac{1}{m_n}\sum_{j=1}^{m_n} \frac{du\left(X^{\theta_n}_{[j]}\right)}{d\theta_n}\frac{dw(z)}{dz}|_{z=\frac{m_n-j-1}{m_n}}$. If the following conditions are satisfied, then the main claim will be proved by invoking Theorem 2.1 from [50]:

(C1): $\sup_n \mathbb{E}|Y_n|^2 < \infty$.

(C2): $\nabla\mathbb{C}(X^{\theta})$ is continuous w.r.t θ when $\theta \in \Theta$.

(C3): $\sum_{i=1}^{n} \gamma_i^2 < \infty, \forall n$.

(C4): $\sum_{i=1}^{n} \gamma_i|\beta_i| < \infty, \forall n$, with β_i defined as $\beta_n = \mathbb{E}[Y_n|\mathcal{F}_{n-1}] - \nabla\mathbb{C}(X^{\theta_n})$.

Conditions **(C1)**, **(C2)** and **(C3)** hold by the assumptions of our problem setting, and the rest of the proof amounts to justifying **(C4)**. Since the simulation at round n is independent of the early iterations, β_n of **C4** can be simplified as $\mathbb{E}[Y_n]-\nabla\mathbb{C}(X^{\theta_n})$. In order to validate **(C4)**, we first realize that, $\forall \theta \in \Theta$, $\frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}\frac{dw(z)}{dz}|_{z=\frac{m_n-j-1}{m_n}} = \frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}\frac{dw(z)}{dz}|_{z=\hat{G}_\theta(X^{\theta}_{[j]})}$, with $\hat{G}_\theta$ denoting the empirical estimate of $G_\theta(\cdot)$, and $X_{[j]}$ being the jth order statistic of the samples $X_1, \ldots, X_{m_n}$ from the distribution X^θ. Moreover, we have

$$\sum_{j=1}^{m_n} \frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=\hat{G}_\theta(X^{\theta}_{[j]})} = \sum_{j=1}^{m_n} \frac{du\left(X^{\theta}_{j}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=\hat{G}_\theta(X^{\theta}_{j})}, \tag{2.68}$$

since the summation of the LHS of (2.68) is a reordered form of the RHS of (2.68).

The combination of Glivenko-Cantelli theorem and continuous mapping theorem implies that for each $j \in \{1, 2, \ldots, m_n\}$,

$$\lim_{m_n\to\infty} \frac{du\left(X^{\theta}_{j}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=\hat{G}_\theta(X^{\theta}_{j})} = \frac{du\left(X^{\theta}_{j}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=G_\theta(X^{\theta}_{j})}. \tag{2.69}$$

The law of large numbers implies that

$$\lim_{m_n\to\infty} \frac{1}{m_n}\frac{du\left(X^{\theta}_{j}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=G_\theta(X^{\theta}_{j})} = \mathbb{E}[\frac{du\left(X^{\theta}\right)}{d\theta}\frac{dw(z)}{dz}|_{z=G_\theta\left(X^{\theta}\right)}]. \tag{2.70}$$

In lieu of the Cesaro's Lemma 9, together with equations (2.69) and (2.70), we can conclude the following convergence property that

$$\lim_{m_n\to\infty} \frac{1}{m_n}\sum_{j=1}^{m_n} \frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}\frac{dw\left(z\right)}{dz}|_{z=\frac{m_n-j-1}{m_n}} = \lim_{m_n\to\infty} \frac{d\mathbb{C}(X^{\theta})}{d\theta} \ a.s. \tag{2.71}$$

The above convergence property of (2.71) implies that if we choose the sequence m_n properly, the sequence of approximation errors $\beta_n = \mathbb{E}[Y_n|\mathcal{F}_{n-1}] - \nabla\mathbb{C}(X^{\theta_n})$ can be controlled such that the condition of **(C4)** hold.

□

Remark 5. *The choice of m_n depends on the condition of the specific problem setting. Specifically, we need to look into the properties of the random variable $\frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}$. However,* *in the usual case where* $\frac{du\left(X^{\theta}_{[j]}\right)}{d\theta}$ *is bounded a.s., the convergence rate of the average in* (2.71) *is* $O\left(\frac{1}{\sqrt{n}}\right)$*, which is indicated by the central limit theorem. Therefore, given that we select* γ_n *as* $\frac{1}{n}$*, a safe choice of* m_n *which will guarantee the convergence of algorithm 4 is* n^{α} *with* $\alpha > 1$.

2.6.5 Model-based parameter search algorithm (CPT-MPS)

In this section, we provide a gradient-free algorithm (CPT-MPS) for maximizing the CPT-value that is based on the MRAS$_2$ algorithm proposed by Chang et al. [20]. While CPT-SPSA is a local optimization scheme, CPT-MPS converges to the global optimum, say θ^*, for the problem (2.43), assuming one exists.

The crucial difference between CPT-MPS and MRAS$_2$ is that the latter has an expected function value objective, i.e., it aims to minimize a function by using sample observations that have zero-mean noise. On the other hand, the objective in our setting is the CPT-value, which distorts the underlying transition probabilities. The implication here is that MRAS$_2$ can estimate the expected value using sample averages, while we have to resort to integrating the empirical distribution, which results in biased estimates.

Basic algorithm

To illustrate the main idea in the algorithm, assume we know the form of $\mathbb{C}(X^{\theta})$. Then, the idea is to generate a sequence of reference distributions $g_k(\theta)$ on the parameter space Θ, such that it eventually concentrates on the global optimum θ^*. One simple way, suggested in Chapter 4 of [20] is

$$g_k(\theta) = \frac{\mathcal{H}(\mathbb{C}(X^\theta))g_{k-1}(\theta)}{\int_\Theta \mathcal{H}(\mathbb{C}(X^{\theta'}))g_{k-1}(\theta')\nu(d\theta')}, \;\; \forall\, \theta \in \Theta, \tag{2.72}$$

where ν is the Lebesgue/counting measure on Θ and $\mathcal{H}$ is a strictly decreasing function. The above construction for g_k's assigns more weight to parameters having higher CPT-values. Meanwhile, it is easy to show that g_k converges to a point-mass concentrated at θ^*.

Next, consider a setting where one can obtain the CPT-value $\mathbb{C}(X^\theta)$ (without any noise) for any parameter θ. In this case, we consider a family of parameterized distributions, say $\{f(\cdot, \eta),\, \eta \in H\}$ on Θ and incrementally update the distribution parameter η such that it minimizes the following KL divergence: $\mathcal{D}(g_k, f(\cdot, \eta)) := \int_\Theta \ln \frac{g_k(\theta)}{f(\theta,\eta)} g_k(\theta)\nu(d\theta)$. As recommended in [20], we employ the natural exponential family (NEF) for the family of distributions $f(\cdot, \eta)$, since it ensures that the KL divergence above can be computed analytically. A parameterized family $\{f(\cdot, \eta), \eta \in H \subseteq \Re^m\}$ on $\mathcal{X}$ is said to to NEF if there exist $h : \Re^n \to \Re$, $\Upsilon : \Re^n \to \Re^m$, and $K : \Re^m \to \Re$ such that

$$f(\mathbf{x}, \eta) = \exp\left\{\eta^T \Upsilon(\mathbf{x}) - K(\eta)\right\} h(\mathbf{x}), \;\; \forall\, \eta \in H, \tag{2.73}$$

where $K(\eta) = \ln \int_{\mathbf{x}\in\mathcal{X}} \exp\left\{\eta^T \Upsilon(\mathbf{x})\right\} h(\mathbf{x})\nu(d\mathbf{x})$, H is the natural parameter space $H = \{\eta \in \Re^m : |K(\eta)| < \infty\}$, and the superscript "$T$" denotes the vector transposition. An algorithm to optimize CPT-value in this *noiseless* setting would perform the following update:

$$\eta_{n+1} \in \arg\max_{\eta \in H} \mathbb{E}_{\eta_n} \left[\frac{[\mathcal{H}(\mathbb{C}(X^\theta)]^n}{f(\theta, \eta_n)} \ln f(\theta, \eta) \right], \tag{2.74}$$

where for any given θ and transformation F on X^θ, $\mathbb{E}_{\eta_n}[F(X^\theta)] = \int_\Theta F(X^\theta) f(\theta, \eta_n)\nu(d\theta)$.

Assuming this setup, the CPT-MPS algorithm would involve the following steps:

Step 1 (Candidate parameters): Generate N_n parameters $\{\theta_n^1, \ldots, \theta_n^{N_n}\}$ using the distribution $f(\cdot, \eta_n)$.

Step 2 (CPT-value estimation): Obtain CPT-value estimates $\overline{\mathbb{C}}_n^{\theta_n^i}$, corresponding to the parameters $\theta_n^i, i = 1, \ldots, N_n$.

Step 3 (Parameter update):

$$\eta_{n+1} \in \arg\max_{\eta \in \mathbb{C}} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{[\mathcal{H}(\overline{\mathbb{C}}_n^{\theta_n^i})]^n}{f(\theta_n^i, \eta_n)} \ln f(\theta_n^i, \eta). \tag{2.75}$$

Algorithm 5 presents the pseudocode for the CPT-value optimization setting where we obtain only asymptotically unbiased estimates of the CPT-value $\mathbb{C}(X^\theta)$ for any parameter θ. As in [20], we use only an elite portion of the candidate parameters that have been sampled, as this guides the parameter search procedure towards better regions more efficiently in comparison to an alternative that uses all the candidate parameters for updating η. This can be achieved by using a quantile estimate of the CPT-value function corresponding to candidate policies that were estimated in a particular iteration. The intuition here is that using policies that have performed well guides the parameter search procedure towards better regions more efficiently in comparison to an alternative that uses all the candidate parameters for updating η. Additionally, the CPT-MPS algorithm includes a smoothing function $\widetilde{I}(\cdot, \chi)$ at the final step of updating η, in order to make itself robust against the biasedness inherited from the CPT-estimator. Readers can refer to [20] for a detailed discussion of the choice $f(\cdot, \eta)$, elite sampling and the effect of $\widetilde{I}(\cdot, \chi)$.

The main convergence result is stated below.

Algorithm 5 Structure of CPT-MPS algorithm.

```
Input: family of distributions $\{f(\cdot,\eta)\}$, initial parameter vector $\eta_0$ s.t. $f(\theta,\eta_0) > 0 \; \forall\, \theta \in \Theta$, trajectory lengths $\{m_n\}$, $\rho_0 \in (0,1]$, $N_0 > 1$, $\varepsilon > 0$, $\varsigma > 1$, $\lambda \in (0,1)$, strictly increasing function $\mathcal{H}$and $\chi_{-1} = -\infty$.
for $n = 0, 1, 2, \ldots$ do
  Generate $N_n$ parameters $\Lambda_n = \{\theta_n^1, \ldots, \theta_n^{N_n}\}$ using the mixture distribution $\widetilde{f}(\cdot,\eta_n) = (1-\lambda) f(\cdot,\widetilde{\eta}_n) + \lambda f(\cdot,\eta_0)$.
  for $i = 1, 2, \ldots, N_n$ do
    Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n^i}$ using $m_n$ samples.
  end for
  Elite Sampling:
  Order the CPT-value estimates as $\{\overline{\mathbb{C}}_n^{\theta_n^{(1)}}, \ldots, \overline{\mathbb{C}}_n^{\theta_n^{(N_n)}}\}$.
  Compute the $(1-\rho_n)$-quantile $\widetilde{\chi}_n(\rho_n, N_n) = \overline{\mathbb{C}}_n^{\theta_n^{\lceil(1-\rho_n)N_n\rceil}}$.
  find largest $\bar{\rho} \in (0, \rho_n)$ such that $\widetilde{\chi}_n(\bar{\rho}, N_n) \geq \bar{\chi}_{n-1} + \varepsilon$ (thresholding step);
  if $\bar{\rho}$ exists then
    Set $\bar{\chi}_n = \widetilde{\chi}_n(\bar{\rho}, N_n)$, $\rho_{n+1} = \bar{\rho}$, $N_{n+1} = N_n$, $\theta_n^* = \theta_{1-\bar{\rho}}$.
  else
    Set $\bar{\chi}_n = \overline{\mathbb{C}}_n^{\theta_{n-1}^*}$, $\rho_{n+1} = \rho_n$, $N_{n+1} = \lceil \varsigma N_n \rceil$, $\theta_n^* = \theta_{n-1}^*$.
  end if
  Sampling distribution update:
    $$\eta_{n+1} \in \arg\max_{\eta \in \mathbb{C}} \sum_{i=1}^{N_n} \frac{[\mathcal{H}(\overline{\mathbb{C}}^{\theta_n^i})]^n)}{\widetilde{f}(\theta,\eta_n)} \widetilde{I}\big(\overline{\mathbb{C}}^{\theta_n^i}, \bar{\chi}_n\big) \ln f(\theta,\eta),$$
  where $\widetilde{I}(z,\chi) := 0$ if $z \leq \chi - \varepsilon$, $(z - \chi + \varepsilon)/\varepsilon$ if $\chi - \varepsilon < z < \chi$ and $1$ if $z \geq \chi$.
end for
```

Theorem 13. *Let $\varphi > 0$ be a positive constant satisfying the condition that the set $\big\{\theta : \mathcal{H}(\mathbb{C}(X^\theta) \geq \frac{1}{\varphi}\big\}$ has a strictly positive Lebesgue/counting measure. Assume (A1), (A2) and that $m_n \to \infty$ as $n \to \infty$. Suppose that multivariate normal densities are used for the sampling distribution, i.e., $\eta_n = (\mu_n, \Sigma_n)$, where μ_n and Σ_n denote the mean and covariance of the normal densities. Then,*

$$\lim_{n\to\infty} \mu_n = \theta^* \textit{ and } \lim_{n\to\infty} \Sigma_n = 0_{d\times d} \ \textit{a.s.} \tag{2.76}$$

Proofs for CPT-MPS

Since we obtain samples of the objective (CPT) in a manner that differs from MRAS$_2$, we need to establish that the thresholding step in Algorithm 5 achieves the same effect as it did in MRAS$_2$. This is achieved by the following lemma, which is a variant of Lemma 4.13 from [20], adapted to our setting.

