Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control

Prashanth L.A.*[1], Cheng Jie†[2], Michael Fu‡[3], Steve Marcus§[4] and Csaba Szepesvári¶[5]

[1]Institute for Systems Research, University of Maryland
[2]Department of Mathematics, University of Maryland
[3]Robert H. Smith School of Business & Institute for Systems Research, University of Maryland
[4]Department of Electrical and Computer Engineering & Institute for Systems Research, University of Maryland
[5]Department of Computing Science, University of Alberta

Abstract

Cumulative prospect theory (CPT) is known to model human decisions well, with substantial empirical evidence supporting this claim. CPT works by distorting probabilities and is more general than the classic expected utility and coherent risk measures. We bring this idea to a risk-sensitive reinforcement learning (RL) setting and design algorithms for both estimation and control. The RL setting presents two particular challenges when CPT is applied: estimating the CPT objective requires estimations of the *entire distribution* of the value function and finding a *randomized* optimal policy. The estimation scheme that we propose uses the empirical distribution to estimate the CPT-value of a random variable. We then use this scheme in the inner loop of policy optimization procedures for a stochastic shortest path problem. We propose both gradient-based as well as gradient-free policy optimization algorithms. The former includes both first-order and second-order methods that are based on the well-known simulation optimization idea of simultaneous perturbation stochastic approximation (SPSA), while the latter is based on a reference distribution that concentrates on the global optima. Using an empirical distribution over the policy space in conjunction with Kullback-Leibler (KL) divergence to the reference distribution, we get a global policy optimization scheme. We provide theoretical convergence guarantees for all the proposed algorithms and also empirically demonstrate the usefulness of our algorithms.

1 Introduction

In this paper we consider human-based decision and more specifically reinforcement learning (RL) problems where the reinforcement learning agent controls a system to produce outcomes ("rewards") that are maximally aligned with the preferences of one or multiple humans, an arrangement shown on Figure 1. To support this arrangement, one possibility is to model human preferences with the help of some *risk metrics* mapping random returns (e.g., the total discounted reward) to some scalar deterministic quantity. Popular approaches that use such risk metrics include the exponential utility formulation (cf. **?**) that implicitly controls the variance. An alternative is a to consider constrained formulations with explicit constraints on the

*prashla@isr.umd.edu
†cjie@math.umd.edu
‡mfu@isr.umd.edu
§marcus@umd.edu
¶szepesva@cs.ualberta.ca

variance of the return (cf. **??**). Another constraint alternative is to bound a coherent risk measure such as Conditional Value-at-Risk (CVaR), while minimizing the usual cost objective (cf. **??**).

The risk metrics underlying the above-mentioned works are based on the assumption that human decision makers are rational and/or consistent. While this may hold in certain restricted settings, a large body of literature indicates that humans are neither rational, nor consistent (which, in fact, is an unsurprising fact, at least in the experience of the authors of the paper). In other words, traditional approaches are based on the belief that optimizing the expected utility (EU) is appealing for human subjects. However, there is substantial evidence that this is not case - see the survey article **?** and Chapter 4 of the book **?**. In particular, the aforementioned references describe the Allais and Ellsberg paradoxes popular among economists for arguing against EU. Thus, if the goal is to produce outcomes that are best aligned with human preferences, an alternative approach is required. A singularly popular and successful approach in behavioral science and economics is based on *prospect theory (PT)* **?** and its later enhancement, the so-called *cumulative prospect theory* (CPT) **?**. CPT is a rank dependent expected utility model **?** that incorporates decision weights to distort probabilities. The suitability of this approach to model human decision making (and thus preferences) has been widely documented **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**. PT/CPT has been applied in a variety of domains, for e.g., healthcare **?**, seismic design **?**, transportation **?**,**?**, **?**, online auctions **?**, insurance **?** and finance **?**, **?**, **?**.

Cs: Add literature supporting this. At least three books:)

Cs: Again, add some books. Actually, the examples that I am reading about, e.g., in the book "Against the Gods", chapter 13 by Bernstein, suggest that some big exploits are possible. And most examples are about that the framing of a decision problem influences what humans choose. This is kinda dirty. I can totally see how policy makers can exploit this. Is this good? Can we have some positive (in the sense of being ethical) examples? E.g., you mentioned the arrangement of shelves in shops or something. Some examples will be badly needed. E.g., google on prospect theory decision making gives me http://link.springer.com/article/10.1007%2Fs11424-015-204 Perhaps add some to the appendix as background on CPT.

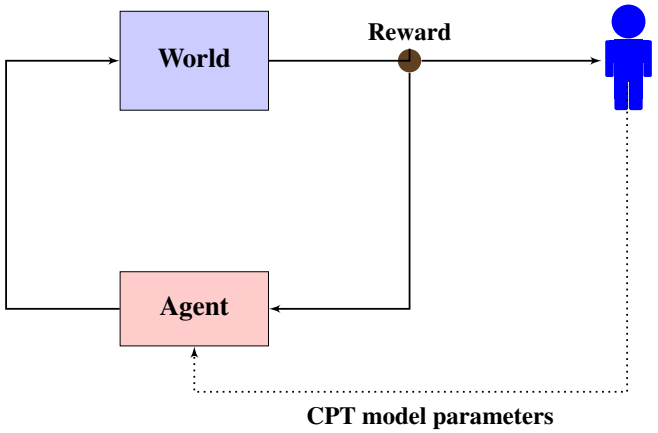


Figure 1: Operational flow of a human-based decision making system

As illustrated in Figure 1, we consider a typical RL setting where the environment is unknown, but can be experimented with and propose a CPT based risk metric as the long-term performance objective. CPT is a non-coherent and non-convex measure that is well known among psychologists and economists to be a good model for human decision-making systems, with strong empirical support. To put it differently, CPT captures well the way humans evaluate outcomes and hence, we offer a CPT-variant of the RL notion of "value function". Unlike the regular value function which is the expectation of the return random variable, CPT-value employs a functional that distorts the underlying probabilities. The latter is achieved by fitting CPT model parameters to capture human preferences. The goal then for the learning system is to find a policy that maximizes the CPT-value of "return".

Cs: This para may need to be rewritten in light that I rewrote the previous one.

In terms of research contributions, this is the first work to combine CPT with RL, and although on the surface it might seem straightforward, in fact there are many research challenges that arise from trying to apply a CPT objective in the RL framework. We outline these challenges as well as our solution approach below.

Prediction: In the case of the classic value function, which is an expectation, a simple sample means can be used for estimation, facilitating the use of temporal difference type algorithms. On the other hand, estimating

Cs: Risk measure is a technical term according to wikipedia. Risk metric does not seem to have this technical meaning so I propose using risk metric everywhere.

Cs: I find it strange to emphasize only these aspects. What's the goal of announcing these here?

the CPT-value for a given policy is challenging, because CPT-value involves a distribution that is distorted using non-linear weight functions and hence, requires that the *entire* distribution to be estimated.
Solution: We use a quantile based approach to estimate the CPT-value. Assuming that the weight functions are Hölder continuous with constant α, we establish convergence (asymptotic) of our CPT-value estimate to the true CPT-value and provide a sample complexity result that establishes that $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ samples are required to be ϵ-close to the CPT-value with high probability. If the weights are Lipschitz (i.e., $\alpha = 1$), the resulting sample complexity is the canonical rate $O\left(\frac{1}{\epsilon^2}\right)$ for Monte carlo type schemes.
Control: Designing policy optimization algorithms in order to find a *CPT-optimal* policy is challenging because CPT-value is a non-coherent and non-convex risk measure that does not lend itself to dynamic programming approaches such as value/policy iteration due to the lack of a "Bellman equation". Thus, it is necessary to design new simulation optimization scheme that use sample CPT-value estimates to optimize the policy, which is generally *randomized*. While classic simulation optimization settings usually have a zero mean noise in function evaluations, our setting one has to tradeoff simulation cost with the bias in a manner such that the resulting policy optimization scheme cancels the bias effect and converges.
Solution: We derive the condition that specifies the rate at which the number of samples for predicting the CPT-value should increase such that the bias of CPT-value estimates vanishes asymptotically (see (A3) later).

Using two well-known ideas from the *simulation optimization* literature **?**, we propose three optimization algorithms for solving (2). These methods overcome the second and third problems mentioned above and are summarized as follows:
Gradient-based methods: We propose two algorithms in this class. The first is a gradient algorithm that employs simultaneous perturbation stochastic approximation (SPSA)-based estimates for the gradient of the CPT-value, while the second is a Newton algorithm that also uses SPSA-based estimates of the gradient and also the Hessian. We remark again that, unlike traditional settings for SPSA, our estimates for CPT-value have a non-zero (albeit controlled) bias. We establish that our algorithms converge to a locally CPT-value optimal policy.
Gradient-free method: We perform a non-trivial adaptation of the algorithm from **?** to devise a globally CPT-value optimizing scheme. The idea is to use a reference model that eventually concentrates on the global minimum and then empirically approximate this reference distribution well-enough. The latter is achieved via natural exponential families in conjunction with Kullback-Leibler (KL) divergence to measure the "distance" from the reference distribution. Unlike the setting of **?**, we neither observe the objective function (CPT-value) perfectly nor with zero-mean noise. We establish that our algorithm converges to a globally CPT-value optimal parameter (assuming it exists).

To put things in context, risk-sensitive reinforcement learning problems are generally hard to solve. For a discounted MDP, **?** showed that there exists a Bellman equation for the variance of the return, but the underlying Bellman operator is not necessarily monotone. The latter observation rules out policy iteration as a solution approach for variance-constrained MDPs. Further, even if the transition dynamics are known, **?** show that finding a globally mean-variance optimal policy in a discounted MDP is NP-hard. For average reward MDPs, **?** consider a variance definition that measures how far the instantaneous reward is away from its average, unlike the discounted setting where the variance was of the return r.v. However, for average reward MDPs, **?** motivate their variance definition well and then provide NP-hardness results for finding a globally optimal policy with the variance constraint. CVaR as a risk measure is equally (if not more) complicated as the measure here is a conditional expectation, where the conditioning is on a low probability event. Apart from the hardness of finding CVaR-optimal solutions, estimating CVaR for a fixed policy in a typical RL setting itself is a challenge considering CVaR relates to rare events and to the best of our knowledge, there is no algorithm with theoretical guarantees to estimate CVaR without wasting a lot of samples. There are proposals based on importance sampling (cf. **??**), but they lack theoretical guarantees.

In contrast, we derive a *provably* sample-efficient scheme for estimating the CPT-value (see next section for a precise definition) for a given policy and use this as the inner loop in a policy optimization schemes that include gradient-based as well as gradient free approaches. Finally, we point out that the CPT-value that we define is a generalization in the sense that one can recover the regular value function and the risk measures such as VaR and CVaR by appropriate choices of a certain weight function used in the definition of CPT value (see the next section for precise details).

Closest related work is **?**, where the authors propose a CPT-measure for an abstract MDP setting (see **?**). We differ from **?** in several ways: **(i)** We do not have a nested structure for the CPT-value (3) and this implies the lack of a Bellman equation for our CPT measure; and **(ii)** We do not assume model information, i.e., we operate in a model-free RL setting. Moreover, we develop both estimation and control algorithms with convergence guarantees for the CPT-value function.

The rest of the paper is organized as follows: In Section 2, we introduce the notion of CPT-value of a general random variable X and make a special case illustration when X is the return of a stochastic shortest path problem. In Section 3, we describe the empirical distribution based scheme for estimating the CPT-value of any random variable. In Sections 4–5, we present the gradient-based algorithms for optimizing the CPT-value. Next, in Section 6, we present a gradient-free model-based algorithm for CPT-value optimization in an MDP. We provide the proofs of convergence for all the proposed algorithms in Section 7. We present the results from numerical experiments for the CPT-value estimation scheme in Section 8 and finally, provide the concluding remarks in Section 9.

2 CPT-value

For a real-valued random variable X, we first introduce a "CPT-functional" that replaces the traditional expectation. Subsequently, we specialize X to be the return of stochastic shortest path problem.

2.1 General definition

The CPT-value of the random variable X is a functional defined as

$$\mathbb{C}_{u,w}(X) = \int_0^{+\infty} w^+(P(u^+(X) > z)dz - \int_0^{+\infty} w^-(P(u^-(X) > z)dz, \tag{1}$$

where $u = (u^+, u^-)$, $w = (w^+, w^-)$, $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ and $w^+, w^- : [0,1] \to [0,1]$ are continuous (see assumptions (A1)-(A2) in Section 3 for precise requirements on the u and w). For notational convenience, we drop the dependence on u, w and use $\mathbb{C}(X)$ to denote the CPT-value.

Let us deconstruct the above definition:

Utility functions: u^+, u^- are utility functions corresponding to gains ($X \geq 0$) and losses ($X \leq 0$), respectively. For example, consider a scenario where one can either earn $\$500$ 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. 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 - see Fig 2. In contrast, the traditional value function makes no such distinction between gains and losses.

Weight functions: w^+, w^- are functions corresponding to gains and losses, respectively. The main idea is that humans deflate high-probabilities and inflate low-probabilities and this is the rationale behind using a weight function in CPT. For example, humans usually choose a stock that gives one million dollars w.p.

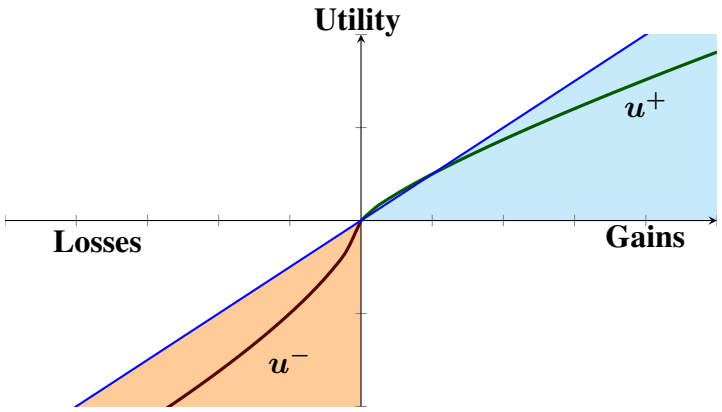


Figure 2: Utility function

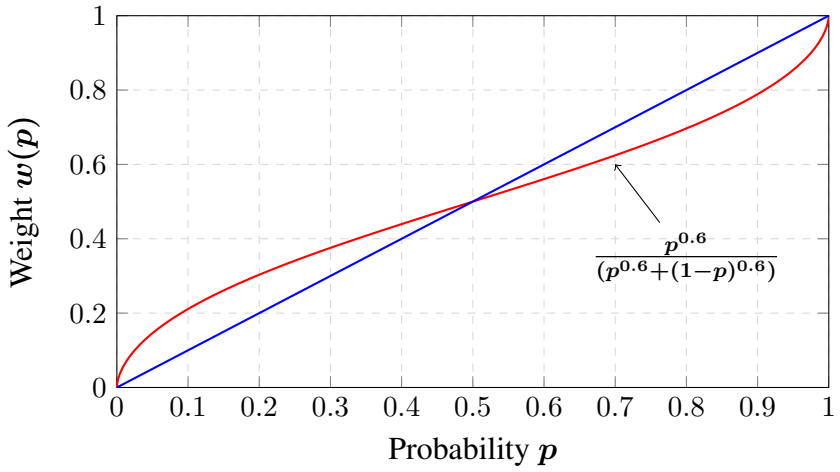


Figure 3: Weight function

$1/10^6$ over one that gives \$1 w.p. 1 and the reverse when signs are flipped. Thus the value seen by the human subject is non-linear in the underlying probabilities - an observation with strong empirical evidence that used human subjects (see **?** or 8000+ papers that follow). In contrast,the traditional value function is linear in the underlying probabilities. As illustrated with $w = w^+ = w^-$ in Fig 3, the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. The authors in **?** recommend $w(p) = \frac{p^\delta}{(p^\delta+(1-p)^\delta)^{1/\delta}}$, while **?** recommends $w(p) = \exp(-(-\ln p)^\delta)$, with $0 < \delta < 1$ and in both cases, the weight function has the inverted-s shape, which is seen to be a good fit from empirical tests on human subjects - see **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**. Weight functions can explain non-linear probability distortions, as illustrated by the following example:

[Stock 1] This investment results in a gain of \$10 with probability (w.p.) 0.1 and a loss of \$500 w.p. 0.9. The expected return is \$$-449$, but this does not necessarily imply that "human" investors' evaluation of the stock is \$$-449$. Instead, it is very likely that the humans evaluate it to a higher value, e.g. \$$-398$ (= gain w.p. 0.2 and loss w.p. 0.8).[1]

[1] See Table 3 in **?** to know why such a human evaluation is likely.

[Stock 2] loss of \$10 w.p. 0.9, gain \$500 w.p. 0.1. Expected return: \$41; Human evaluation: \$92 (= loss w.p. 0.8).

[Stock 3] loss of \$10 w.p. 0.1, gain \$500 w.p. 0.9. Expected return: \$449; Human evaluation: \$398 (= loss w.p. 0.2).

These references also include experimental tests on human subjects and conclude that the weight function in non-linear and inverted-S (such as that in Fig 3) is a good fit from empirical data. The CPT paper **?** recommends $w(p) = \frac{p^\delta}{(p^\delta+(1-p)^\delta)^{1/\delta}}$, while **?** recommends $w(p) = \exp(-(-\ln p)^\delta)$, with $0 < \delta < 1$ in both cases and both give the inverted-s shape for weight function.

Optimization objective: Suppose the r.v. X is a function of a d-dimensional parameter θ. The goal then is to solve the following problem:

$$\text{Find } \theta^* = \arg\min_{\theta\in\Theta} \mathbb{C}(X^\theta), \tag{2}$$

where Θ is a compact and convex subset of $\mathbb{R}^d$. The above optimization problem has several applications in RL. For instance, X could be the total reward/discounted reward/average reward r.v. for a fixed policy (given as θ) in the context of a stochastic shortest path/discounted/average reward MDP, respectively. We illustrate one of these applications next.