Lemma 10. *The sequence of random variables $\{\theta_n^*, n = 0, 1, \ldots\}$ in Algorithm 5 converges w.p.1 as $n \to \infty$.*

Proof. Let $\mathcal{A}_n$ be the event that the first if statement is true within the *thresholding* step of Algorithm 5. Let $\mathcal{B}_n := \{\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*}) \leq \frac{\varepsilon}{2}\}$. Whenever $\mathcal{A}_n$ holds, we have $\overline{\mathbb{C}}_n^{\theta_n^*} - \overline{\mathbb{C}}_n^{\theta_{n-1}^*} \geq \varepsilon$ and hence, we obtain

$$\begin{aligned}
&\mathbb{P}(\mathcal{A}_n \cap \mathcal{B}_n)\\
&\leq \mathbb{P}\left(\{\overline{\mathbb{C}}_n^{\theta_n^*} - \overline{\mathbb{C}}_{n-1}^{\theta_{n-1}^*} \geq \varepsilon\} \cap \left\{\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*}) \leq \frac{\varepsilon}{2}\right\}\right)\\
&\leq P\Big(\bigcup_{\theta\in\Lambda_n, \theta'\in\Lambda_{n-1}} \Big\{\{\overline{\mathbb{C}}_n^{\theta} - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\}\\
&\qquad\qquad \cap \left\{\mathbb{C}(X^\theta) - \mathbb{C}(X^{\theta'}) \leq \frac{\varepsilon}{2}\right\} \Big\}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{\substack{\theta\in\Lambda_n,\\ \theta'\in\Lambda_{k-1}}} \mathbb{P}\left(\left\{\overline{\mathbb{C}}_n^{\theta} - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\right\} \cap \left\{\mathbb{C}(X^{\theta}) - \mathbb{C}(X^{\theta'}) \leq \frac{\varepsilon}{2}\right\}\right) \\
&\leq |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} \mathbb{P}\left(\left\{\overline{\mathbb{C}}_n^{\theta} - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\right\}\right. \\
&\qquad\qquad \left. \cap \left\{\mathbb{C}(X^{\theta}) - \mathbb{C}(X^{\theta'}) \leq \frac{\varepsilon}{2}\right\}\right) \\
&\leq |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} \mathbb{P}\left(\overline{\mathbb{C}}_n^{\theta} - \overline{\mathbb{C}}_{n-1}^{\theta'} - \mathbb{C}(X^{\theta}) + \mathbb{C}(X^{\theta'}) \geq \frac{\varepsilon}{2}\right) \\
&\leq |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} \left(\mathbb{P}\left(\overline{\mathbb{C}}_n^{\theta} - \mathbb{C}(X^{\theta}) \geq \frac{\varepsilon}{4}\right)\right. \\
&\qquad \left. + \mathbb{P}\left(\overline{\mathbb{C}}_{n-1}^{\theta'} - \mathbb{C}(X^{\theta'}) \geq \frac{\varepsilon}{4}\right)\right) \\
&\leq 4|\Lambda_n||\Lambda_{n-1}| e^{-\frac{m_n\epsilon^2}{8L^2M^2}},
\end{aligned}
$$

where $|\Lambda_n|$ denotes the cardinality of the set Λ_n. From the foregoing, we have $\sum_{n=1}^{\infty} \mathbb{P}\left(\mathcal{A}_n \cap \mathcal{B}_n\right) < \infty$ since $m_n \to \infty$ as $n \to \infty$. Applying the Borel-Cantelli lemma, we obtain $\mathbb{P}\left(\mathcal{A}_n \cap \mathcal{B}_n \text{ i.o.}\right) = 0$. Hence, if $\mathcal{A}_n$ happens infinitely often, then $\mathcal{B}_n^c$ will also happen infinitely often and we have

$$
\begin{aligned}
\sum_{n=1}^{\infty} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] &= \sum_{n:\ \mathcal{A}_n \text{occurs}} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] \\
&\quad + \sum_{n:\ \mathcal{A}_n^c \text{occurs}} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] \\
&= \sum_{n:\ \mathcal{A}_n \text{occurs}} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] \\
&= \sum_{\substack{n:\\ \mathcal{A}_n\cap\mathcal{B}_n\\ \text{occurs}}} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] + \sum_{\substack{n:\\ \mathcal{A}_n\cap\mathcal{B}_n^c\\ \text{occurs}}} \left[\mathbb{C}(X^{\theta_n^*}) - \mathbb{C}(X^{\theta_{n-1}^*})\right] \\
&= \infty \ \text{ w.p.1, since } \varepsilon > 0.
\end{aligned}
$$

In the above, the first equality follows from the fact that if the else clause in thresholding step in Algorithm 5 is hit, then $\theta_n^* = \theta_{n-1}^*$. From the last equality above, we conclude that it is a contradiction because, $\mathbb{C}(X^{\theta}) < \mathbb{C}(X^{\theta^*})$ for any θ (since θ^* is the global maximum). The main claim now follows, since $\mathcal{A}_n$ can happen only a finite number of times. □

Proof. (***Theorem 13***)

Once we have established Lemma 10, the rest of the proof follows in an identical fashion as the proof of Corollary 4.18 of [20], because our algorithm operates in a similar manner as MRAS$_2$ w.r.t. generating the candidate solution using a parameterized family $f(\cdot, \eta)$ and updating the distribution parameter η. The difference, as mentioned earlier, is the manner in which the samples are generated and the objective (CPT-value) function is estimated. The aforementioned lemma established that the elite sampling and thresholding achieve the same effect as that in MRAS$_2$, and hence the rest of the proof follows from [20]. □

2.7 Conclusions

CPT has been a very popular paradigm for modeling human decisions among psychologists/economists, but has escaped the radar of the Machine Learning and Control community. The work of this chapter is the first step in incorporating CPT-based criteria into an RL framework. However, both estimation and control of CPT-based value is challenging. Using temporal-difference learning type algorithms for estimation was ruled out for CPT-value, since the underlying probabilities get (non-linearly) distorted by a weight function. Using empirical distributions, we proposed an estimation scheme that converges at the optimal rate. Next, for the problem of control, since CPT-value does not conform to any Bellman equation, we employed SPSA - a popular simulation optimization scheme and designed both first and second-order algorithms for optimizing the CPT-value function. We provided theoretical convergence guarantees for all the proposed algorithms. We will illustrate the usefulness of CPT-based criteria on numerical examples in Chapter 4.

Chapter 3

Dynamic Programming in Cumulative Prospect Theory

3.1 Overview

Historically, the study of risk-sensitive criteria has focused on their normative applications – i.e., what should be done. The classic example is expected utility function which produce deterministic policy. Recently, literature on dynamic coherent risk measures has broadened the choices for risk-sensitive performance evaluation. In this chapter, we apply the CPT-functional introduced in Section 2.4 as the transitional measure in a nested dynamic programming structure. As compared to the dynamic coherent risk measure proposed in the previous literature, the CPT-driven measure is risk sensitive but non-convex. We rigorously formulate a CPT-driven dynamic programming problem and analyze two infinite horizon problems, namely, discounted and transient. In both cases, we investigate the assumptions needed in order to yield the strongest results of dynamic programming: value and policy iteration converge to a unique value function and an optimal policy. The content of this chapter based on joint work with Lin et al. [53].

3.2 Related work

In many applications, risk-sensitive measures are more appropriate than risk-neutral measures [42]. In standard MDPs, the performance measures are frequently expressed as expected utility functions that are risk-sensitive [23]. For example, many problems evaluate their outcomes by using $E\left[u(X)\right]$, where u is a risk-sensitive utility function (e.g., exponential), and X is a random variable representing the total reward or cost. A notable advantage of this approach is its robustness with respect to modeling errors [27]. Using this approach, if an optimal policy exists, then a deterministic optimal policy exists [16].

Another important class of risk-sensitive criteria is the class of coherent risk measures, which are convex risk measures with the additional property of positive homogeneity (see [45], Def. 2.3). Prominent examples include mean-semideviation and conditional value-at-risk [3, 25]. Recently, dynamic coherent risk measures have received much attention in the literature [64, 21]. In particular, Ruszczyński in [67] concludes that time-consistent coherent risk measures [68] are suitable for solving the dynamic optimization problem.

In problems involving a human decision maker, it is desirable to use criteria that are beyond expected utility and coherent risk measures. A well-known example of a non-coherent performance measure is suggested by Tversky and Kahneman in their cumulative prospect theory (CPT) [78]. A detailed description of cumulative prospect theory is referred to Section 2.4. Although CPT is empirically proved to be able to capture human decision dynamics under uncertainty (e.g., lotteries) [80], its incorporation into dynamic systems is still nascent. Recently, He and Zhou [38] have studied a portfolio choice problem using a CPT-based approach. The problem maximizes the terminal wealth of a self-financing portfolio, a constraint on the action space of the MDP, driven by a financial market that is uncontrollable from the perspective of the investor (see [38], Eq. 3). These results become more difficult, if not impossible, to obtain if these assumptions are eliminated. The motivation of this chapter is to widen the application of CPT-based criteria to more general dynamic problems, paying attention to the structure of optimal policies obtained.

3.3 Motivational Example

CPT is proved to produce optimal randomized policies, which are more robust against modeling errors (see Section 2.4). In financial terms, CPT is used to balance our need for optimal portfolio return, while acknowledging that our model/information is not perfect (see [38]). While the advantage of producing a randomized policy is shared by coherent risk measures, in some instances, CPT produces substantially more randomized policies than those of coherent risk measures.

The following example applies a CPT criterion to an organ transplant problem:

3.3.1 The Organ Transplant Example: A Comparative Analysis

The organ transplant example is from [19]. The problem considers the discrete-time absorbing Markov chain depicted in Figures 3.1 and 3.2.

The initial state S (i.e., sick) represents a patient demanding an organ transplant. 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) W stands for wait, in which case the next state could either be D or S probabilistically; 2) one can choose to transplant (T), which concentrates the tran-

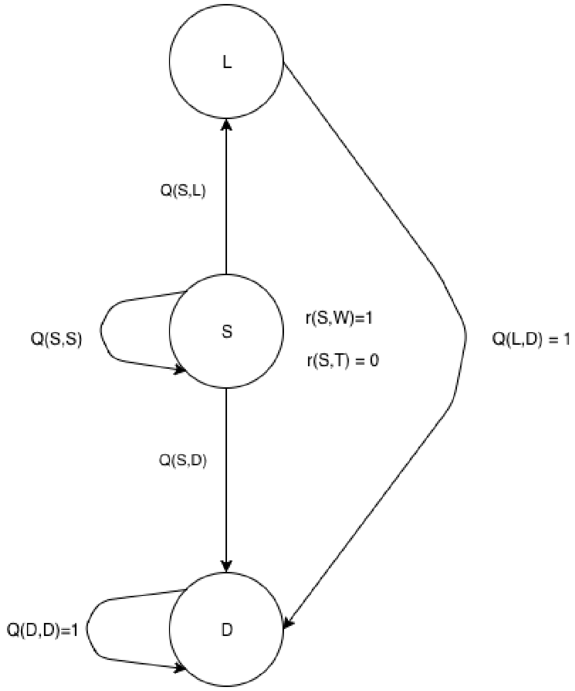


Figure 3.1: Organ transplant model: Transition rate graph on states L, S, and D

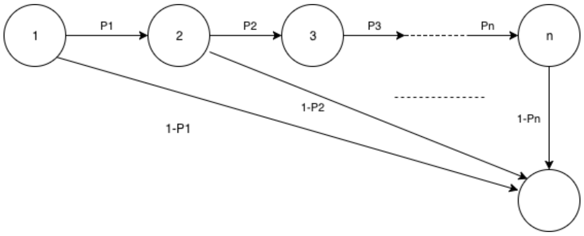


Figure 3.2: Transition graph within state L

Method	$r(L)$	Optimum Value	Optimal λ_W
Expected Value	610.46	846.611	1.000000
Semideviation	515.33	426.139	0.9866
CPT	702.32	104.438	0.868232

Table 3.1: Organ Transplant Optimal Value and Policy Comparison

sition 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 the 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 $r\left(L\right)$. The state L is an aggregation of n states in a survival model representing months of life after the transplant. 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 [19], 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 [19] by negating the rewards.

The policy of the organ transplant problem setting entails a simple probability distribution on choosing between two actions W and T on state S. Formally, for a given policy, we denote λ_W and λ_T as the probability of taking action W and T respectively. Under different criteria(expected-value, semideviation, CPT), we seek the optimal policy regarding the performance measure on the reward of the organ transplant model. See [52] for a detailed discussion of the process including defining the performance measure and optimizing the corresponding objective function.

As is evident from Table 3.1, the CPT performance measure produces a more randomized optimal policy than the other two approaches. The λ_W value of 0.987 for semideviation is close to the deterministic policy of W (i.e., to wait). The ease with which the CPT performance measure is able to obtain an randomized optimal 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. In this problem, the CPT approach yields a vastly different randomized policy, while the mean-semideviation yields a policy that is only marginally randomized. A further discussion of the organ transplant example can be found in [52].

3.4 Contribution of the chapter

In this chapter, our main contribution is on proving the suitability of dynamic programming for solving CPT-based risk-sensitive problems. In particular, we are interested in the case of discounted and transient infinite-horizon problems. Our proof strategy and conclusion have many parallels with that of Ruszczyński's [19, 67].

The greatest technical challenge induced by the probability distortion of CPT is that of nonconvexity of the resulting risk measure; this is in stark contrast to both the expected utility and coherent risk measures. The reason why previous work in this area has insisted on the convexity of the risk measure is because of the diversification principle: the fact a portfolio is less risky than its individual parts. While this constraint makes sense when asking what a rational agent should do, it falls short when we are trying to model what the decision would be in reality.

3.5 Dynamic Programming problems

Dynamic programming, introduced by Bellman [5], has been the subject of intense research in the past five decades; see for example [6]. 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 techniques to lift MDPs' curse-of-dimensionality (e.g., approximate dynamic programming [9]), which enables the widespread application of dynamic programming in many fields.