2.2 Application: Stochastic Shortest Path

We consider a stochastic shortest path (SSP) problem with states $\{0,\ldots,\mathcal{L}\}$ and actions $\{1,\ldots,\mathcal{M}\}$. An *episode* is a simulated sample path that starts in state x^0 and ends in the cost-free absorbing state 0. Let $\theta = \left(\theta^1,\ldots,\theta^{\mathcal{L}\mathcal{M}}\right)^\intercal$ be a randomized policy, where θ^i denotes the probability of choosing action $(i\%\mathcal{M})$ in state $\lceil i/\mathcal{M}\rceil$, with $\sum_{j=(i-1)\mathcal{M}+1}^{i\mathcal{M}} \theta^j = 1$, for $i = 1,\ldots,\mathcal{L}$. Let $D^\theta(x^0)$ be a random variable (r.v) that denotes the total cost from an episode simulated using policy θ starting from state x^0, i.e.,

$$D^\theta(x^0) = \sum_{m=0}^{\tau} g(x_m, a_m),$$

where the actions a_m are chosen using θ and τ is the first passage time to state 0.

The traditional RL objective for an SSP is to minimize the expected value $\mathbb{E}(D^\theta(x^0))$ and this can be written as

$$\min_{\theta\in\theta} \int_0^{+\infty} P(D^\theta(x^0)) > z)dz,$$

where Θ is the set of admissible policies that are *proper*[2].
In this paper, we adopt the CPT approach and aim to solve the following problem:

$$\min_{\theta\in\theta} \mathbb{C}(D^\theta(x^0)),$$

where the CPT-value function $\mathbb{C}(D^\theta(x^0))$ is defined as

$$\mathbb{C}(D^\theta(x^0)) = \int_0^{+\infty} w^+(P(u^+(D^\theta(x^0))) > z)dz - \int_0^{+\infty} w^-(P(u^-(D^\theta(x^0))) > z)dz. \tag{3}$$

[2]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. **?**.

Generalization: It is easy to see that the CPT-value is a generalization of the traditional value function, as a choice of identity map for the weight and utility functions in (3) makes it the expectation of the total cost D^θ). It is also possible to get (3) to coincide with coherent risk measures (e.g. CVaR) by the appropriate choice of weight functions.

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 the numerical experiments section).

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

3 CPT-value estimation

For the sake of notational simplicity, we let X denote the r.v. X^θ, i.e., where the parameter θ is assumed to be fixed for the purpose of CPT-value estimation in this section.

On integrability Observe that the first integral in (3), i.e.,

$$\int_0^{+\infty} w^+(P(u^+(X) > z))dz \tag{4}$$

may diverge even if the first moment of random variable $u^+(X)$ is finite. For example, suppose U has the tail distribution function

$$P(U > z) = \frac{1}{z^2}, z \in [1, +\infty),$$

and $w^+(z)$ takes the form $w(z) = z^{\frac{1}{3}}$. Then, the integral (4) with respect to the distorted tail, i.e.,

$$\int_1^{+\infty} \frac{1}{z^{\frac{2}{3}}} dz$$

does not even exist. A similar argument applies to the second integral in (3) as well.

To overcome the above integrability issues, we make different assumptions on the weight and/or utility functions. In particular, we assume that the weight functions w^+, w^- are either **(i)** Lipschitz continuous, or **(ii)** Hölder continuous, or **(iii)** locally Lipschitz. We devise a scheme for estimating (3) given only samples from X and show that, under each of the aforementioned assumptions, our estimator (presented next) converges almost surely. We also provide sample complexity bounds assuming that the utility functions are bounded.

3.1 Estimation scheme for Hölder continuous weights

Recall the Hölder continuity property first in definition 1:

Definition 1. ***(Hölder continuity)*** *If $0 < \alpha \leq 1$, a function $f \in C([a,b])$ is said to satisfy a Hölder condition of order α (or to be Hölder continuous of order α) if $\exists \;\; K$ s.t.*

$$\sup_{x \neq y} \frac{|f(x) - f(y)|}{|x-y|^{\alpha}} \leq K.$$

In order to ensure integrability of the CPT-value (3), we make the following assumption:

Assumption (A1). The support of the r.v.s $u^+(X)$ and $u^-(X)$ are $[0, +\infty)$, and the weight functions w^+, w^- are Hölder continuous with common order α. Further, $\exists \gamma \leq \alpha$ s.t,

$$\int_0^{+\infty} P^{\gamma}(u^+(X) > z)dz < +\infty \text{ and } \int_0^{+\infty} P^{\gamma}(u^-(X) > z)dz < +\infty.$$

The above assumption ensures that the CPT-value as defined by (3) is finite - see Proposition 4 in Section 7.1.1 for a formal proof.

Approximating CPT-value using quantiles: Let ξ_{α}^+ denote the αth quantile of the r.v. $u^+(X)$ Then, it can be seen that (see Proposition 5 in Section 7.1.1)

$$\lim_{n \to \infty} \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) = \int_0^{+\infty} w^+(P(u^+(X) > z))dz. \tag{5}$$

The identical property holds for the pairs $u^-(X)$, ξ_{α}^- and w^- , with ξ_{α}^- denote the αth quantile of the r.v. $u^-(X)$.

However, we do not know the distribution of $u^+(X)$ or $u^-(X)$ and hence, we develop a procedure that uses order statistics for estimating quantiles, which in turn assists in estimating the CPT-value along the lines of (5). The estimation scheme is presented in Algorithm 1.

Ch: Need some text here to connect the quantile business to EDFs, i.e., to say that step 5 in Algo 1 is the same as estimating EDFs and then doing empirical integration using EDFs to estiamete CPT-value

Algorithm 1 CPT-value estimation for Hölder continuous weights

1: Simulate n i.i.d. replications follows the distribution of r.v. X, sort them in ascending order and denote them as $X_{[1]}, X_{[2]}, \ldots X_{[n]}$.
2: Calculate $u^+(X_{[1]}), \ldots u^+(X_{[n]})$.
3: Order the simulated samples and label them as follows: $u^+(X_{[1]}), \ldots, u^+(X_{[n]})$.
4: Use $u^+(X_{[i]}), i \in \mathbb{N} \cap (0,n)$ as an approximation for the $\frac{i}{n}th$ quantile of $u^+(X)$, i.e, $\xi_{\frac{i}{n}}, i \in \mathbb{N} \cap (0,n)$.
5: Denote the statistic $\overline{\mathbb{C}}_n^+ := \sum_{i=1}^{n-1} u^+(X_{[i]})(w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))$
6: Repeat the procedure on the sequence $X_{[1]}, X_{[2]}, \ldots X_{[n]}$, with respect to the function u^-, and denote the statistic $\overline{\mathbb{C}}_n^- := \sum_{i=1}^{n-1} u^-(X_{[i]})(w^-(\frac{n-i}{n}) - w^-(\frac{n-i-1}{n}))$
7: Return the statistic $\overline{\mathbb{C}}_n = \overline{\mathbb{C}}_n^+ - \overline{\mathbb{C}}_n^-$.

Main results

We make the following assumptions on the utility functions:

Assumption (A2). The utility functions $u^+(X)$ and $u^-(X)$ are continuous and strictly increasing.

Assumption (A2'). In addition to (A2), the utility functions $u^+(X)$ and $u^-(X)$ are bounded above by $M < \infty$.

Ch: Check if increasing is necessary

For the convergence rate results below, we require (A2'), while (A2) is sufficient to prove asymptotic convergence.

Proposition 1. ***(Asymptotic convergence.)*** *«««< Updated upstream Assume (A1) and also that $F^+(\cdot)$ and $F^-(\cdot)$, the distribution functions of $u^+(X)$, and $u^-(X)$, are Lipschitz continuous with constants L^+ and L^-, respectively, on the interval $(0, +\infty)$, and $(-\infty, 0)$. Then, we have that*

$$\lim_{n\to+\infty} \overline{\mathbb{C}}_n = \mathbb{C}(X) \text{ a.s.}, \tag{6}$$

======= Assume (A1'). (A2) and also that $F^+(\cdot), F^-(\cdot)$ - the distribution functions of $u^+(X)$ and $u^-(X)$ - are Lipschitz continuous. Then, we have

$$\widehat{V}_n(X) \to V(X) \text{ a.s. as } n \to \infty, >>>>>>> Stashedchanges \tag{7}$$

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

Proof. See Section 7.1.1. □

While the above result establishes that $\overline{\mathbb{C}}_n$ is an unbiased estimate in the asymptotic sense, it is important to know the rate at which the estimate $\overline{\mathbb{C}}_n$ converges to the CPT-value $\mathbb{C}(X)$. The following sample complexity result shows that $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ number of samples are required to be ϵ-close to the CPT-value with high probability.

Proposition 2. ***(Sample complexity.)*** *Assume (A1) and (A2'). Then, $\forall \epsilon, \delta$, we have*

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

Proof. Notice the the following equivalence:

$$\sum_{i=1}^{n-1} u^+(X_{[i]})(w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) = \int_0^M w^+(1 - \widehat{F_n^+}(x))dx,$$

and also,

$$\sum_{i=1}^{n-1} u^-(X_{[i]})(w^-(\frac{n-i}{n}) - w^-(\frac{n-i-1}{n})) = \int_0^M w^-(1 - \widehat{F_n^-}(x))dx,$$

where $\widehat{F_n^+}(x)$ and $\widehat{F_n^-}(x)$ are the empirical distributions of $u^+(X)$ and $u^-(X)$,respectively defined as follows:

$$\hat{F}_n^+(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(u^+(X_i)\leq x)}, \hat{F}_n^-(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(u^-(X_i)\leq x)}. \tag{8}$$

The main claim follows fromt the equivalence mentioned above together with the well-known DKW inequality. The detailed proof is available in Section 7.1.1. □

Special case: Lipschitz continuous weights

Assumption (A1'). The weight functions w^+, w^- are Lipschitz with common constant L, and $u^+(X)$ and $u^-(X)$ both have bounded first moments.

Corollary 1 (Lipschitz case)**.** *Assume (A1') and (A2). Then, we have that*

$$\lim_{n\to+\infty} \overline{\mathbb{C}}_n = \mathbb{C}(X) \text{ a.s.}$$

In addition, if we assume (A2'), we have

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

Proof. Setting $\alpha = \gamma = 1$ in the proof of Proposition 4, it is easy to see that the CPT-value (3) is finite. Thus, the claims regarding asymptotic convergence and sample complexity are special cases of Proposition 1–2, with $\alpha = 1$. □

3.2 Estimation scheme for locally Lipschitz weights and discrete X

P: Would be better if the background been put to the introduction part

Background. Here we assume that the r.v. X is discrete valued and uses a definition for CPT-value that is equivalent to (1). 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 $\mathbb{C}(X) = \sum_{i=1}^{K} u(x_i)w(p_i)$. However, PT is lacking in some theoretical aspects as it violates first-order *stochastic dominance*.[3]

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

$$\mathbb{C}(X) = (u^-(x_1)) \cdot w^-(p_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\sum_{j=1}^{i} p_j) - w^-(\sum_{j=1}^{i-1} p_j)\Big) \tag{9}$$

$$+ \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\sum_{j=i}^{K} p_j) - w^+(\sum_{j=i+1}^{K} p_j)\Big) + u^+(x_K) \cdot w^+(p_K), \tag{10}$$

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.[4]

[3]Consider the following example from **?**: 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.

[4]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.

Estimation scheme. Let $\hat{p_k} = \frac{1}{n}\sum_{i=1}^{n} I_{\{U=x_k\}}$. Then, we estimate $\mathbb{C}(X)$ as follows:

$$\overline{\mathbb{C}}_n = u^-(x_1)\cdot w^-(\hat{p}_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\sum_{j=1}^{i}\hat{p}_j) - w^-(\sum_{j=1}^{i-1}\hat{p}_j)\Big) \tag{11}$$

$$+ \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\sum_{j=i}^{K}\hat{p}_j) - w^+(\sum_{j=i+1}^{K}\hat{p}_j)\Big) + u^+(x_K)\cdot w^+(\hat{p}_K). \tag{12}$$

Because $\hat{p_k}$ converge a.e to $p_k = P(X_i = x_k)$, with X_i be the ith sample of X, the above estimator is strong consistent property by the continuous mapping theorem.

Main result

The following proposition presents a sample complexity result for the discrete-valued X under the following assumption:

Assumption (A3). The weight functions $w^+(X)$ and $w^-(X)$ are locally Lipschitz continuous, i.e., for any x, there exist $L < \infty$ and $\delta > 0$, such that

$$|w^+(x) - w^+(y)| \leq L_x|x - y|, \text{ for all } y \in (x-\delta, x+\delta).$$

We denote $L = \max\{L_k, k = 2...K\}$, where L_k is the Lipschitz constant at $F_k = \sum_{i=k}^{K} p_i$.
Let

$$F_k = \begin{cases} \sum_{i=1}^{k} p_k & \text{if } k \leq l \\ \sum_{i=k}^{K} p_k & \text{if } k > l. \end{cases} \tag{13}$$

The main result for discrete-valued X is given below.

Proposition 3 (Sample Complexity: discrete case)**.** *Denote $L = \max\{L_k, k = 2...K\}$, where L_k is the local Lipschitz constant of function $w^-(x)$ at points F_k, where $k = 1, ...l$, and of function $w^+(x)$ at points $k = l+1, ...K$. And let $A = \max\{x_k, k = 1...K\}$, $\delta = \min\{\delta_k\}$, where δ_k is the half length of the interval centered at point F_k where the locally Lipschitz property with constant L_k holds. For any ϵ, ρ,let $M = \min(\delta^2, \epsilon^2/(KLA)^2)$, and we have*

$$P(\big|\overline{\mathbb{C}}_n - \mathbb{C}(X)\big| \leq \epsilon) > 1 - \rho \,, \forall n > \frac{\ln(\frac{4K}{a})}{M} \tag{14}$$

Proof. See Section 7.1.2. □

4 Gradient-based algorithm for CPT optimization (CPT-SPSA)

4.1 Gradient estimation

Given that we operate in a learning setting and only have biased estimates of the CPT-value from Algorithm 1, we require a simulation optimization scheme that estimates $\nabla\mathbb{C}(X^\theta)$. Simultaneous perturbation methods are a general class of stochastic gradient schemes that optimize a function given only noisy sample values - see **?** for textbook introduction. SPSA is a well-known scheme that estimates the gradient using two sample

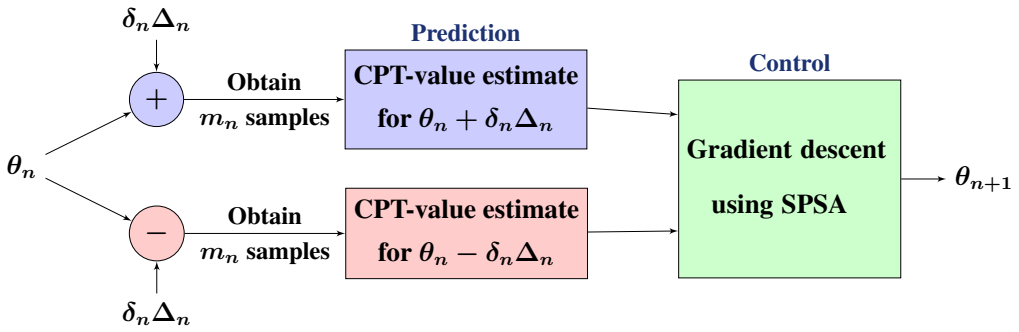


Figure 4: Overall flow of CPT-SPSA-G.

(a) Simulation optimization

(b) CPT-value optimization

Figure 5: Illustration of difference between classic simulation optimization and CPT-value optimiziation settings

values. In our context, at any iteration n of CPT-SPSA-G, with parameter θ_n, the gradient $\nabla\mathbb{C}(X^{\theta_n})$ is estimated as follows: For any $i = 1, \ldots, d$,

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

where δ_n is a positive scalar that satisfies (A3) below, $\Delta_n = \left(\Delta_n^1, \ldots, \Delta_n^{\mathcal{LM}}\right)^\mathsf{T}$, where $\{\Delta_n^i, i = 1, \ldots, \mathcal{LM}\}$, $n = 1, 2, \ldots$ are i.i.d. Rademacher, 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. $\overline{X}^{\theta_n-\delta_n\Delta_n}$). The (asymptotic) unbiasedness of the gradient estimate is proven in Lemma 6.

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. **?**, **?**. However, the idea applies to non-convex functions as well - cf. **?**, **?**.

4.2 Update rule

We incrementally update the parameter θ in the descent direction as follows: For every state $i = 1, \ldots, \mathcal{LM}$,

$$\theta_{n+1}^i = \Gamma_i\left(\theta_n^i - \gamma_n\widehat{\nabla}_i\mathbb{C}(X^{\theta_n})\right), \tag{15}$$

where γ_n is a step-size chosen to satisfy (A3) below and $\Gamma = (\Gamma_1, \ldots, \Gamma_d)$ is an operator that ensures that the update (15) stays bounded within a compact and convex set Θ. Fig. 4 illustrates the overall flow of the gradient algorithm based on SPSA, while Algorithm 2 presents the pseudocode.

On the number of samples m_n per iteration: As illustrated in Figure 5, the CPT-value estimation scheme is biased, i.e., providing samples with parameter θ_n at instant n, we obtain its CPT-value estimate as $\mathbb{C}(X^{\theta_n}) + \epsilon_n^\theta$ with ϵ_n^θ denoting the bias. The bias can be controlled by increasing the number of samples m_n in each iteration of CPT-SPSA (see Algorithm 2). This is unlike classic simulation optimization settings

Algorithm 2 Structure of CPT-SPSA-G algorithm.

Input: initial parameter θ_0, perturbation constants $\delta_n > 0$, sample size $\{m_n\}$, step-sizes $\{\gamma_n\}$, operator Γ.

for $n = 0, 1, 2, \ldots$ **do**

Generate $\{\Delta_n^i, i = 1, \ldots, \mathcal{LM}\}$ using Rademacher distribution, independent of $\{\Delta_m, m = 0, 1, \ldots, n-1\}$

CPT-value Estimation (Trajectory 1)

Simulate m_n samples using parameter $(\theta_n + \delta_n \Delta_n)$

Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n}$

CPT-value Estimation (Trajectory 2)

Simulate m_n samples using parameter $\theta_n - \delta_n \Delta_n$

Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}$

Gradient descent

$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i - \gamma_n \widehat{\nabla}_i \mathbb{C}(X^{\theta_n}) \right)$

end for

Return θ_n

where one only sees function evaluations with zero mean noise and there is no question of deciding on m_n to control the bias as we have in our setting.