3.5.1 Abstract Problem Formulation

We are interested in nonempty Borel spaces X and A of states and controls such that for each $x \in X$ there is a nonempty feasible control Borel set $A(x) \subset A$. We denote the set of probability measures over A. We denote by $\mathcal{S}$ the set of all measurable functions $\mu : X \to A$ satisfying $\mu(x) \in A(x)\,,\ \forall x \in X$, which we refer to as policies.

The nonempty Borel space of disturbances is denoted by Δ, and given a state-action pair $(x_k, a_k) \in X \times A$, an element $\delta_k \in \Delta(x_k, a_k) \subset \Delta$ drives the system to its next state through a measurable function $f : X \times A \times \Delta \to X$ by $x_{k+1} = f(x_k, a_k, \delta_k)$. At each time k, a per-step cost is accumulated and denoted by a measurable function $g : X \times A \times \Delta \to \mathbb{R}$. The stochastic kernel $P(\cdot|x, a)$ is defined over $\Delta(x, a)$. Furthermore, we denote both the realization and the random variable disturbance at time k by δ_k. We denote by $R(X)$ the set of real-valued measurable functions $J : X \to \mathbb{R}$. Also, let $H : X \times A \times R(X)$ be a given mapping. For each policy $\mu \in \mathcal{S}$, we consider the mapping $T_\mu : R(X) \to R(X)$ defined by

$$(T_\mu J)(x) = H(x, \mu(x), J), \quad \forall x \in X, J \in R(X),$$

and we also consider the mapping T defined by

$$(TJ)(x) = \inf_{a \in A(x)} H(x, a, J), \quad \forall x \in X, J \in R(X).$$

The mappings T_μ and T serve to define a multistage optimization problem and a Dynamic Programming like methodology for its solution. Particularly, for some function $J \in R(X)$, and nonstationary Markov policy $\pi \in \Pi$ where $\pi = \{\mu_0, \mu_1, \mu_2, \ldots\}$ and Π denotes the set of all feasible non-stationary Markov policies, we define for each integer $N \geq 1$ the functions

$$J_{\pi,N}(x) = \left(T_{\mu_0} T_{\mu_1} T_{\mu_2} \cdots T_{\mu_{N-1}} J\right)(x), \quad \forall x \in X,$$

where $T_{\mu_0} T_{\mu_1} T_{\mu_2} \cdots T_{\mu_{N-1}}$ denotes the composition of mappings $T_{\mu_0}, T_{\mu_1}, T_{\mu_2}, \cdots T_{\mu_{N-1}}$, i.e.,

$$T_{\mu_0} T_{\mu_1} T_{\mu_2} \cdots T_{\mu_{N-1}} J = \left(T_{\mu_0}\left(T_{\mu_1}\left(\cdots\left(T_{\mu_{N-2}}\left(T_{\mu_{N-1}} J\right)\right)\right)\cdots\right)\right).$$

Consider also the function

$$J_\pi\left(x\right) = \limsup_{N\to\infty}\left(T_{\mu_0}T_{\mu_1}T_{\mu_2}\cdots T_{\mu_{N-1}}J\right)\left(x\right), \tag{3.1}$$

which we view as the "infinite horizon cost function" of π. We want to minimize J_π over $\pi \in \Pi$, i.e., to find

$$J^\star\left(x\right) = \inf_{\pi} J_\pi(x), \quad \forall x \in X, \tag{3.2}$$

and a policy $\pi^\star$ that attains the infimum, if it exists. Notice that $J^\star$ can usually be shown to satisfy the "fixed point" property that

$$J^\star(x) = \inf_{a\in A(x)} H(x, a, J^\star). \tag{3.3}$$

We refer to (3.3) as *Bellman's equation*. Another fact is that if an optimal policy $\pi^\star$ exists, it "typically" can be selected to be stationary, $\pi = \mu^\star, \mu^\star, \cdots$, with $\mu^\star \in \mathcal{S}$ satisfying an optimality condition, such as

$$T_{\mu^\star}J^\star = TJ^\star.$$

3.5.2 Example: Markovian Decision Problems with expected cost function

Consider the stationary discrete-time dynamic system

$$x_{k+1} = f(x_k, a_k, \delta_k), \quad k = 0, 1, \cdots,$$

where for all k, the state x_k is an element of a space X, the control a_k is an element of a space A, and δ_k is a random "disturbance", and $\delta_k \in \Delta(x_k, a_k)$. We assume that

$\Delta(x_k, a_k)$ is a countable set and we consider problems with infinite state and control spaces. Follow the notation in Section 3.5.1, we constrain the control a_k to take values in a given nonempty set $A(x_k)$ of A, which depends on the current state x_k. The random disturbance δ_k is characterized by probability distributions $P(\cdot|x_k, a_k)$ that are identical for all k, where $P(\delta_k|x_k, a_k)$ is the probability of occurrence of δ_k, when the current state and control are x_k and a_k, respectively.

Given an initial state x_0, we want to find a policy $\pi = \mu_0, \mu_1, \cdots$, where $\mu_k : X \to A, \mu_k(x_k) \in A(x_k)$, for all $x_k \in X, k = 0, 1, ...$, that minimizes the cost function

$$J_\pi(x_0) = \limsup_{N\to\infty} \mathbb{E}\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), \delta_k)\}, \tag{3.4}$$

subject to the system equation constraint

$$x_{k+1} = f(x_k, \mu_k(x_k), \delta_k), \quad k = 0, 1, ...$$

To make connection with abstract Dynamic Programming, we define

$$H(x, a, J) = \mathbb{E}\{g(x, a, \delta) + \alpha J\left(f(x, a, \delta)\right)\}.$$

Therefore, for any given policy μ, we have

$$\left(T_\mu J\right)(x) = \mathbb{E}\{g(x, \mu(x), \delta) + \alpha J\left(f\left(x, \mu(x), \delta\right)\right)\},$$

and

$$\left(TJ\right)(x) = \inf_{a\in A(x)} \mathbb{E}\{g\left(x, a, \delta\right) + \alpha J\left(f\left(x, a, \delta\right)\right)\}.$$

The N-stage cost can be expressed in terms of T_μ :

$$\left(T_{\mu_0} \cdots T_{\mu_{N-1}} \bar{J}\right)(x_0) = \mathbb{E}\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), \delta_k)\}.$$

The same is true for the infinite stages cost, i.e.:

$$J_\pi(x_0) = \limsup_{N\to} \left(T_{\mu_0}\cdots T_{\mu_{N-1}}\bar{J}\right)(x_0) = \limsup_{N\to\infty} \mathbb{E}\{\sum_{k=0}^{N-1} \alpha^k g(x_k, \mu_k(x_k), \delta_k)\}, \quad (3.5)$$

where $\bar{J}(\cdot)$ is simply the zero-function, i.e., $\bar{J}(x) = 0, \forall x \in X$.

3.5.3 Dynamic programming with CPT-transitional measure

Notice that the expected value in (3.5) can be rewritten as the nested expectation

$$\mathbb{E}\left[g(x_0, a_0, \delta_0) + \mathbb{E}\left[g\left(x_1, a_1, \delta_1\right) + \cdots |x_1\right]|x_0\right].$$

In this section, we generalize the dynamic programming structure introduced in Section 3.5.2 by replacing expected cost with CPT-functional as the transitional measure.

We briefly recall the CPT-functional defined in Section 2.4 below: Given random variable X, the functional, denoted by $\mathbb{C}$, depends on function pairs $u = (u^+, u^-)$ and $w = (w^+, w^-)$. Suppose X is a continuous random variable and the CPT-functional is defined as

$$\begin{aligned}\mathbb{C}(X) = &\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz \\ &- \int_0^\infty w^-\left(\mathbb{P}\left(u^-(X) > z\right)\right) dz\,. \end{aligned} \quad (3.6)$$

The pair $u = (u^+, u^-)$ is the pair of utility functions and the pair $w = (w^+, w^-)$ is the pair of probability weighting functions. Appropriate integrability assumptions has been discussed in Section 2.5, which is skipped in this chapter.

Following the abstract dynamic programming structure constructed in Section 3.5.1, CPT-dynamic programming problem is formulated by introducing the corresponding H

mapping as $H\left(x,a,J\right) =$

$$\int_0^\infty w^+\left(P\left(u^+\left(g\left(x,a,\delta\right)+\alpha J\left(f\left(x,a,\delta\right)\right)\right)>z\right)\right)dz$$
$$\text{-}\int_0^\infty w^-\left(P\left(u^-\left(g\left(x,a,\delta\right)+\alpha J\left(f\left(x,a,\delta\right)\right)\right)>z\right)\right)dz, \tag{3.7}$$

where the system evolves according to the equation $x_{k+1} = f\left(x_k, a_k, \delta_k\right)$, and the discount factor $\alpha \in (0,1)$. Without loss of generality, the reference point is assumed to be zero.

3.5.4 Assumptions for the existence of optimality

To justify whether infinite horizon CPT-driven dynamic programming will attain optimal point which satisfy (3.2), we need to test whether such framework meets the following three assumptions:

Assumption 10. *(Monotonicity) If $J, J' \in R\left(X\right)$ and $J \leq J'$, then $H\left(x,a,J\right) \leq H(x,a,J'),\ \forall x \in X,\ a \in A\left(x\right).$*

Since the contraction assumption requires a Banach space, we introduce the space of real-valued functions J on X embedded with the essential sup-norm $\left\|J\right\|_\infty$.

Assumption 11. *(Contraction) For all $J \in R\left(X\right)$ and $\mu \in \mathcal{S}$, the functions $T_\mu J$ and TJ belong to $R\left(X\right)$. Furthermore, for some $\alpha \in (0,1)$, we have $\left\|T_\mu J - T_\mu J'\right\|_\infty \leq \alpha\left\|J - J'\right\|_\infty,\ \forall J, J' \in R\left(X\right),\ \mu \in \mathcal{S}.$*

We assume our per-step cost function g satisfies the following assumption.

Assumption 12. *There exists a constant $c > 0$ such that $\sup_{x\in X, a\in A(x), \delta\in\Delta}\left|g\left(x,a,\delta\right)\right| \leq c.$*

Note that assumptions 10 and 11 are conditions on the H mapping, which can be obtained from the problem description. Assumption 12 is standard for the classical discounted infinite horizon problem and can be easily verified for practical problems. Additionally, we follow the same procedures in ([8], page 201, Appendix C) to impose appropriate restrictions on $A(x)$ and transitional probability kernel $P(\cdot|x, a)$ to address the measurability issue. In the next section, we seek conditions on either the functions w^+, w^-, u^+, u^- or the underlying model that will satisfy assumptions 10 and 11 for the discounted infinite horizon and transient cases.

3.6 Conditions for the existence of optimal policies

3.6.1 Discounted Infinite Horizon

With the transitional measure introduced in Section 3.5, we first try to explore the properties for the pair of $u = (u^+, u^-)$ and $w = (w^+, w^-)$ in order to satisfy the monotonicity assumption.

Theorem 14. *If $u^+, u^- : \mathbb{R}^+ \to \mathbb{R}^+$ are both monotonically non-decreasing functions, and w^+, w^- are probability weighting functions, then Eq.* (3.7) *satisfies Assumption 10.*

Proof. Let $J \leq J' \in R\left(X\right)$; since u^+ is monotonically non-decreasing,

$$u^+\left(g\left(x, a, \delta\right) + \alpha J\left(f\left(x, a, \delta\right)\right)\right) \leq u^+\left(g\left(x, a, \delta\right) + \alpha J'\left(f\left(x, a, \delta\right)\right)\right),$$

and along with the fact that w_+ is also monotonically non-decreasing, we have

$$\int_0^\infty w^+\left(P\left(u^+\left(g\left(x, a, \delta\right) + \alpha J\left(f\left(x, a, \delta\right)\right)\right) > z\right)\right) dz \tag{3.8}$$

$$\leq \int_0^\infty w^+ \left(P\left(u^+\left(g\left(x,a,\delta\right)+\alpha J'\left(f\left(x,a,\delta\right)\right)\right)>z\right)\right)dz. \tag{3.9}$$

A similar argument can be made about the term on the right hand side of the minus sign in (3.7). □

The next theorem explores the conditions such that the T_μ defined by H is a contraction.

Theorem 15. *Assume the following conditions hold: 1) the assumptions in Theorem 14 hold; 2) u^+, u^- are invertible (denoted by $(u^+)^{-1}$ and $(u^-)^{-1}$), differentiable (denoted by $(u^+)'$ and $(u^-)'$) with $u^+(0) = u^-(0) = 0$; 3) $(u^+)', (u^-)'$ are monotonically non-increasing; 4) there exists a $\beta \in (0,1)$ such that*

$$\int_0^{\alpha c} w^+\left(P\left(Z<z\right)\right)\left(u^+\right)'\left(\alpha c - z\right)dz + \int_0^{\alpha c} w^-\left(P\left(Z>z\right)\right)\left(u^-\right)'\left(z\right)dz \leq \beta c,\ c>0$$

holds for any non-negative real-valued random variable Z. Then Assumption 11 is satisfied.

Proof. Letting μ be any policy and x be any state, we simplify our notation as follows: we use g to denote $g\left(x,\mu\left(x\right),\delta\right)$, J to denote $J\left(f\left(x,\mu\left(x\right),\delta\right)\right)$, and assume $J, J' \in R\left(X\right)$. By the monotonicity property of the mapping H (cf. Theorem 14), we have

$$\begin{aligned}\left(T_\mu J\right)(x) \leq &\int_0^\infty w^+\left(P\left(u^+\left(g+\alpha J' + \alpha\left\|J-J'\right\|_\infty\right)>z\right)\right)dz \\ &- \int_0^\infty w^-\left(P\left(u^-\left(g+\alpha J' + \alpha\left\|J-J'\right\|_\infty\right)>z\right)\right)dz.\end{aligned} \tag{3.10}$$

To simplify the presentation further, we let $c = \left\|J-J'\right\|_\infty$. By using the fact that the utility functions are invertible (i.e., condition 2), the fact that for any $a \in \mathbf{R}$ and

$$b, \xi \in (0,\infty),$$

$$(a+b)^{+} > \xi \iff a > \xi - b,$$

and

$$(a+b)^{-} > \xi \iff -a > \xi + b,$$

and

$$-a > \xi + b \iff \max(-a, 0) > \xi + b,$$

we have

$$\begin{aligned} T_\mu J \le &\int_0^\infty w^+ \left(P\left((g+\alpha J') > (u^+)^{-1}(z) - \alpha c\right)\right) dz \\ &- \int_0^\infty w^- \left(P\left((g+\alpha J')_- > (u^-)^{-1}(z) + \alpha c\right)\right) dz, \end{aligned}$$

where the functions $(x)^+ = \max(x, 0)$ and $(x)^- = -\min(x, 0)$.