To motivate the choice for m_n, we first rewrite the update rule (15) as follows:

$$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i - \gamma_n \left(\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i} \right) + \underbrace{\frac{(\epsilon_n^{\theta_n + \delta_n \Delta_n} - \epsilon_n^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i}}_{\kappa_n} \right).$$

Let $\zeta_n = \sum_{l=0}^{n} \gamma_l \kappa_l$. Then, a critical requirement that allows us to ignore the bias term ζ_n is the following condition (see Lemma 1 in Chapter 2 of **?**):

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

While Theorems 1–2 show that the bias ϵ^θ is bounded above, to establish convergence of the policy gradient recursion (15), we increase the number of samples m_n so that the bias vanishes asymptotically. Assumption (A3) provides a condition on the rate at which m_n has to increase.

Assumption (A3). The step-sizes γ_n and the perturbation constants δ_n are positive $\forall n$ and satisfy

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

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 $\gamma > \nu\alpha/2$.

4.3 Convergence result

Theorem 2. *Assume (A1)-(A3). Consider the ordinary differential equation (ODE):*

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

where

$$\check{\Gamma}_i(f(\theta)) := \lim_{\alpha \downarrow 0} \frac{\Gamma_i(\theta + \alpha f(\theta)) - \theta}{\alpha}, \textit{ for any continuous } f(\cdot).$$

Let $\mathcal{K} = \{\theta \mid \check{\Gamma}_i \left(\nabla_i V^\theta(x^0)\right) = 0, \forall i = 1, \ldots, d\}$. *Then,*

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

Proof. See Section 7.2. □

5 Newton algorithm for CPT-value optimization (CPT-SPSA-N)

5.1 Need for second-order methods

While stochastic gradient descent methods are useful in minimizing 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 $\gamma_n = \gamma_0/n$, if a_0 is not chosen to be greater than $1/3\lambda_{min}(\nabla^2 \mathbb{C}(X^{\theta^*}))$, then the optimum rate of convergence is not achieved. Here $\lambda_{\min}$ denotes the minimum eigenvalue, while $\theta^* \in \mathcal{K}$ (see Theorem 2). A standard approach to overcome this step-size dependency is to use iterate averaging, suggested independently by Polyak **?** and Ruppert **?**. The idea is to use larger step-sizes $\gamma_n = 1/n^\varsigma$, where $\varsigma \in (1/2, 1)$, and then combine it with averaging of the iterates. 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 **?**). Thus, it is optimal to average iterates only after a sufficient number of iterations have passed and all the iterates are very close to the optimum. However, the latter situation serves as a stopping condition in practice.

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.

5.2 Gradient and Hessian estimation

We estimate the Hessian of the CPT-value function using the scheme suggested by **?**. As in the case of the first-order method, we use Rademacher 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. Rademacher and independent of $\theta_0, \ldots, \theta_n$. 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

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

$$\widehat{H}_n^{i,j} = \frac{\overline{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \widehat{\Delta}_n} + \overline{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \widehat{\Delta}_n} - 2\widehat{V}_n^{\theta_n}(x^0)}{\delta_n^2 \Delta_n^i \widehat{\Delta}_n^j}.$$

Further, set $\widehat{H}_n^{i,j} = \widehat{H}_n^{j,i}$, for $i, j = 1, \ldots, d$. Notice that the above estimates require three samples, while the second-order SPSA algorithm proposed first in **?** required four. Both the gradient estimate $\widehat{\nabla} V_n^{\theta_n}(x^0)$ and the Hessian estimate $\widehat{H_n}$ can be shown to be an $O(\delta_n^2)$ term away from the true gradient $\nabla V_n^\theta(x^0)$ and Hessian $\nabla^2 V_n^\theta(x^0)$, respectively (see Lemmas 7–8).

Algorithm 3 Structure of CPT-SPSA-N algorithm.

Input: initial parameter θ_0, perturbation constants $\delta_n > 0$, trajectory lengths $\{m_n\}$, step-sizes $\{\gamma_n, \xi_n\}$.
for $n = 0, 1, 2, \ldots$ **do**
Generate $\{\Delta_n^i, \widehat{\Delta}_n^i, i = 1, \ldots, d\}$ using Rademacher distribution, independent of $\{\Delta_m, \widehat{\Delta}_m, m = 0, 1, \ldots, n-1\}$
CPT-value Estimation (Trajectory 1)
Simulate m_n samples using parameter $(\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n))$
Obtain CPT-value estimate $\widehat{V}_n^{(\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n))}(x^0)$ using Algorithm 1
CPT-value Estimation (Trajectory 2)
Simulate m_n samples using parameter $(\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n))$
Obtain CPT-value estimate $\widehat{V}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}(x^0))$ Algorithm 1
CPT-value Estimation (Trajectory 3)
Simulate m_n samples using parameter θ_n
Obtain CPT-value estimate $\widehat{V}_n^{\theta_n}(x^0))$ 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 (16)–(17)
end for
Return θ_n

5.3 Update rule

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

$$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i - \gamma_n \sum_{j=1}^{\mathcal{LM}} M_n^{i,j} \widehat{\nabla}_j \mathbb{C}(X_n^\theta) \right), \tag{16}$$

$$\overline{H}_n = (1 - \xi_n)\overline{H}_{n-1} + \xi_n \widehat{H}_n, \tag{17}$$

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 (16) - see (**?**, Chapter 6). Further, Γ is a projection operator as in CPT-SPSA-G and $M_n = \Upsilon(\overline{H}_n)^{-1}$. 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 (A4). For any $\{A_n\}$ and $\{B_n\}$, $\lim_{n\to\infty} \|A_n - B_n\| = 0 \Rightarrow \lim_{n\to\infty} \| \Upsilon(A_n) - \Upsilon(B_n) \| = 0$. Further, for any $\{C_n\}$ with $\sup_n \| C_n \| < \infty$, $\sup_n \left(\| \Upsilon(C_n) \| + \| \{\Upsilon(C_n)\}^{-1} \| \right) < \infty$ as well.

A simple way to ensure the above is to have $\Upsilon(\overline{H}_n)$ as a diagonal matrix and then add a positive scalar δ_n to the diagonal elements so as to ensure invertibility - see **?**, **?** for a similar operator.

The overall flow on CPT-SPSA-N is similar to Fig. 4, except that three system trajectories with a different perturbation sequence are used. Algorithm 3 presents the pseudocode.

5.4 Convergence result

Theorem 3. *Assume (A1)-(A4). Consider the ODE:*

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

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

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

Proof. See Section 7.3. □

6 Gradient-free CPT-value optimization algorithm (GF-CPT)

We perform a non-trivial adaptation of the algorithm from **?** to our setting of optimizing CPT-value in MDPs. We require that there exists a unique global optimum θ^* for the problem $\min_{\theta \in \Theta} \mathbb{C}(X^\theta)$.

6.1 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 **?** is

$$g_k(\theta) = \frac{\mathcal{H}(\mathbb{C}(X^\theta)) g_{k-1}(\theta)}{\int_\theta \mathcal{H}(\mathbb{C}(X^{\theta'})) g_{k-1}(\theta') \nu(d\theta')}, \ \forall \, \theta \in \theta, \tag{18}$$

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 policies having lower CPT-values and 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 policy θ. In this case, we consider a family of parameterized distributions, say $\{f(\cdot, \eta), \, \eta \in \mathbb{C}\}$ and incrementally update the distribution parameter η such that it minimizes the following KL divergence:

$$\mathcal{D}(g_k, f(\cdot, \eta)) := E_{g_k} \left[\ln \frac{g_k(\mathcal{R}(\theta))}{f(\mathcal{R}(\theta), \eta)} \right] = \int_\theta \ln \frac{g_k(\theta)}{f(\theta, \eta)} g_k(\theta) \nu(d\theta),$$

where $\mathcal{R}(\theta)$ is a random vector taking values in the policy space θ. An algorithm to optimize CPT-value in this *noise-less* setting would perform the following update for the parameter η_n:

$$\eta_{n+1} \in \arg\min_{\eta \in \mathbb{C}} E_{\eta_n} \left[\frac{[\mathcal{H}(\mathbb{C}(X^{\mathcal{R}(\theta)}))]^n}{f(\mathcal{R}(\theta), \eta_n)} \ln f(\mathcal{R}(\theta), \eta) \right], \tag{19}$$

where $E_{\eta_n}[\mathbb{C}(X^{\mathcal{R}(\theta)})] = \int_\theta V^\theta(x^0) f(\theta, \eta_n) \nu(d\theta)$.

[1]Here $\widehat{V}_n^{\theta_n^{(i)}}(x^0)$ denotes the ith order statistic.

[2]Here $\widetilde{I}(z, \chi) := \begin{cases} 0 & \text{if } z \leq \chi - \varepsilon, \\ (z - \chi + \varepsilon)/\varepsilon & \text{if } \chi - \varepsilon < z < \chi, \\ 1 & \text{if } z \geq \chi. \end{cases}$

Algorithm 4 Structure of GF-CPT-MPS algorithm.

Input: family of distributions $\{f(\cdot,\eta)\}$, initial parameter vector η_0 s.t. $f(\theta,\eta_0)>0\ \forall\,\theta\in\theta$, trajectory lengths $\{m_n\}$, $\rho_0\in(0,1]$, $N_0>1$, $\varepsilon>0$, $\alpha>1$, $\lambda\in(0,1)$, strictly decreasing function $\mathcal{H}$

for $n=0,1,2,\ldots$ **do**

Candidate Policies

Generate N_n policies using the mixed distribution $\widetilde{f}(\cdot,\eta_n)=(1-\lambda)f(\cdot,\widetilde{\eta}_n)+\lambda f(\cdot,\eta_0)$.
Denote these candidate policies by $\Lambda_n=\{\theta_n^1,\ldots,\theta^{N_n}\}$.

CPT-value Estimation

for $i=1,2,\ldots,N_n$ **do**

Simulate m_n samples using policy θ_n^i

Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n^i}$ using Algorithm 1

end for

Elite Sampling

Order the CPT-value estimates[1] $\{\overline{\mathbb{C}}_n^{\theta_n^{(1)}},\ldots,\overline{\mathbb{C}}_n^{\theta_n^{(N_n)}}\}$.
Compute the $(1-\rho_n)$-quantile from the above samples as follows:

$$\widetilde{\chi}_n(\rho_n,N_n)=\overline{\mathbb{C}}_n^{\theta_n^{\lceil(1-\rho_k)N_k\rceil}}. \tag{20}$$

Thresholding

if $n=0$ or $\widetilde{\chi}_n(\rho_n,N_n)\geq\bar{\chi}_{n-1}+\varepsilon$ **then**

Set $\bar{\chi}_k=\widetilde{\chi}_k(\rho_n,N_n),\ \rho_{k+1}=\rho_n,\ N_{k+1}=N_k$ and
Set $\theta_n^*=\theta_{1-\rho_n}$, where $\theta_{1-\rho_n}$ is the policy that corresponds to the $(1-\rho_n)$-quantile in (20).

else

find the largest $\bar{\rho}\in(0,\rho_n)$ such that $\widetilde{\chi}_n(\bar{\rho},N_n)\geq\bar{\chi}_{n-1}+\varepsilon$;

if $\bar{\rho}$ exists **then**

Set $\bar{\chi}_n=\widetilde{\chi}_n(\bar{\rho},N_n),\ \rho_{k+1}=\bar{\rho},\ N_{n+1}=N_n$ and $\theta_n^*=\theta_{1-\bar{\rho}}$

else

Set $\bar{\chi}_n=\widehat{V}_n^{\theta_{n-1}^*}(x^0)$,$\rho_{n+1}=\rho_n$,$N_{n+1}=\lceil\alpha N_n\rceil$, and $\theta_n^*=\theta_{n-1}^*$.

end if

end if

Sampling Distribution Update

Parameter update[2]:

$$\eta_{n+1}\in\arg\min_{\eta\in\mathbb{C}}\frac{1}{N_n}\sum_{i=1}^{N_n}\frac{[\mathcal{H}(\overline{\mathbb{C}}_n^{\theta_n^i})]^n}{\widetilde{f}(\theta,\eta_n)}\widetilde{I}\big(\overline{\mathbb{C}}_n^{\theta_n^i},\bar{\chi}_n\big)\ln f(\theta,\eta).$$

end for

Return θ_n

Finally, we get to our setting where we only obtain a biased estimate of the CPT-value $\mathbb{C}(X^\theta)$ for any policy θ. Recall that the bias is due to a finite sample run followed by the estimation scheme outlined in Algorithm 1. As in the case of SPSA-based algorithms, it is easy to see that the number of samples m_n (in iteration n) should asymptotically increase to infinity. Assuming this setup, the gradient-free model-based policy search algorithm would involve the following steps (see Algorithm 4 for the pseudocode):

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

Step 2 (CPT-value estimation): Obtain m_n samples for each of the parameters in $\theta_n^i, i = 1, \ldots, N_n$ and return CPT-value estimates $\overline{\mathbb{C}}_n^{\theta_n^i}$.

Step 3 (Parameter update):

$$\eta_{n+1} \in \arg\min_{\eta \in \mathbb{C}} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{[\mathcal{H}(\overline{\mathbb{C}}_n^{\theta_n^i})]^n}{f(\theta_n^i, \eta_n)} \ln f(\theta_n^i, \eta). \tag{21}$$

A few remarks are in order.

Remark 1. ***(Choice of sampling distribution)*** *A natural question is how to compute the KL-distance* (19) *in order to update the policy. A related question is how to choose the family of distributions $f(\cdot, \theta)$, so that the update* (19) *can be done efficiently. One choice is to employ the natural exponential family (NEF) since it ensures that the KL distance in* (19) *can be computed analytically.*

Remark 2. ***(Elite sampling)*** *In practice, it is efficient to use only an elite portion of the candidate policies that have been sampled in order to update the sampling distribution $f(\cdot, \eta)$. 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 policy search procedure towards better regions more efficiently in comparison to an alternative that uses all the candidate policies for updating η.*

6.2 Convergence result

Theorem 4. *Assume (A1)-(A2). 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,*

$$\lim_{n\to\infty} \mu_n = \theta^* \text{ and } \lim_{n\to\infty} \Sigma_n = 0_{d\times d} \ \ a.s. \tag{22}$$

Proof. See Section 7.4. □

7 Convergence Proofs

7.1 Proofs for CPT-value estimator

7.1.1 Hölder continuous weights

Proposition 4. *Under (A1'), the CPT-value $\mathbb{C}(X)$ as defined by* (3) *is finite.*

Proof. Hölder continuity of w^+ together with the fact that $w^+(0) = 0$ imply that

$$\int_0^{+\infty} w^+(P(U^+ > t))dz \leq C \int_0^{+\infty} P^\alpha(U^+ > z)dz \leq C \int_0^{+\infty} P^\gamma(U^+ > z)dz < +\infty.$$

The second inequality is valid since $P(U^+ > z) \leq 1$. The claim follows for the first integral in (3) and the finiteness of the second integral in (3) can be argued in an analogous fashion. □

Proposition 5. *Assume (A1'). Let $\xi^+_{\frac{i}{n}}$ and $\xi^-_{\frac{i}{n}}$ denote the $\frac{i}{n}$th quantile of U^+ and U^-, respectively. Then, we have*

$$\begin{aligned}\lim_{n\to\infty}\sum_{0}^{n-1}\xi^+_{\frac{i}{n}}(w^+(\frac{n-i}{n})-w^+(\frac{n-i-1}{n}))&=\int_0^{+\infty}w^+(P(U^+>z))dz<+\infty,\\ \lim_{n\to\infty}\sum_{0}^{n-1}\xi^-_{\frac{i}{n}}(w^-(\frac{n-i}{n})-w^-(\frac{n-i-1}{n}))&=\int_0^{+\infty}w^-(P(U^->z))dz<+\infty\end{aligned} \tag{23}$$

Ch: Substitute U, tdt, w, ξ with U^+, zdz, w^+ and ξ^+, and say similar argument holds for $-$ parts,done

Proof. We shall focus on proving the first part of equation (28). Consider the following linear combination of simple functions:

$$\sum_{i=0}^{n-1}w^+(\frac{i}{n})\cdot I_{[\xi^+_{\frac{n-i-1}{n}},\xi_{\frac{n-i}{n}}]}(t), \tag{24}$$

which will converge almost everywhere to the function $w(P(U > t))$ in the interval $[0,+\infty)$, and also notice that

$$\sum_{i=0}^{n-1}w^+(\frac{i}{n})\cdot I_{[\xi^+_{\frac{n-i-1}{n}},\xi^+_{\frac{n-i}{n}}]}(t)<w(P(U>t)),\ \forall t\in[0,+\infty). \tag{25}$$

The integral of (4) equals

$$\int_0^{+\infty}\sum_{i=0}^{n-1}w^+_{\frac{i}{n}}\cdot I_{[\xi^+_{\frac{n-i-1}{n}},\xi^+_{\frac{n-i}{n}}]}(t) \tag{26}$$

$$=\sum_{i=0}^{n-1}w^+_{\frac{i}{n}}(t)\cdot(\xi^+_{\frac{n-i}{n}}-\xi^+_{\frac{n-i-1}{n}}) \tag{27}$$

$$=\sum_{i=0}^{n-1}\xi^+_{\frac{i}{n}}\cdot(w^+_{\frac{n-i}{n}}-w^+_{\frac{n-i-1}{n}}). \tag{28}$$

The Hölder continuity property assures the fact that $\lim_{n\to\infty}|w^+(\frac{n-i}{n})-w^+(\frac{n-i-1}{n})|=0$, and the limit in (28) holds through a typical application of the dominated convergence theorem. The second part of equation (28) can be justified in a similar fashion. □

Proof of Proposition 1

Ch: Substitute U, tdt, w, ξ^+ with U^+, zdz, w^+ and ξ^+, resply and say similar argument holds for $-$ parts, done

Proof. We prove the w^+ part, and the w^- part is proved in a similar fashion.