By performing a change of variables using $y^+ = (u^+)^{-1}(z) - \alpha c$ and $y^- = (u^-)^{-1}(z) + \alpha c$, we obtain

$$\begin{aligned} T_\mu J \le &\int_{-\alpha c}^\infty w^+ \left(P\left((g+\alpha J') > y^+\right)\right) (u^+)' \left(y^+ + \alpha c\right) dy^+ \\ &- \int_{\alpha c}^\infty w^- \left(P\left((g+\alpha J')_- > y^-\right)\right) (u^-)' \left(y^- - \alpha c\right) dy^-. \end{aligned}$$

Next, we rewrite the equation above by adding sum-to-zero terms:

$$\begin{aligned} T_\mu J \le &\int_0^\infty w^+ \Big(P\Big((g+\alpha J')^+ > y\Big)\Big) (u^+)'(y)\, dz \\ &- \int_0^\infty w^+ \Big(P\Big((g+\alpha J')^+ > y\Big)\Big) (u^+)'(y)\, dy \\ &- \int_0^\infty w^- \Big(P\Big((g+\alpha J')^- > y\Big)\Big) (u^-)'(y)\, dz \\ &+ \int_0^\infty w^- \Big(P\Big((g+\alpha J')^- > y\Big)\Big) (u^-)'(y)\, dy \end{aligned}$$

$$+\int_{-\alpha c}^{\infty} w^{+}\left(P\left(\left(g+\alpha J'\right)>y\right)\right)\left(u^{+}\right)'\left(y+\alpha c\right)dy$$
$$-\int_{\alpha c}^{\infty} w^{-}\left(P\left(\left(g+\alpha J'\right)^{-}>y\right)\right)\left(u^{-}\right)'\left(y-\alpha c\right)dy. \qquad (3.11)$$

Condition 3 implies that $\left(u^{+}\right)'(y+\alpha c)-\left(u^{+}\right)'(y)\leq 0, \forall y\geq 0$ and $\left(u^{-}\right)'(y)-\left(u^{-}\right)'(y-\alpha c)\leq 0, \forall y\geq \alpha c$; therefore we know the following inequalities hold:

$$\int_{-\alpha c}^{\infty} w^{+}\left(P\left(\left(g+\alpha J'\right)>y\right)\right)\left(u^{+}\right)'\left(y+\alpha c\right)dy$$
$$-\int_{0}^{\infty} w^{+}\left(P\left(\left(g+\alpha J'\right)^{+}>y\right)\right)\left(u^{+}\right)'\left(y\right)dy$$
$$\leq\int_{-\alpha c}^{0} w^{+}\left(P\left(\left(g+\alpha J'\right)>y\right)\right)\left(u^{+}\right)'\left(y+\alpha c\right)dy, \qquad (3.12)$$

$$\int_{0}^{\infty} w^{-}\left(P\left(\left(g+\alpha J'\right)^{-}>y\right)\right)\left(u^{-}\right)'\left(y\right)dy$$
$$-\int_{\alpha c}^{\infty} w^{-}\left(P\left(\left(g+\alpha J'\right)^{-}>y\right)\right)\left(u^{-}\right)'\left(y-\alpha c\right)dy$$
$$\leq\int_{0}^{\alpha c} w^{-}\left(P\left(\left(g+\alpha J'\right)^{-}>y\right)\right)\left(u^{-}\right)'\left(y\right)dy. \qquad (3.13)$$

By substituting the inequalities in (3.12) and (3.13) into (3.11), and applying the facts that

$$1)\int_{0}^{\infty} w^{+}\left(P\left(u^{+}\left(g+\alpha J'\right)>z\right)\right)dz$$
$$=\int_{0}^{\infty} w^{+}\left(P\left(\left(g+\alpha J'\right)^{+}>z\right)\right)\left(u^{+}\right)'\left(z\right)dz,$$
$$2)\int_{0}^{\infty} w^{-}\left(P\left(u^{-}\left(g+\alpha J'\right)>z\right)\right)dz$$
$$=\int_{0}^{\infty} w^{-}\left(P\left(\left(g+\alpha J'\right)^{-}>z\right)\right)\left(u^{-}\right)'\left(z\right)dz,$$
$$3)\int_{-\alpha c}^{0} w^{+}\left(P\left(\left(g+\alpha J'\right)>y\right)\right)\left(u^{+}\right)'\left(y+\alpha c\right)dy$$
$$=\int_{0}^{\alpha c} w^{+}\left(P\left(\left(g+\alpha J'\right)>-y\right)\right)\left(u^{+}\right)'\left(\alpha c-y\right)dy,$$

and -$a < z \implies (a)_{-} < z,\ z \in (0, \infty)$, we have

$$\begin{aligned} T_{\mu} J \leq & \int_{0}^{\infty} w^{+}\left(P\left(u^{+}\left(g+\alpha J^{\prime}\right)>z\right)\right) d z \\ & -\int_{0}^{\infty} w^{-}\left(P\left(u^{-}\left(g+\alpha J^{\prime}\right)>z\right)\right) d z \\ & +\int_{0}^{\alpha c} w^{+}\left(P\left(\left(g+\alpha J^{\prime}\right)^{-}<y\right)\right)\left(u^{+}\right)^{\prime}(\alpha c-y) d y \\ & +\int_{0}^{\alpha c} w^{-}\left(P\left(\left(g+\alpha J^{\prime}\right)^{-}>y\right)\right)\left(u^{-}\right)^{\prime}(y) d y . \end{aligned}$$

By using condition 4 of the theorem, we have $\left(T_{\mu} J\right)(x) \leq\left(T_{\mu} J^{\prime}\right)(x)+\beta\left\|J-J^{\prime}\right\|_{\infty}$, which holds for all $x \in X$; the conclusion follows. □

Remark 6. *The conditions in the previous theorem might seem unnatural at first. However, let u^{+} be the identity function and u^{-} be the identity function scaled by $\epsilon \in [0,1]$, and let $w^{-}=w^{+}=w$. In this case, we have $H(x, a, J)=$*

$$\begin{aligned} \int_{0}^{\infty} w\left(P\left((g(x, a, \delta)+\alpha J(f(x, a, \delta)))^{+}>z\right)\right) d z & \\ -\epsilon \int_{0}^{\infty} w\left(P\left((g(x, a, \delta)+\alpha J(f(x, a, \delta)))^{-}>z\right)\right) d z, & \quad (3.14) \end{aligned}$$

which is often used in practice. One can easily check that (3.14) *indeed satisfies all the conditions in Theorems 14 and 15, given the fact that there exists a $\beta \in (0,1)$ such that*

$$w(p)+\epsilon w(1-p) \leq \frac{\beta}{\alpha},\ \forall p \in[0,1] .$$

A very special case is when w is the identity function (i.e., no distortion) in (3.14)*, for which condition 4 in the previous theorem simplifies to*

$$\int_{0}^{\alpha c}[P(Z<z)+P(Z>z)] d z \leq \alpha c \leq \beta c .$$

On the other hand, if we allow w^+ and w^- to be any weighting functions, a simple way to check condition 4 is by using the fact that because w^+ and w^- are probability weighting functions, they are bounded by 1. Applying this knowledge, we can rewrite condition 4 as follows: $\beta \in (0,1)$ such that

$$\int_0^{\alpha c} \left(u^+\right)' (\alpha c - z)\, dz + \int_0^{\alpha c} \left(u^-\right)' (z)\, dz \le \beta c.$$

This condition can be satisfied if we let $(u^+)'$ and $(u^-)'$ be bounded by b_1 and b_2 respectively, and requiring

$$(b_1 + b_2)\alpha < 1,$$

where α is the discount factor. This means that we can satisfy condition 4 for any w^+, w^- if $\alpha < \frac{1}{(b_1+b_2)}$.

3.6.2 Transient Markov Control Model

In this section, we prove the optimality of the dynamic programming equation for transient Markov control models. A transient Markov model evolves according to the equation $x_{k+1} = f(x_k, a_k, \delta_k)$ and has some absorbing state $x_A \in X$, such that if $x_k = x_A$, then $f(x_A, a, \delta) = x_A$ and $g(x_A, a, \delta) = 0$ for all $a \in A(x_A)$, $\delta \in \Delta$. In other words, once an absorbing state is reached, no action can be taken to leave the state and the cost is zero in perpetuity. We denote the first hitting time of the absorbing state with a policy $\pi \in \Pi$ by $\tau_A^\pi := \inf\{t \ge 0 | x_t^\pi = x_A\}$. A transient Markov model reaches its absorbing state in a finite amount of time starting from an initial state x_0, i.e., $\sup_{\pi\in\Pi} E[\tau_A^\pi | x_0] < \infty$. The corresponding H mapping for the systems is: $H(x, a, J) =$

$$\int_0^\infty w^+ \left(\tilde{P}\left(u^+\left(g(x,a,\delta) + J(f(x,a,\delta))\right) > z\right)\right) dz$$

$$-\int_0^\infty w^-\left(\tilde{P}\left(u^-\left(g\left(x,a,\delta\right)+J\left(f\left(x,a,\delta\right)\right)\right)>z\right)\right)dz. \tag{3.15}$$

where $\tilde{P}$ is defined as $\tilde{P}\left(\cdot\right)=P\left(\cdot\cap f\left(x,a,\delta\right)\in X\setminus x_A\right)\leq 1,\ \forall a\in A\left(x\right),\ x\in X.$

Definition 3. *A policy $\pi=\{\mu_0,\mu_1,\dots\}\in\Pi$ is transient with respect to a Markov control model, if there exists a constant c such that $\sum_{k=0}^{\infty}P\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right)\leq c$. If the inequality above holds for all $\pi\in\Pi$, then the model is called uniformly transient. The inequality is also known as the Pliska condition [57].*

The next two theorems give the conditions needed to satisfy the monotonicity and contraction assumptions.

Theorem 16. *If $u^+,u^-:\mathbb{R}^+\to\mathbb{R}^+$ are both monotonically non-decreasing functions, and w^+,w^- are probability weighting functions, then* (3.15) *satisfies assumption 10.*

Proof. Use the same argument as in Theorem 14. □

Theorem 17. *Assume the following conditions hold: 1) the Markov control model is uniformly transient; 2) $\exists C>0$ such that $(u^+)'\left(0\right)\leq C,\ (u^-)'\left(0\right)\leq C$; 3) conditions 1-3 in Theorem 15 hold; 4) $\exists\xi>0$ such that $w^+\left(x\right)\leq\xi x$ and $w^-\left(x\right)\leq\xi x$. Then the operator T_μ defined by using* (3.15) *is a K-step contraction.*

Proof. Observe that $\forall J\in R\left(X\right),\ T_\mu^kJ=T_\mu(T_\mu^{k-1}J)$ As in the proof for the discounted case, fixing $x\in X$ and using condition 3, we can arrive at the following conclusion for the transient case:

$$T_\mu\left(T_\mu^{k-1}J\right)\leq\int_0^\infty w^+\left(\tilde{P}_k\left(u^+\left(g+T_\mu^{k-1}J'\right)\right)>y\right)dz$$
$$-\int_0^\infty w^-\left(\tilde{P}_k\left(u^-\left(g+T_\mu^{k-1}J'\right)\right)>y\right)dz$$

$$+\int_0^{c_{k-1}} w^+\left(\tilde{P}_k\left(\left(g+T_\mu^{k-1}J'\right)^- < y\right)\right)(u^+)'(c_{k-1}-y)\,dy$$
$$+\int_0^{c_{k-1}} w^-\left(\tilde{P}_k\left(\left(g+T_\mu^{k-1}J'\right)^- > y\right)\right)(u^-)'(y)\,dy,$$

where $c_{k-1}=\left\|T_\mu^{k-1}J-T_\mu^{k-1}J'\right\|_\infty$, and $\tilde{P}_k$ is defined as $\tilde{P}_k(\cdot)=P(\cdot\cap f(x_k,a_k,\delta_k\in X\setminus x_A))$.

Using conditions 2 and 4, it is easy to see that

$$\begin{aligned}T_\mu^kJ-T_\mu^kJ' &\leq \int_0^{c_{k-1}} w^+\left(\tilde{P}_k\left(\left(g+T_\mu^{k-1}J'\right)^- < y\right)\right)Cdy\\ &+\int_0^{c_{k-1}} w^-\left(\tilde{P}_k\left(\left(g+T_\mu^{k-1}J'\right)^- > y\right)\right)Cdy\\ &\leq 2\xi CP\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right)c_{k-1}.\end{aligned}$$

The inequality holds because by the preceding definition of $\tilde{P}_k$, i.e.,

$$\tilde{P}_k(\cdot)\leq P\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right).$$

As a result, $\forall k$, the sup-norm of $\left\|T_\mu^kJ-T_\mu^kJ'\right\|_\infty$ is bounded by

$$2\xi CP\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right)\left\|T_\mu^{k-1}J-T_\mu^{k-1}J'\right\|_\infty.$$

In other words, at each time step $k+1$, $\tilde{P}$ has at most $P\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right)$ non-absorbed probability measure.

A similar relationship between $\left\|T_\mu^{k-1}J-T_\mu^{k-1}J'\right\|_\infty$ and $\left\|T_\mu^{k-2}J-T_\mu^{k-2}J'\right\|_\infty$ holds. Therefore by applying the one-step bound relationship repeatedly for k times, one can find a bound of $\left\|T_\mu^kJ-T_\mu^kJ'\right\|_\infty$ with $\|J-J'\|_\infty$:

$$\left\|T_\mu^kJ-T_\mu^kJ'\right\|_\infty\leq(2\xi C)^k\prod_{i=1}^kP\left(f\left(x_i,\mu_i\left(x_i\right),\delta_i\right)\in X\setminus x_A\right)\|J-J'\|_\infty. \qquad (3.16)$$

We denote P_k as $P\left(f\left(x_k,\mu_k\left(x_k\right),\delta_k\right)\in X\setminus x_A\right)$. In the transient Markov model setting, P_k will converge to 0 as k goes to infinity since $\sum_{k=0}^\infty P_k$ is bounded, and this fact can actually enable us to find a constant K such that $(2\xi C)^K\prod_{i=1}^KP_i<1$.

To prove the preceding claim, first observe that since $\lim_{k\to\infty} P_k = 0$, we can find a constant M such that $\forall k > M$, $P_k < \frac{1}{2\xi C}$. Meanwhile, $\forall k > M$, $(2\xi C)^k \prod_{i=1}^{k} P_i$ can be written as the product of two terms:

$$(2\xi C)^k \prod_{i=1}^{k} P_i = (2\xi C)^M \prod_{i=1}^{M} P_i \cdot (2\xi C)^{k-M} \prod_{i=M+1}^{k} P_i. \tag{3.17}$$

In (3.17), $P_i \leq 1 \, \forall i$, and as a matter of fact, $(2\xi C)^M \prod_{i=1}^{M} P_i \leq (2\xi C)^M$. The property that $\forall k > M$, $P_k < \frac{1}{2\xi C}$ implies

$$\lim_{k\to\infty} (2\xi C)^{k-M} \prod_{i=M+1}^{k} P_i = 0.$$

The above convergence property indicates that we can find a constant K which satisfies

$$(2\xi C)^{K-M} \prod_{i=M+1}^{K} P_i \leq 1 / \left((2\xi C)^M \prod_{i=1}^{M} P_i \right).$$

With such K being found which satisfies $(2\xi C)^K \prod_{i=1}^{K} P_i < 1$, we denote γ the constant of $(2\xi C)^K \prod_{i=1}^{K} P_i$, and rewrite (3.16) as

$$\left\| T_\mu^K J - T_\mu^K J' \right\|_\infty \leq \gamma \left\| J - J' \right\|_\infty .$$

The operator T_μ is thus a K-step contraction, and the statement of the theorem is thus proved. □

A K-step contraction is comparable to a contraction, in that the mapping is a contraction in a finite number of steps. The results of dynamic programming are still applicable by substituting the operators T_μ^K and T^K, where the superscript K represents applying the operator K times. The reader may refer to [43] for a more detailed discussion of k-step contractions.

3.7 Conclusion and future work

Non-convexity is the key feature when using CPT to model human decisions. CPT can also be applied to many real-life problems where underestimating rare event probabilities is a concern. In particular, CPT will provide a robust randomized policy, while incorporating our confidence in the model. Based on our proof of the suitability of CPT for dynamic problems, future work will involve the application of CPT to other dynamic problems.

Chapter 4

Experiments

4.1 Overview

In this chapter, we present simulation experiments regarding the CPT-measure introduced in Section 2.4. The chapter is organized as follows: In Section 4.2, we apply the CPT-measure to a few random variables and show that, the optimal parameter values using CPT-value are different from the optimal parameter values using traditional expectation. Section 4.2 tests the sample complexity properties of our proposed CPT-estimator. In Sections 4.3 and 4.4, we incorporate the CPT-measure into real-life simulation examples to investigate the difference between CPT-based decision making problems and traditional expected-utility-based decision making problems. The simulation examples included in this chapter can also be found in the joint paper with Prashanth et al. [44].

4.2 Numerical Experiments

In this section, we show that the optimal CPT-value reacts differently to the change of parameters of the underlying distribution as compared to the optimal expected value. In other words, there are families of random variables $\{X_\theta, \theta \in \Theta\}$ where $\arg\max_\theta \mathbb{E}\left(X_\theta\right)$ is radically different from $\arg\max_\theta \mathbb{C}\left(X_\theta\right)$. This finding would make a case for specialized algorithms that optimize CPT-based criteria, since expected value optimizing algorithms cannot be used as surrogates.

The CPT-value in this section is aligned with the form proposed in (2.1) and uses the following choices for utility and weight functions:

$$
\begin{aligned}
&u^+(x) = |x|^\sigma, \quad u^-(x) = \lambda|x|^\sigma,\\
&w^+(p) = \frac{p^{\eta_1}}{(p^{\eta_1} + (1-p)^{\eta_1})^{\frac{1}{\eta_1}}}, w^-(p) = \frac{p^{\eta_2}}{(p^{\eta_2} + (1-p)^{\eta_2})^{\frac{1}{\eta_2}}},
\end{aligned}
$$

where $\lambda = 0.25$, $\sigma = 0.88$, $\eta_1 = 0.61$ and $\eta_2 = 0.69$. The choices for σ and $w^+(\cdot)$, $w^-(\cdot)$ are based on the recommendations given by [78].

4.2.1 Comparison between CPT and expectation

Since it is usually hard to obtain an analytical expression for the CPT-value, we use numerical integration via the trapezoidal rule. Meanwhile, since we usually have little knowledge about the property of CPT-functional, gradient descent algorithms usually won't guarantee the convergence to the global optimal within the feasible region. Therefore, we consider two settings where the feasible region is triangle shaped over two distribution parameters. In each setting, the expected value optima is calculated analytically, while for the CPT-value, we perform a grid search, where the distance between points in the grid is 0.05.

Example 1. *We consider normally distributed r.v.s with mean μ and variance σ. As shown in Figure 4.1, the feasible region for (μ, σ) is the triangle with vertices $(0.5, 2), (0.5, 6)$ and $(2.5, 2)$. The expected value takes its maximum analytically at $(2.5, 2)$, while a numerical optimization of the CPT-value returned a maximum at $(0.5, 6)$, with corresponding*

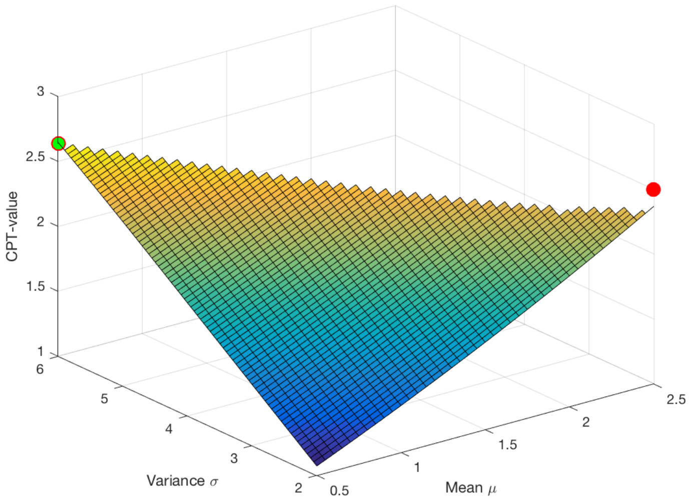


Figure 4.1: CPT-value of normal distributed r.v.s with mean μ and variance σ parameters

CPT-value 2.65. *The CPT value of the r.v.* $N(2.5, 2)$ *was* 2.37.

Example 2. *We consider skew normal distributed r.v.s* $sn(\xi, \omega, \alpha)$ *with location* ξ*, scale* ω *and shape* α*. The mean of* $X_{(\xi,\omega,\alpha)} \sim sn(\xi, \omega, \alpha)$ *is* $\xi + \omega\delta\sqrt{\frac{2}{\pi}}$*, while the variance is* $\omega^2(1 - \frac{2\delta^2}{\pi})$*, with* $\delta = \frac{\alpha}{1+\alpha^2}$*. With* $\alpha = 0.5$*, we set up the feasible region for* (ξ, ω) *to be the triangle with vertices* $(-1, 1), (1, 1)$ *and* $(-1, 5)$ *as shown in Figure 4.3. It turns out that the point* $(-1, 5)$ *returns the largest CPT-value, with* $\mathbb{C}(X_{-1,5,0.5,0.5}) = 2.30$*, while* $\mathbb{E}(X_{-1,5,0.5}) = 0.78$*. On the other hand, the point* $(1, 1)$ *has the largest expected value with* $\mathbb{E}(X_{1,1,0.5}) = 1.36$*, but the CPT value of the same r.v. is* 1.25.

4.2.2 Consistency of CPT estimator

We illustrate the rapid convergence of the estimator in Algorithm 1 for a skew normal distributed r.v. with location, scale and shape parameters set to $2, 1$ and 2, respectively. For calculating the CPT-value, we use the trapezoidal rule. We conducted the

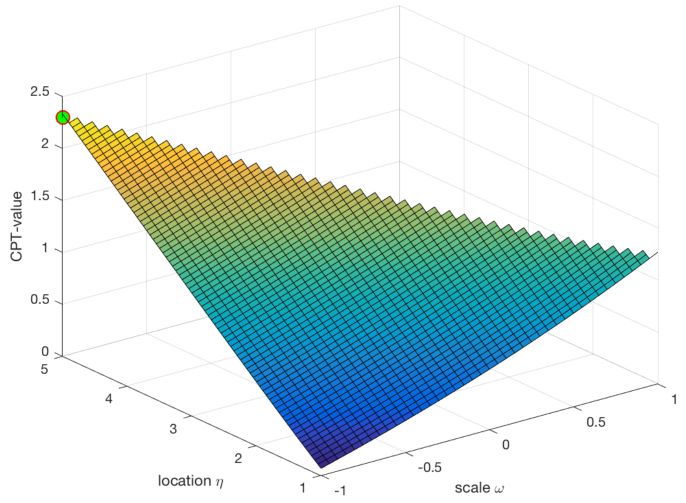


Figure 4.2: Expected value of skew Normal distributed r.v.s with fixed shape $\alpha = 0.5$ and varying location ξ and scale ω

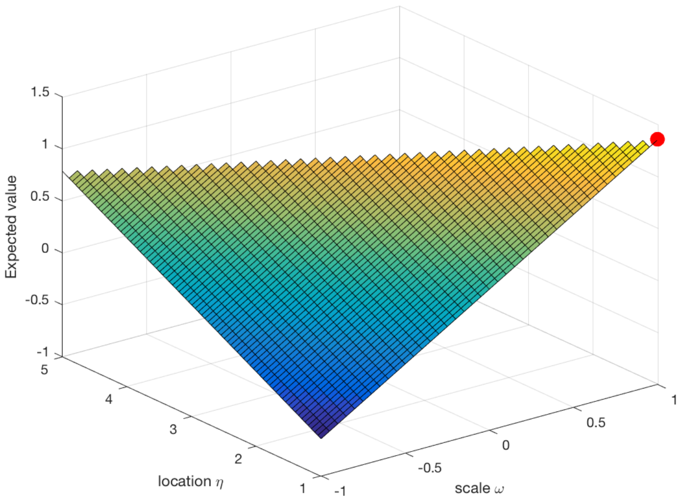


Figure 4.3: CPT value of skew Normal distributed r.v.s with fixed shape $\alpha = 0.5$ and varying location ξ and scale ω

experiment in 100 simulation phases indexed from 1 to 100. In each phase i, we generate i.i.d. estimators $\overline{\mathbb{C}}^{j}_{n_i}(X)$ with n_i samples of skew normal distributed r.v. X, where $j = 1, \ldots, 10$ corresponds to an independent simulation. The number of samples n_i in each phase i ranges from 100 to 10^6. For each phase i, we calculate the estimation error, which is the absolute difference between $\overline{\mathbb{C}}_{n_i} = \frac{1}{10}\sum_{j=1}^{10} \mathbb{C}^{j}_{n_i}$ and the numerically integrated CPT-value. Figure 4.4 the difference between CPT-estimate $\overline{\mathbb{C}}_{n_i}$ using Algorithm 1 and numerically integrated approximation $\tilde{\mathbb{C}}(X)$ to CPT-value $\mathbb{C}(X)$ of a skew normal distributed r.v. X with shape 2, location 2 and scale 1. The shaded bands denote the standard error calculated from ten independent simulations.

Here, the margin of error denotes half the length of the t-confidence interval. It is evident from Figure 4.4 that our CPT-value estimate gets very close to the true CPT-value rapidly.

4.3 House buying at optimal price

We consider a SSP version of an example[1] for buying a house at the optimal price. Suppose the house is priced at x_k any instant k and at the next instant, the price either goes down to $(x_k \times C_{down})$ w.p. p_{down} or goes up to $(x_k \times C_{up})$ w.p. $1 - p_{down}$. The actions are to either wait (denoted w), which results in a holding cost h or to buy (denoted b) at the current price. The horizon is capped at T, with a terminal cost x_T. The goal is to minimize the total cost defined as $D^{\pi}(x^0) = \sum_{k=0}^{\tau} \left(I_{\{a_k=b\}} x_k + I_{\{a_k=w\}} h \right) + I_{\{\tau=T\}} x_T$, where $\tau = \{k | \pi(x_k) = 1\} \wedge T$. We set $T = 20, h = 0.1, C_{up} = 2, C_{down} = 0.5$, and

[1] A similar example has been considered in [22].