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

$$P(\left|\sum_{i=1}^{n-1} U_{[i]}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n}))\right| > \epsilon), \quad (29)$$

for any given $\epsilon > 0$. Observe that

$$\begin{aligned}
&P(\left|\sum_{i=1}^{n-1} U_{[i]}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n}))\right| > \epsilon) \\
&\leq P(\bigcup_{i=1}^{n-1} \{\left|U_{[i]}^{+} \cdot (w_{(}^{+}\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n})) - \xi_{\frac{i}{n}}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}\}) \\
&\leq \sum_{i=1}^{n-1} P(\left|U_{[i]}^{+} \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n})) - \xi_{\frac{i}{n}}^{+} \cdot (w_{(}^{+}\frac{n-i}{n}) - w_{(}^{+}\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}) \\
&= \sum_{i=1}^{n-1} P(\left|(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+}) \cdot (w^{+}(\frac{n-i}{n}) - w^{+}(\frac{n-i-1}{n}))\right| > \frac{\epsilon}{n}) \\
&\leq \sum_{i=1}^{n-1} P(\left|(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+}) \cdot (\frac{1}{n})^{\alpha}\right| > \frac{\epsilon}{n}) \\
&= \sum_{i=1}^{n-1} P(\left|(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+})\right| > \frac{\epsilon}{n^{1-\alpha}}).
\end{aligned}$$

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

$$\begin{aligned}
&P(\left|U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+}\right| > \frac{\epsilon}{n^{(1-\alpha)}}) \\
&= P(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+} > \frac{\epsilon}{n^{(1-\alpha)}}) + P(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+} < -\frac{\epsilon}{n^{(1-\alpha)}}).
\end{aligned}$$

We focus on the term $P(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+} > \frac{\epsilon}{n^{(1-\alpha)}})$. Let $W_t = I_{(U_t^{+} > \xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}})}, t = 1, \ldots, n$. Using the fact that probability distribution function is non-decreasing, we obtain

$$\begin{aligned}
&P(U_{[i]}^{+} - \xi_{\frac{i}{n}}^{+} > \frac{\epsilon}{n^{(1-\alpha)}}) \\
&= P(\sum_{t=1}^{n} W_t > n \cdot (1 - \frac{i}{n^{(1-\alpha)}})) \\
&= P(\sum_{t=1}^{n} W_t - n \cdot [1 - F(\xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F(\xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]).
\end{aligned}$$

Notice that $EW_t = 1 - F(\xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}})$, and by recalling Hoeffding's inequality, one can derive that

$$P(\sum_{i=1}^{n} W_t - n \cdot [1 - F(\xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F(\xi_{\frac{i}{n}}^{+} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]) < e^{-2n \cdot \delta_t'}, \quad (30)$$

with $\delta_i^{'} = F(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}$, and if $F(x)$ is Lipschitz, $\delta_i^{'} \leq L \cdot (\frac{\epsilon}{n^{(1-\alpha)}})$. Therefore we will have

$$P(U^+_{[i]} - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}) < e^{-2n \cdot L \frac{\epsilon}{n^{(1-\alpha)}}} = e^{-2n^{\alpha} \cdot L\epsilon} \tag{31}$$

In a similar fashion, one can show that

$$P(U^+_{[i]} - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{n^{(1-\alpha)}}) \leq e^{-2n^{\alpha} \cdot L\epsilon}$$

Because of the assumption that $F(x)$ is Lipschitz continuous, one can state that

$$P(\left| U^+_{[i]} - \xi^+_{\frac{i}{n}} \right| < -\frac{\epsilon}{n^{(1-\alpha)}}) \leq 2 \cdot e^{-2n^{\alpha} \cdot L\epsilon}, \ \forall i \in \mathbb{N} \cap (0,1)$$

As a result we can derive a bound for the probability in (6) such that

$$P(\left| \sum_{i=1}^{n-1} U^+_{[i]} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) \right| > \epsilon) \leq 2n \cdot e^{-2n^{\alpha} \cdot L}. \tag{32}$$

Notice that $\sum_{n=1}^{+\infty} 2n \cdot e^{-2n^{\alpha} \cdot L\epsilon} < \infty$ since the sequence $2n \cdot e^{-2n^{\alpha} \cdot L}$ will decrease more rapidly than the sequence $\frac{1}{n^k}, \forall k > 1$.

By applying the Borel Cantelli lemma,

$$P(\left| \sum_{i=1}^{n-1} U^+_{[i]} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) \right| > \epsilon, i.o.) = 0, \ \forall \epsilon > 0$$

which implies

$$\sum_{i=1}^{n-1} U^+_{[i]} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) \xrightarrow{n \to +\infty} 0 \text{ w.p } 1,$$

which constitutes the proof of theorem's statement, that

$$\lim_{n \to +\infty} \sum_{i=1}^{n-1} U^+_{[i]} (w^+(\frac{n-i+1}{n}) - w^+(\frac{n-i}{n})) \xrightarrow{n \to \infty} \int_0^{+\infty} w^+(P(U > t))dt, \text{w.p1} \tag{33}$$

□

Proof of Proposition 2

Proof. We prove the w^+ part, and the w^- part is proved in a similar fashion. Since U^+ is bounded above by M and w is Hölder with constant C and power α, we have

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

Now, plugging in the DKW inequality, we obtain

$$P\left(\left|\int_0^{+\infty} w^+(P(U^+) > t)dt - \int_0^{+\infty} w^+(1 - \hat{F}_n^+(t))dt\right| > \epsilon\right)$$
$$\leq P\left(LM \sup_{t\in\mathbb{R}} \left|(P(U^+ < t) - \hat{F}_n^+(t)\right|^\alpha > \epsilon\right) \leq e^{-n\frac{\epsilon^{(2/\alpha)}}{2L^2M^2}}. \tag{34}$$

□

7.1.2 Proofs for discrete valued X

For the sake of notational convenience, we assume $w^+ = w^- = w$. For proving Proposition 3, we require Hoeffding's inequality, which is given below.

Lemma 5. *Let $Y_1, ...Y_n$ be independent random variables satisfying $P(a \leq Y_i \leq b) = 1$, for each i, where $a < b$. Then for $t > 0$,*

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

Let

$$\hat{F}_k = \begin{cases} \sum_{i=1}^k \hat{p}_k & \text{if } k \leq l \\ \sum_{i=k}^K \hat{p}_k & \text{if } k > l. \end{cases} \tag{35}$$

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

Proposition 6. *Let F_k and $\hat{F}_k$ be as defined in* (13)*,* (35)*, Then, we have that, 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. Notice that when $k > l$, $\hat{F}_k = I_{(U_i \geq x_k)}$. Since the random variables U_i are independent of each other and for each i, are bounded above by 1, we can apply Hoeffding's inequality to obtain

$$P(\left|\hat{F}_k - F_k\right| > \epsilon)$$
$$= P(\left|\frac{1}{n}\sum_{i=1}^n I_{\{U_i \geq x_k\}} - \frac{1}{n}\sum_{i=1}^n E(I_{\{U_i \geq x_k\}})\right| > \epsilon)$$
$$= P(\left|\sum_{i=1}^n I_{\{U_i \geq x_k\}} - \sum_{i=1}^n E(I_{\{U_i \geq x_k\}})\right| > n\epsilon)$$
$$\leq 2e^{-2n\epsilon^2}.$$

□

The proof of Proposition 3 requires the following claim which gives the convergence rate under local Lipschitz weights.

Proposition 7. *Under conditions of Proposition 3, with F_k and $\hat{F}_k$ as defined in* (13) (35)*, we have*

$$P(\left|\sum_{i=1}^K x_k w(\hat{F}_k) - \sum_{i=1}^K x_k w(F_k)\right| > \epsilon) < K \cdot (e^{-\delta^2 \cdot 2n} + e^{-\epsilon^2 2n/(KLA)^2})$$

Proof. Observe that

$$
\begin{aligned}
&P(\left|\sum_{k=1}^{K} x_k w(\hat{F}_k) - \sum_{k=1}^{K} x_k w(F_k)\right| > \epsilon) \\
&= P(\bigcup_{k=1}^{K} \left|x_k w(\hat{F}_k) - x_k w(F_k)\right| > \frac{\epsilon}{K}) \\
&\leq \sum_{k=1}^{K} P(\left|x_k w(\hat{F}_k) - x_k w(F_k)\right| > \frac{\epsilon}{K})
\end{aligned}
$$

Notice that $\forall k = 1, ...K$ $[p_k - \delta, p_k + \delta)$, the function w is locally Lipschitz with common constant L. Therefore, for each k, we can decompose the probability as

$$
\begin{aligned}
&P(\left|x_k w(\hat{F}_k) - x_k w(F_k)\right| > \frac{\epsilon}{K}) \\
&= P([\left|F_k - \hat{F}_k\right| > \delta] \bigcap [\left|x_k w(\hat{F}_k) - x_k w(F_k)\right|] > \frac{\epsilon}{K}) + P([\left|F_k - \hat{F}_k\right| \leq \delta] \bigcap [\left|x_k w(\hat{F}_k) - x_k w(F_k)\right|] > \frac{\epsilon}{K}) \\
&\leq P(\left|F_k - \hat{F}_k\right| > \delta) + P([\left|F_k - \hat{F}_k\right| \leq \delta] \bigcap [\left|x_k w(\hat{F}_k) - x_k w(F_k)\right|] > \frac{\epsilon}{K}).
\end{aligned}
$$