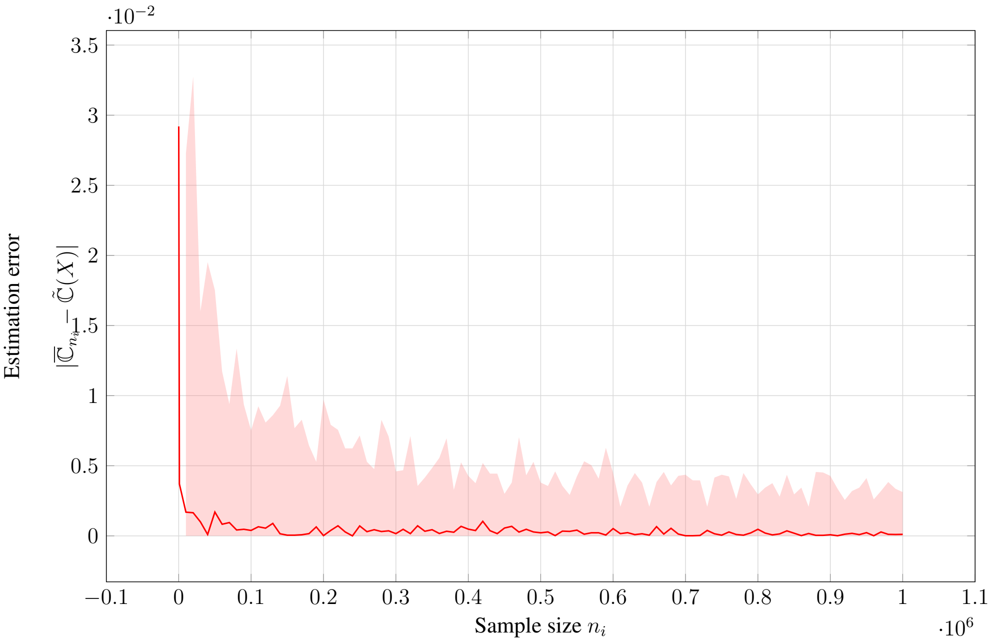


Figure 4.4: Difference between CPT-estimate and numerically integrated approximation of CPT-value

$x_0 = 1$.

4.3.1 Implementation

On this example, we implement the first-order CPT-SPSA and the second-order CPT-SPSA-N algorithms. For the sake of comparison, we also apply value iteration to the SSP example described above. Note that value iteration requires knowledge of the model, while our CPT-based algorithms estimate the CPT-value using simulated episodes. For CPT-SPSA, we set $\delta_n = 1.9/n^{0.101}$ and $\gamma_n = 1/n$, while for CPT-SPSA-N, we set $\delta_n = 3.8/n^{0.166}$ and $\gamma_n = 1/n^{0.6}$. For all algorithms, we set each entry of the initial policy π_0 to 0.1. For CPT-value estimation, we simulate 1000 SSP episodes, with the SSP horizon T set to 20. All algorithms are run with a budget of 1000 samples, which implies

500 iterations of CPT-SPSA and 250 iterations of CPT-SPSA-N. The results presented are averages over 500 independent simulations. For CPT-SPSA/CPT-SPSA-N, the weight functions w^+ and w^- are set to $p^{0.6}/(p^{0.6} + (1-p)^{0.6})$, while the utility functions are identity maps.

4.3.2 Results

Figure 4.5 presents the value function computed using value iteration, while Figures 4.6–4.7 present the CPT-value $\mathbb{C}^{\pi_{end}}(x^0)$ for CPT-SPSA and CPT-SPSA-N, respectively. The performance plots are for various values of p_{down}, the probability of house price going down. From Figure 4.5, we notice that the variations in expected total cost is larger in comparison to that in CPT-value. A similar observation holds true for an SPSA-based algorithm from [10] that optimizes the regular value function. While it is difficult to plot the entire policies, for the expected value minimizing algorithms it was observed that there were drastic changes in the policies with a change of 0.01 in p_{down}, while CPT-SPSA/CPT-SPSA-N resulted in randomized policies that smoothly transitioned with changes in p_{down}. Figure 4.5, 4.6 and 4.7 together verify that CPT-aware SPSA algorithms are less sensitive to the model changes as compared to the expected value minimizing algorithms. It is also evident that the second-order CPT-SPSA-N gives marginally better results than its first-order counterpart CPT-SPSA. Finally, what isn't shown is that the CPT-value obtained for CPT-SPSA/CPT-SPSA-N is much lower than that obtained for an SPSA based algorithm from [10] that optimizes expected value, thus making apparent the need for specialized algorithms that incorporate CPT-based criteria.

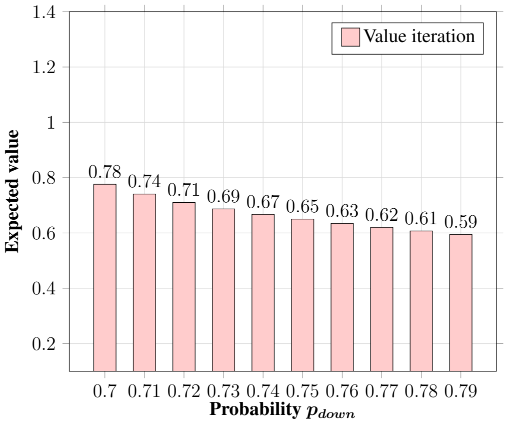


Figure 4.5: Value iteration

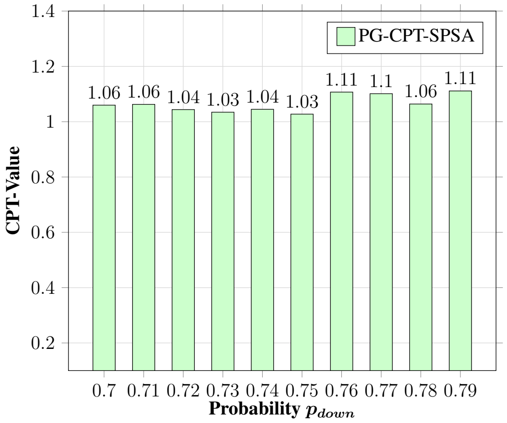


Figure 4.6: First-order SPSA for CPT-value

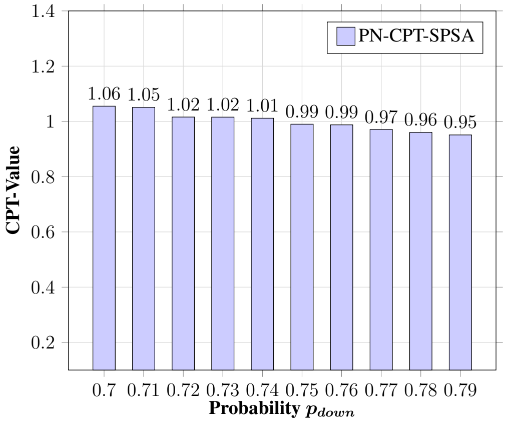


Figure 4.7: Second-order SPSA for CPT-value

4.4 Traffic Control Simulation

We consider a traffic signal control application where the aim is to improve the road user experience by an adaptive traffic light control (TLC) algorithm. We optimize the CPT-value of the delay experienced by road users, since CPT realistically captures the attitude of the road users towards delays. We then optimize the CPT-value of the delay and contrast this approach with traditional expected delay minimizing algorithms. It is assumed that the CPT functional's parameters (u, w) are given (usually, these are obtained by observing human behavior). The experiments are performed using the GLD traffic simulator [83], and the implementation is available at `https://bitbucket.org/prashla/rl-gld`.

4.4.1 Simulation Setup

We consider a road network with $\mathcal{N}$ signalled lanes that are spread across junctions and $\mathcal{M}$ paths, where each path connects (uniquely) two edge nodes, from which the traffic is generated (see Figure 4.8).

At any instant n, let q_n^i and t_n^i denote the queue length and elapsed time since the lane turned red, for any lane $i = 1, \ldots, \mathcal{N}$. Let $d_n^{i,j}$ denote the delay experienced by jth road user on ith path, for any $i = 1, \ldots, \mathcal{M}$ and $j = 1, \ldots, n_i$, where n_i denotes the number of road users on path i. We specify the various components of the traffic control MDP below. The state $s_n = (q_n^1, \ldots, q_n^{\mathcal{N}}, t_n^1, \ldots, t_n^{\mathcal{N}}, d_n^{1,1}, \ldots, d_n^{\mathcal{M}, n_{\mathcal{M}}})^\intercal$ is a vector of lane-wise queue lengths, elapsed times and pathwise delays. The actions are the feasible traffic signal configurations.

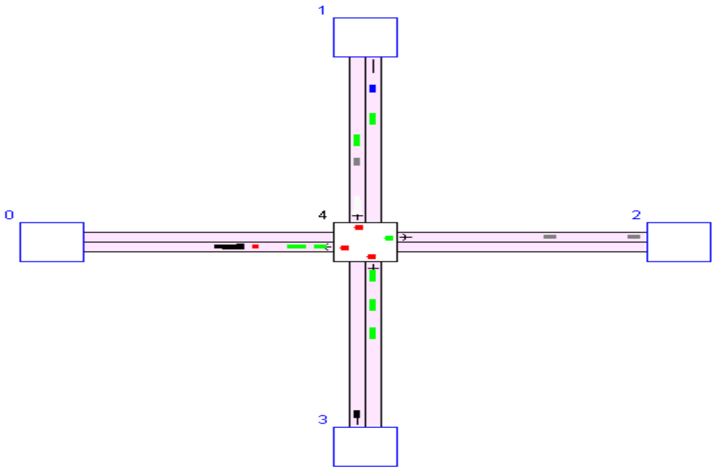


Figure 4.8: Snapshot of the road network from the GLD simulator. The figure shows four edge nodes that generate traffic, one traffic light and two-laned roads carrying automobiles.

We consider Boltzmann policies that have the form

$$\pi_\theta(s,a) = \frac{e^{\theta^\top \phi_{s,a}}}{\sum_{a' \in \mathcal{A}(s)} e^{\theta^\top \phi_{s,a'}}}, \;\; \forall s \in \mathcal{S}, \; \forall a \in \mathcal{A}(s),$$

with features $\phi_{s,a}$ as described in Section V-B of [61].

We consider two different notions of return as follows:

CPT: For any policy θ, let X_i^θ be the delay r.v. and μ_i^θ the proportion of road users along path i, for $i = 1, \ldots, \mathcal{M}$. Any road user along path i will evaluate the delay (s)he experiences in a manner that is captured well by CPT. An important component of CPT is to employ a reference point to calculate gains and losses. Choosing a suitable reference point is challenging, but [78] advocates using the status-quo as the reference point.

With the objective of maximizing the experience of road users across paths, the overall return to be optimized is given by

$$\max_{\theta\in\Theta} \text{CPT}(X_1^\theta, \dots, X_{\mathcal{M}}^\theta) = \sum_{i=1}^{\mathcal{M}} \mu_i^\theta \mathbb{C}(B_i - X_i^\theta), \tag{4.1}$$

where Θ is the d-dimensional hypercube formed by intervals $[0.1, 1.0]$ in each dimension. The rationale behind the objective above is that CPT-value $\mathbb{C}(B_i - X_i^\theta)$ would capture the road user experience/satisfaction for each path i and the goal is to maximize the *average satisfaction* over all paths. In our setting, we use pathwise delays, say B_i for path i, obtained from a pre-timed TLC (cf. the Fixed TLCs in [60]) as the reference point. If the delay of any TLC algorithm is less than that of pre-timed TLC, then the (positive) difference in delays is perceived as a gain and in the complementary case, the delay difference is perceived as a loss. Thus, the CPT-value $\mathbb{C}(B_i - X_i)$ for any path i in Figure 4.8 is to be understood as a *differential delay gain* w.r.t. B_i.

AVG: For the sake of comparison, we consider the traditional objective of minimizing the overall average delay, i.e.,

$$\min_{\theta\in\Theta} \text{AVG}(X_1^\theta, \dots, X_{\mathcal{M}}^\theta) = \sum_{i=1}^{\mathcal{M}} \mu_i^\theta \mathbb{E}(X_i^\theta). \tag{4.2}$$

In comparison to CPT objective, the above does not incorporate baseline delays, makes no distinction between gains and losses via utility functions and does not distort probabilities.

We implement the following TLC algorithms:

CPT-SPSA: This is a first-order algorithm that solves (4.1) using SPSA-based gradient estimates and Algorithm 1 for estimating CPT-value $\mathbb{C}(B_i - X_i)$ for each path $i = 1, \dots, \mathcal{M}$, with $d_n^{i,j}, j = 1, \dots, n_i$ as the samples.

AVG-SPSA: This is SPSA-based first-order algorithm that solves (4.2), while using sample averages of the delays to estimate the expected delay $\mathbb{E}(X_i)$ for each path $i = 1, \dots, \mathcal{M}$.

Table 4.1: AVG and CPT-value estimates for AVG-SPSA and CPT-SPSA.

	AVG-value	**CPT-value**
AVG-SPSA	$\mathbf{111.67}$	53.31
CPT-SPSA	116.21	$\mathbf{59.91}$

The underlying CPT-value $\mathbb{C}(X_i)$ follows the exact form as in Section 4.2, except here we set $\lambda = 2.25$. The choices for λ, σ, η_1 and η_2 are based on median estimates given by [78] and have been used earlier in a traffic application (see [34]). For all the algorithms, motivated by standard guidelines (see [74]), we set $\delta_n = 1.9/n^{0.101}$ and $a_n = 1/(n+50)$. The initial point θ_0 is the d-dimensional vector of ones and $\forall i$, the operator Γ_i keeps the iterate θ_i within $[0.1, 1.0]$.

The experiments involve two phases: first, a training phase where we run each algorithm for 500 iterations, with each iteration involving two perturbed simulations. Each simulation involves running the traffic simulator with a fixed policy parameter for 5000 steps, and this corresponds to approximately 4000 delay samples. The training phase is followed by a test phase where we fix the policy obtained at the end of training and then run the traffic simulator with the aforementioned parameter for 5000 steps. The results presented are averages over ten independent simulations.

4.4.2 Results

Table 4.1 presents the overall AVG and CPT-values for AVG-SPSA and CPT-SPSA, while Figures 4.11 and 4.12 present the expected delay and CPT of differential delay for each of the 12 paths in Figure 4.8.

It is evident that each algorithm converges to a different policy and the difference, in ℓ_1 norm, between policy parameters obtained at the end of training phase for AVG-SPSA and CPT-SPSA was observed to be 6.51.

As shown in Table 4.1, AVG-SPSA results in a TLC policy with lower expected delay, while CPT-SPSA's policy has higher CPT-value. This is expected because AVG-SPSA uses neither utilities nor probability distortions and minimizes overall delay, while CPT-SPSA uses a pre-timed TLC baseline and treats delay gains and losses differently.

Further, Figures 4.9 and 4.10 present the histogram of the delays for the path from 0 to 1, we observe that CPT-SPSA results in a strategy that avoids high delays at the cost of a slightly higher average delay, whereas AVG-SPSA occasionally incurs delays significantly larger than the average delay.

From Figures 4.11 to 4.12, we observe that CPT-SPSA gives significantly better CPT-value at the path $0-1$, $0-3$, $1-3$, while on the remaining paths, the CPT-value of CPT-SPSA and AVG-SPSA are comparable. From an expected delay viewpoint, AVG-SPSA exhibits lower expected delay on paths $0-1$, $0-2$, $0-3$, $3-0$. while on the remaining paths the expected delay of AVG-SPSA and CPT-SPSA are comparable.

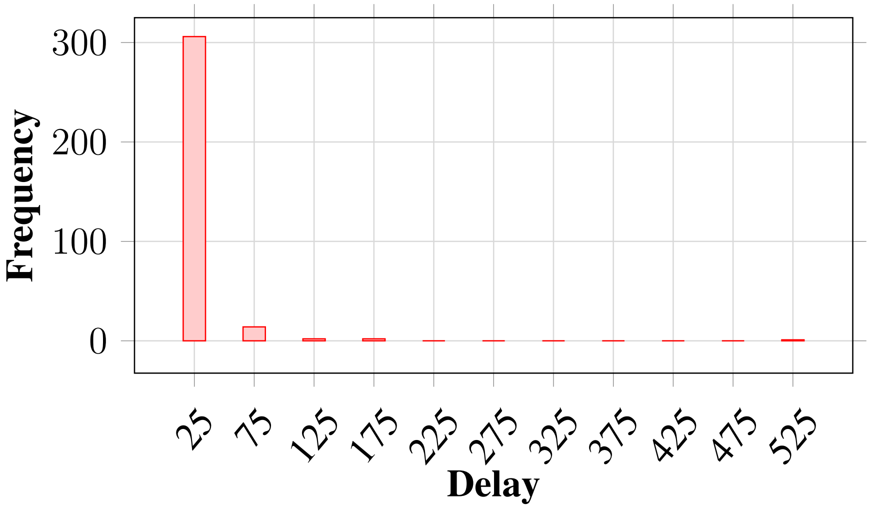


Figure 4.9: Histogram of the sample delays for the path from node 0 to 1 for AVG-SPSA that minimizes overall expected delay

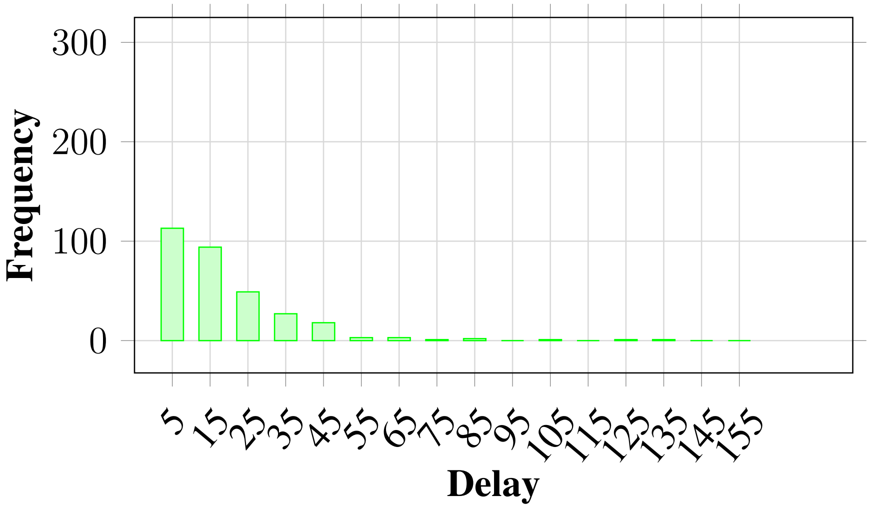


Figure 4.10: Histogram of the sample delays for the path from node 0 to 1 for CPT-SPSA that maximizes CPT-value of differential delay

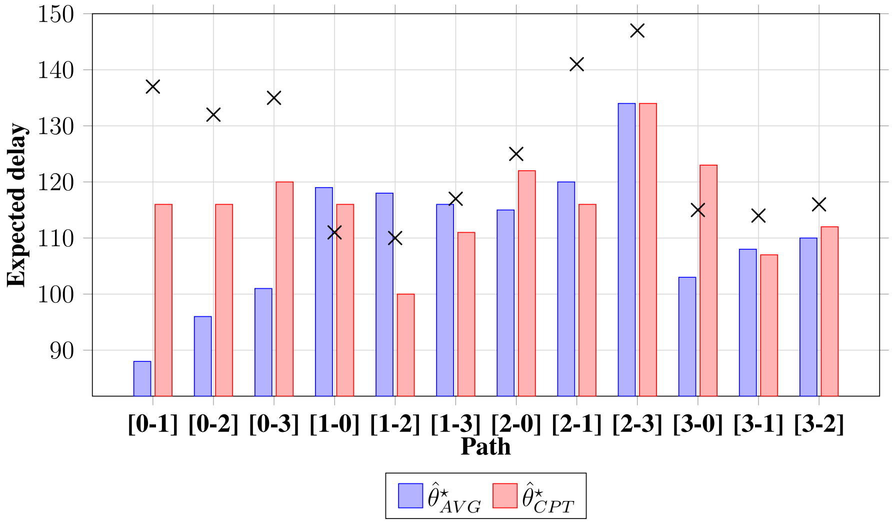


Figure 4.11: Expected delay (path-wise). The cross indicate baseline delays

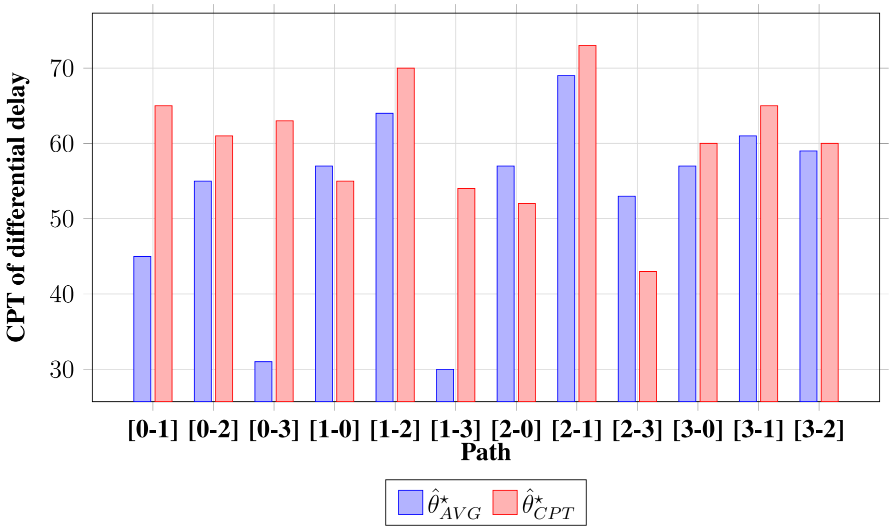


Figure 4.12: CPT of differential delay (pathwise)

Chapter 5

Discussions and Future Work

5.1 How should we apply CPT measure?

Chapter 2 and Chapter 3 present two decision-making frameworks where CPT measure is applied differently. For a given MDP, we evaluate the CPT-measure of its accumulated reward in Chapter 2, while we introduce CPT-measure on each time step of the MDP in Chapter 3. A natural question is raised: How shall we choose between the two frameworks in practice?

The choice between the two frameworks depends on specific decision-making problem settings one tries to solve in practice. For instance, in the problem setting of portfolio choice, a financial institution tends to be interested in changes in the portfolio positions at the end of each trading day and reassesses the portfolio prices once the changes occur [64]. In this setting, CPT-driven dynamic programming framework may be well fitted to model the choice of the financial institution. However, in the Traffic Light Control (TLC) experiment illustrated in Section 4.4, the only metric a road user cares about is the delay time accumulated in the entire path. Therefore, the CPT-value of the total delay experienced by a road user may be an appropriate performance measure to model users' choice in the traffic system.

There is no principle which can guide one to make the right choice at every situation. However, it is usually helpful for one to determine whether his decision is only based on the final accumulated outcome or he needs to derive a performance measure on the random observations at each time step.

5.2 Future Work

Future research can be developed in a few aspects. First, we can investigate the asymptotic normality property of the CPT-value estimator derived in Section 2.5. To this end, a potential approach is to make use of the properties of order statistics. Indeed, the estimator $\overline{\mathbb{C}}_n$ in algorithm 1 is essentially a weighted sum of order statistics, which is termed as a“L-estimate” in statistical literature. The book by Serfling ([69]) as well as the related papers (Stigler [75], Shorack [70], etc) presented asymptotic normality of L-estimates under various restrictions on the weights and the related random variable. Generally, one requires that the extremal order statistics (i.e., the minimum and the maximum of the samples) do not contribute much to the L-estimate. The asymptotic normality of $\overline{\mathbb{C}}_n$ could help us to obtain a better sample complexity result of the estimator as compared to Theorem 3, and to study further on the sample complexity of our optimization algorithms.

In Chapter 3, we propose a CPT-value driven dynamic programming framework and examine its contraction and monotonicity properties. In the future, we would like to develop function approximation schemes to find optimal policies of the CPT dynamic programming problems within a reasonable amount of computational effort. Further, it is interesting to look into the cases where we only have partial information of the dynamic programming structure. Under such cases, some variations of reinforcement learning algorithms (TD-learning, Q-learning etc, [76]) may help us to solve the problems.

Appendix A

Minimax lower bound and LeCam Method

A minimax lower bound is a widely used measure on the performance of classical estimation problems, and minimax bounds can also be used in studying optimization problems. In this chapter we review a classical technique for deriving a minimax lower bound that have proven useful in a variety of problems [85].

A.1 Basic framework of minimax risk

Let us begin by defining the standard minimax risk. We let $\mathcal{P}$ denote a class of probability distributions on a sample space $\mathcal{X}$, and let $\theta : \mathcal{P} \to \Theta$ denote a function defined on $\mathcal{P}$.

For a distribution $P \in \mathcal{P}$, we want to estimate the unknown parameter $\theta(P) \in \Theta$ given i.i.d samples X_i from P. An estimator is denoted as $\hat{\theta}$ and is a function from $\mathcal{X} \to \Theta$. We evaluate the quality of the estimator in terms of the risk

$$\mathbb{E}\left[\Phi\left(\rho\left(\hat{\theta}\left(X_1, \ldots, X_n\right), \theta\left(P\right)\right)\right)\right],$$

where Φ denotes the risk function and ρ is a predefined metric on Θ.

Since designing an estimator on each singleton $P \in \mathcal{P}$ is not practical, it is important to consider the risk functional in a global sense over Θ. One approach, suggested by Wald [81], is to choose the estimator $\hat{\theta}$ minimizing the maximum risk

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\Phi \left(\rho \left(\hat{\theta} \left(X_1, \ldots, X_n \right), \theta \left(P \right) \right) \right) \right].$$

An optimal estimator for this metric then gives the *minimax risk*, which is defined as

$$\mathcal{M} \left(\theta \left(\mathcal{P} \right), \Phi \circ \rho \right) := \inf_{\hat{\theta}} \sup_{P\in\mathcal{P}} \mathbb{E}_P \left[\Phi \left(\rho \left(\hat{\theta} \left(X_1, \ldots, X_n \right), \theta \left(P \right) \right) \right) \right]. \tag{A.1}$$

A.2 Statistical test and estimation lower bound

It can be shown that estimation risk can be lower bounded by the probability of error in testing problems [84]. To see it, we formulate the statistical testing problem as follows: Given an index set $\mathcal{V}$ of finite cardinality, consider a family of distributions $\{P_v\}_{v\in\mathcal{V}} \subset \mathcal{P}$. The family $\{P_v\}_{v\in\mathcal{V}}$ is associated with a set of parameters $\{\theta \left(P_v \right)\}_{v\in\mathcal{V}}$. We define a $\rho-$semimetric such that

$$\rho \left(\theta \left(P_v \right), \theta \left(P_{v'} \right) \right) \geq 2\delta \ \ \forall v \neq v'.$$

Given a vector of sample $X = (X_1, \ldots, X_n)$ with X_i i.i.d. and drawn from one distribution in $\{P_v\}_{v\in\mathcal{V}}$, we wish to find the value of the underlying index v. To this end, we establish a measurable mapping $\Psi : \mathcal{X}^n \to \mathcal{V}$ as a test function. The error probability associated with Ψ is $\mathbb{P} \left(\Psi \left(X_1^n \right) \neq V \right)$, where $\mathbb{P}$ is jointly determined by the probability measure on the random index V and X. The following proposition, proved in [84], presents a classical relationship between estimation and testing.

Proposition 1. *The minimax error defined in Section A.1 has lower bound*

$$\mathcal{M}_n \left(\theta \left(\mathcal{P} \right), \Phi \circ \rho \right) \geq \Phi \left(\delta \right) \inf_{\Psi} \mathbb{P} \left(\Psi \left(X_1, \ldots, X_n \right) \neq V \right),$$

where the infimum ranges over all testing functions.

A.3 Le Cam Method

The Le Cam method provides lower bounds on the error in simple binary hypothesis testing problems. Suppose that we have a Bayesian hypothesis testing problem where V is chosen with equal probability to be 1 or 2, and given $V = v$, the sample X is drawn from the distribution P_v. We have for any test $\Psi : \mathcal{X} \to \{1, 2\}$, the probability of error is

$$\mathbb{P}\left(\Psi\left(X \neq V\right)\right) = \frac{1}{2}P\left(\Psi\left(X\right) \neq 1\right) + \frac{1}{2}P\left(\Psi\left(X\right) \neq 2\right). \tag{A.2}$$

A standard result of Le Cam, which builds the relationship between total variation and testing error, is the following lemma:

Lemma 11. *For any distributions P_1 and P_2 on $\mathcal{X}$,*

$$\inf_{\Psi}\{P_1\left(\Psi\left(X\right) \neq 1\right) + P_2\left(\Psi\left(X\right) \neq 2\right)\} = 1 - \left\|P_1, P_2\right\|_{\mathrm{TV}}, \tag{A.3}$$

where the infimum is taken over all tests $\Psi : \mathcal{X} \to \{1, 2\}$, and the total variation operator $\left\|P_1, P_2\right\|_{\mathrm{TV}}$ is defined in definition 2 at Section 2.5.3.

Proof. Denote A as the set that $\Psi\left(X\right) = 1$, and A^c stands for the set that $\Psi\left(X\right) = 2$. Naturally,

$$P_1\left(\Psi\left(X\right) \neq 1\right) + P_2\left(\Psi\left(X\right) \neq 2\right) = P_1\left(A^c\right) + P_2\left(A\right) = 1 - P_1\left(A\right) + P_2\left(A\right).$$

Taking the infimum over all possible tests Ψ, we have

$$\inf_{\Psi}\{P_1\left(\Psi\left(X\right) \neq 1\right) + P_2\left(\Psi\left(X\right) \neq 2\right)\} = 1 - \sup_{A \subset \mathcal{X}}\left(P_1\left(A\right) - P_2\left(A\right)\right),$$

which yields (A.3). □

Revisiting the setting where we have n i.i.d. samples X_i, and V has equal probability to be 1 or 2, the probability of test error then satisfies

$$\inf_{\Psi} \mathbb{P}\left(\Psi\left(X_1, \ldots, X_n\right) \neq V\right) = \frac{1}{2} - \frac{1}{2}\left\|P_1^n, P_2^n\right\|_{\mathrm{TV}}. \tag{A.4}$$

The expressions (A.4) and (A.3), together with Proposition 1, imply the following proposition of the lower bound on minimax risk:

Proposition 2. *For any family $\mathcal{P}$ of distributions for which there exists a pair $P_1, P_2 \in \mathcal{P}$ satisfying $\rho\left(\theta\left(P_1\right), \theta\left(P_2\right)\right) \geq 2\delta$, the minimax risk after n observations has lower bound*

$$\mathcal{M}\left(\theta\left(\mathcal{P}\right), \Phi \circ \rho\right) \geq \Phi\left(\delta\right)\left[\frac{1}{2} - \frac{1}{2}\left\|P_1^n, P_2^n\right\|_{\mathrm{TV}}\right].$$

Appendix B

Empirical distribution

The empirical distribution function can be used to approximate the underlying CDF based on which the samples are generated. The function's value at a specific point is the fraction of observations less than or equal to the value of that point.

B.1 Definition and fundamental theorems

We begin with problem of estimating a CDF. Let $(U_1, \ldots U_n) \sim F$ where $F(u) = P\left(U \leq u\right)$ is a distribution function on the real line. The empirical function is defined as follows:

Definition 4. *The* ***empirical distribution function*** $\hat{F_n}$ *is the CDF that puts mass* $1/n$ *at each data point* U_i*. Formally,*

$$\hat{F_n}(u) = \frac{1}{n}\sum_{i=1}^{n} I\left(U_i \leq u\right) \tag{B.1}$$

where

$$I\left(U_i \leq u\right) = \begin{cases} 1 & \textit{if} \;\; U_i \leq u \\ \\ 0 & \textit{if} \;\; U_i \geq u. \end{cases}$$

Some classical theorems are summarized below:

Theorem 18. *Let* $(U_1, \ldots U_n) \sim F$ *and let* $\hat{F_n}$ *denote the empirical CDF defined in* (B.1). *Then we have:*

1. *At any fixed value of* u*,* $\mathbb{E}\left(\hat{F}_n(u)\right) = F\left(u\right)$ *and* $\mathbb{V}\left(\hat{F}_n(u)\right) = \frac{F(x)(1-F(x))}{n}$*.*

   Therefore, the mean square error equals $\frac{F(x)(1-F(x))}{n} \to 0$ *and* $\hat{F}_n \to F\left(x\right)$ *w.p.1.*

2. ***(Glivenko-Cantelli Theorem)****.*

$$\sup_x \left|\hat{F}_n\left(x\right) - F\left(x\right)\right| \to 0 \text{ a.s.}$$

3. ***(Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)****.*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |\hat{F}_n(x) - F(x)| > \epsilon\right) \leq 2e^{-2n\epsilon^2} \ \ \forall\epsilon > 0.$$

Remark 7. *From the DKW inequality, we can construct a confidence set. Let* $\epsilon_n^2 = \log\left(2/\alpha\right)/\left(2n\right)$*,* $L\left(x\right) = \max\{\hat{F}_n\left(x\right) - \epsilon_n, 0\}$ *and* $U\left(x\right) = \min\{\hat{F}_n\left(x\right) + \epsilon_n, 1\}$*. It follows that for any CDF* F*,*

$$P\left(L\left(x\right) \leq F\left(x\right) \leq U\left(x\right)\forall x\right) \geq 1 - \alpha.$$

Bibliography

[1] M. Allais. Le comportement de l'homme rationel devant le risque: Critique des postulats et axioms de l'ecole americaine. *Econometrica*, 21:503–546, 1953.

[2] K. J. Arrow. *Essays in the Theory of Risk Bearing*. Markham, Chicago, IL, 1971.

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

[4] N. C. Barberis. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, pages 173–196, 2013.

[5] R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716–719, 1952.

[6] R. Bellman. *Applied Dynamic Programming*. Princeton University Press, Princeton, 1957.

[7] D. P. Bertsekas. *Dynamic Programming and Optimal Control, vol. II, 3rd edition*. Athena Scientific, 2007.

[8] D. P. Bertsekas. *Abstract Dynamic Programming*. Athena Scientific, Belmont, MA, 2013.

[9] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[10] S. Bhatnagar and S. Kumar. A simultaneous perturbation stochastic approximation-based actor-critic algorithm for Markov decision processes. *IEEE Transactions on Automatic Control*, 49(4):592–598, 2004.

[11] S. Bhatnagar, H. L. Prasad, and L. A. Prashanth. *Stochastic Recursive Algorithms for Optimization*, volume 434. Springer, 2013.

[12] S. Bhatnagar and L. A. Prashanth. Simultaneous perturbation Newton algorithms for simulation optimization. *Journal of Optimization Theory and Applications*, 164(2):621–643, 2015.

[13] P. Billingsley. *Probability and Measure*. John Wiley and Sons, 2nd edition, 1995.

[14] V. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[15] V. Borkar and R. Jain. Risk-constrained Markov decision processes. In *IEEE Conference on Decision and Control*, pages 2664–2669, 2010.

[16] V. S. Borkar. *Topics in Controlled Markov Chains*. CRC Press, 1991.

[17] X. Cao and X. Wan. Sensitivity analysis of nonlinear behavior with distorted probability. *Mathematical Finance*, 2014.

[18] M. Cary, A. Das, B. Edelman, I. Giotis, K. Heimerl, A. R. Karlin, C. Mathieu, and M. Schwarz. Greedy bidding strategies for keyword auctions. *Proceedings of the 8th ACM conference on Electronic commerce - EC 07*, 2007.

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

[20] H. S. Chang, J. Hu, M. C. Fu, and S. I. Marcus. *Simulation-based Algorithms for Markov Decision Processes*. Springer, 2nd edition, 2013.

[21] P. Cheridito, F. Delbaen, and M. Kupper. Coherent and convex monetary risk measures for bounded cádlág processes. *Stochastic Processes and their Applications*, 112(1):1–22, 2004.

[22] Y. Chow and M. Ghavamzadeh. Algorithms for CVaR optimization in MDPs. In *Advances in Neural Information Processing Systems*, pages 3509–3517, 2014.

[23] K. J. Chung and M. J. Sobel. Discounted MDP's: distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

[24] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[25] F. Delbaen and E. T. Hochschule. Coherent risk measures on general probability spaces. In *Essays in Honor of Dieter Sondermann*, pages 1–37. Springer, 2002.

[26] J. C. Duchi, M. I. Jordan, M. J. Wainwright, and A. Wibisono. Optimal rates for zero-order convex optimization: the power of two function evaluations. *arXiv preprint arXiv:1312.2139*, 2013.

[27] P. Dupuis, M. R. James, and I. Petersen. Robust properties of risk-sensitive control. *Mathematics of Control, Signals, and Systems*, 13(4):318–332, 2000.

[28] D. Ellsberg. Risk, ambiguity and the Savage's axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.

[29] M. Fathi and N. Frikha. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *Electronic Journal of Probability*, 18(67):1–36, 2013.

[30] W. Feller. *An Introduction to Probability Theory and Its Applications*, volume 1. Wiley, January 1968.

[31] H. Fennema and P. Wakker. Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making*, 10:53–64, 1997.

[32] J. Filar, L. Kallenberg, and H. Lee. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.

[33] M. C. Fu. Stochastic gradient estimation. In *Handbook of Simulation Optimization*, pages 105–147. Springer, 2015.

[34] S. Gao, E. Frejinger, and M. Ben-Akiva. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportation Research Part C: Emerging Technologies*, 18(5):727–740, 2010.

[35] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization*. Academic Press, 1981.

[36] A. Gopalan, L.A. Prashanth, M. C. Fu, and S. I. Marcus. Weighted bandits or: How bandits learn distorted values that are not expected. In *AAAI Conference on Artificial Intelligence*, pages 1941–1947, 2017.

[37] E. Hazan. *Introduction to Online Convex Optimization*. Now Publishers, 2016.

[38] X. He and X. Zhou. Portfolio choice via quantiles. *Mathematical Finance*, 21(2):203–231, 2011.

[39] P. Heidelberger, X. Cao, M. A. Zazanis, and R. Suri. Convergence properties of infinitesimal perturbation analysis estimates. *Management Science*, 34:1281–1302, 1988.

[40] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Encyclopedia of Statistical Sciences*, 2004.

[41] L. J. Hong. Estimating quantile sensitivities. *Operations Research*, 57:118–130, 2009.

[42] R. A. Howard and J. E. Matheson. Risk-sensitive Markov decision processes. *Management Science*, 18(7):356–369, March 1972.

[43] H. W. James and E. J. Collins. An analysis of transient Markov decision processes. *Journal of Applied Probability*, 43(3):603–621, 2006.

[44] C. Jie, L. A. Prashanth, M. C. Fu, S. I. Marcus, and C. Szepesvari. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*. accepted for publication.

[45] A. Jobert and L. C. G. Rogers. Valuations and dynamic convex risk measures. *Mathematical Finance*, 18(1):1–22, 2007.

[46] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pages 263–291, 1979.

[47] B. Kitts and B. Leblanc. Optimal bidding on keyword auctions. *Electronic Markets*, 14(3):186–201, 2004.

[48] M. J. Kochenderfer. *Decision Making Under Uncertainty: Theory and Application*. The MIT Press, 2015.

[49] H. Kushner and D. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.

[50] H. Kushner and G. G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer, 2003.

[51] P. L'Ecuyer and G. Perron. On the convergence rates of IPA and FDC derivative estimators. *Operations Research*, 42(4):643–656, 1994.

[52] K. Lin. *Stochastic Systems with Cumulative Prospect Theory*. PhD thesis, University of Maryland, August 2013.

[53] K. Lin, C. Jie, and S. I. Marcus. Probabilistically distorted risk-sensitive infinite-horizon dynamic programming. *IEEE Automatica*. submitted for review.

[54] S. Mannor and J. N. Tsitsiklis. Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.

[55] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7(1):77–91, 1952.

[56] J. V. Neumann and O. Morgenstern. *Theory of Games and Economic Behavior (Commemorative Edition)*. Princeton University Press, March 2007.

[57] S. R. Pliska. Dynamic programming and its applications. In Martin L. Puterman, editor, *Markov Decision Processes Discrete Stochastic Dynamic Programming*, chapter 10, pages 335–349. Academic Press, 1978.

[58] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[59] L. A. Prashanth. Policy gradients for CVaR-constrained MDPs. In *Algorithmic Learning Theory*, pages 155–169, 2014.

[60] L. A. Prashanth and S. Bhatnagar. Reinforcement learning with function approximation for traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):412–421, 2011.

[61] L. A. Prashanth and S. Bhatnagar. Threshold tuning using stochastic optimization for graded signal control. *IEEE Transactions on Vehicular Technology*, 61(9):3865–3880, 2012.

[62] L. A. Prashanth, C. Jie, M. C. Fu, S. I. Marcus, and C. Szepesvári. Cumulative prospect theory meets reinforcement learning: Prediction and control. In *International Conference on Machine Learning*, pages 1406–1415, 2016.

[63] D. Prelec. The probability weighting function. *Econometrica*, pages 497–527, 1998.

[64] F. Riedel. Dynamic coherent risk measures. *Stochastic Processes and their Applications*, 112(2):185–200, August 2004.

[65] R. T. Rockafellar and S. Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[66] D. Ruppert. Stochastic approximation. *Handbook of Sequential Analysis*, pages 503–529, 1991.

[67] A. Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010.

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

[69] R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, 2002.

[70] G. R. Shorack. Asymptotic normality of linear combinations of functions of order statistics. *The Annals of Mathematical Statistics*, 40(6):2041âĂŞ2050, 1969.

[71] M. Sobel. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pages 794–802, 1982.

[72] J. C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Auto. Cont.*, 37(3):332–341, 1992.

[73] J. C. Spall. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Autom. Contr.*, 45:1839–1853, 2000.

[74] J. C. Spall. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65. John Wiley & Sons, 2005.

[75] S. M. Stigler. Linear functions of order statistics with smooth weight functions. *The Annals of Statistics*, 2(4):676âĂŞ693, 1974.

[76] R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

[77] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press, 1st edition, 1998.

[78] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[79] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[80] P. P. Wakker. *Prospect Theory: For Risk and Ambiguity*. Cambridge University Press, July 2010.

[81] A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.*, 10(4):299–326, 12 1939.

[82] L. A. Wasserman. *All of Nonparametric Statistics*. Springer, 2015.

[83] M. Wiering, J. Vreeken, J. van Veenen, and A. Koopman. Simulation and optimization of traffic in a city. In *IEEE Intelligent Vehicles Symposium*, pages 453–458, June 2004.

[84] Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5):1564–1599, 1999.

[85] B. Yu, F. Assouad, and L. Cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer, 1997.