According to the property of locally Lipschitz continuous, we have

$$
\begin{aligned}
&P([\left|F_k - \hat{F}_k\right| \leq \delta] \bigcap [\left|x_k w(\hat{F}_k) - x_k w(F_k)\right|] > \frac{\epsilon}{K}) \\
&\leq P(x_k L \left|F_k - \hat{F}_k\right| > \frac{\epsilon}{K}) \leq e^{-\epsilon \cdot 2n/(KLx_k)^2} \leq e^{-\epsilon \cdot 2n/(KLA)^2} \text{ for } \forall k.
\end{aligned}
$$

And similarly,

$$
\begin{aligned}
&P(\left|F_k - \hat{F}_k\right| > \delta) \\
&\leq e^{-\delta^2/2n} \text{ for } \forall k.
\end{aligned}
$$

And as a result,

$$
\begin{aligned}
&P(\left|\sum_{k=1}^{K} x_k w(\hat{F}_k) - \sum_{k=1}^{K} x_k w(F_k)\right| > \epsilon) \\
&\leq \sum_{k=1}^{K} P(\left|x_k w(\hat{F}_k) - x_k w(F_k)\right| > \frac{\epsilon}{K}) \\
&\leq \sum_{k=1}^{K} \left(e^{-\delta^2 \cdot 2n} + e^{-\epsilon^2 \cdot 2n/(KLA)^2}\right) \\
&= K \cdot (e^{-\delta^2 \cdot 2n} + e^{-\epsilon^2 \cdot 2n/(KLA)^2})
\end{aligned}
$$

□

Proof of Proposition 3

Proof. We need to prove that

$$P(\left|\sum_{i=1}^{K} u(x_k)\cdot(w(\hat{F}_k)-w(\hat{F}_{k+1}))-\sum_{i=1}^{K} u(x_k)\cdot(w(F_k)-w(F_{k+1}))\right|\leq\epsilon)>1-\rho\,,\forall n>\frac{\ln(\frac{4K}{a})}{M}, \tag{36}$$

where 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 Proposition 7, it is easy to infer that

$$P(\left|\sum_{i=1}^{K} x_k w(\hat{F}_{k+1})-\sum_{i=1}^{K} x_k w(F_{k+1})\right|>\epsilon)<K\cdot(e^{-\delta^2\cdot 2n}+e^{-\epsilon^2 2n/(KLA)^2})$$

Hence,

$$\begin{aligned}
&P(\left|\sum_{i=1}^{K} x_k\cdot(w(\hat{F}_k)-w(\hat{F}_{k+1}))-\sum_{i=1}^{K} x_k\cdot(w(F_k)-w(F_{k+1}))\right|>\epsilon)\\
&\leq P(\left|\sum_{i=1}^{K} x_k\cdot(w(\hat{F}_k))-\sum_{i=1}^{K} x_k\cdot(w(F_k))\right|>\epsilon/2)+P(\left|\sum_{i=1}^{K} x_k\cdot(w(\hat{F}_{k+1}))-\sum_{i=1}^{K} x_k\cdot(w(F_{k+1}))\right|>\epsilon/2)\\
&\leq 2K(e^{-\delta^2\cdot 2n}+e^{-\epsilon^2 2n/(KLA)^2})
\end{aligned}$$

The main claim follows. □

7.2 Proofs for CPT-SPSA-G

To prove the main result in Theorem 2, 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 **?**) since we have a non-zero bias in the function evaluations, while the regular SPSA assumes the noise is zero-mean. Following this lemma, we complete the proof of Theorem 2 by invoking the well-known Kushner-Clark lemma **?**.

P: The bias control is with high prob and so probably all conv claims shoud be with high prob. Probably there is a simpler way out, but i dont know (yet)

Lemma 6. *Let $\mathcal{F}_n=\sigma(\theta_m, m\leq n)$, $n\geq 1$. Then, for any $i=1,\ldots,d$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}-\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_i\mathbb{C}(X^{\theta_n})\right|\to 0 \text{ as } n\to\infty. \tag{37}$$

Proof. Recall that the CPT-value estimation scheme is biased, i.e., providing samples with policy θ, we obtain its CPT-value estimate as $V^\theta(x_0)+\epsilon^\theta$. Here ϵ^θ denotes the bias.

We claim

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}-\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\,|\,\mathcal{F}_n\right]=\mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n})-\mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i}\,|\,\mathcal{F}_n\right]+\mathbb{E}\left[\eta_n\,|\,\mathcal{F}_n\right], \tag{38}$$

where $\eta_n=\left(\frac{\epsilon^{\theta_n+\delta_n\Delta}-\epsilon^{\theta_n-\delta_n\Delta}}{2\delta_n\Delta_n^i}\right)$ is the bias arising out of the empirical distribution based CPT-value estimation scheme. From Proposition 2 and the fact that $\frac{1}{m_n^{\alpha/2}\delta_n}\to 0$ by assumption (A3), we have that η_n

goes to zero asymptotically. In other words,

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] \xrightarrow{n\to\infty} \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right]. \tag{39}$$

We now analyse the RHS of (39). By using suitable Taylor's expansions,

$$\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) = \mathbb{C}(X^{\theta_n}) + \delta_n\Delta_n^\intercal\nabla\mathbb{C}(X^{\theta_n}) + \frac{\delta^2}{2}\Delta_n^\intercal\nabla^2\mathbb{C}(X^{\theta_n})\Delta_n + O(\delta_n^3),$$
$$\mathbb{C}(X^{\theta_n-\delta_n\Delta_n}) = \mathbb{C}(X^{\theta_n}) - \delta_n\Delta_n^\intercal\nabla\mathbb{C}(X^{\theta_n}) + \frac{\delta^2}{2}\Delta_n^\intercal\nabla^2\mathbb{C}(X^{\theta_n})\Delta_n + O(\delta_n^3).$$

From the above, it is easy to see that

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

Taking conditional expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] =& \nabla_i\mathbb{C}(X^{\theta_n}) + \mathbb{E}\left[\sum_{j=1,j\neq i}^{N}\frac{\Delta_n^j}{\Delta_n^i}\right]\nabla_j\mathbb{C}(X^{\theta_n}) + O(\delta_n^2)\\ =& \nabla_i\mathbb{C}(X^{\theta_n}) + O(\delta_n^2).\end{aligned} \tag{40}$$

The first equality above follows from the fact that Δ_n is distributed according to a d-dimensional vector of Rademacher random variables 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 of Theorem 2

Proof. We first rewrite the update rule (15) as follows: For $i = 1, \ldots, d$,

$$\theta_{n+1}^i = \theta_n^i - \gamma_n(\nabla_i\mathbb{C}(X^{\theta_n}) + \beta_n + \xi_n), \tag{41}$$

where

$$\begin{aligned}\beta_n &= \mathbb{E}\left(\frac{(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right) - \nabla\mathbb{C}(X^{\theta_n}), \text{ and}\\ \xi_n &= \left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right) - \mathbb{E}\left(\frac{(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right).\end{aligned}$$

In the above, β_n is the bias in the gradient estimate due to SPSA and ξ_n is a martingale difference sequence..

Convergence of (41) can be inferred from Theorem 5.3.1 on pp. 191-196 of **?**, provided we verify the necessary assumptions given as (B1)-(B5) below:

(B1) $\nabla\mathbb{C}(X^\theta)$ is a continuous $\mathbb{R}^d$-valued function.

P: Don't know how this gets verfied in our setting. Help!

(B2) The sequence $\beta_n, n \geq 0$ is a bounded random sequence with $\beta_n \to 0$ almost surely as $n \to \infty$.

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

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

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

(B5) There exists a compact subset K which is the set of asymptotically stable equilibrium points for the following ODE:

$$\dot{\theta}_t^i = \check{\Gamma}_i\left(-\nabla\mathbb{C}(X^{\theta_t^i})\right), \text{ for } i = 1, \ldots, d, \tag{42}$$

In the following, we verify the above assumptions for the recursion (15):

- (B1) holds by assumption in our setting.
- Lemma 6 above establishes that the bias β_n is $O(\delta_n^2)$ and since $\delta_n \to 0$ as $n \to \infty$, it is easy to see that (B2) is satisfied for β_n.
- (B3) holds by assumption (A3).
- We verify (B4) using arguments similar to those used in **?** for the classic SPSA algorithm:
  We first recall Doob's martingale inequality (see (2.1.7) on pp. 27 of **?**):

$$P\left(\sup_{m\geq 0}\|W_l\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\lim_{l\to\infty}\mathbb{E}\|W_l\|^2. \tag{43}$$

Applying the above inequality to the martingale sequence $\{W_l\}$, where $W_l := \sum_{n=0}^{l-1}\gamma_n\eta_n$, $l \geq 1$, we obtain

$$P\left(\sup_{l\geq k}\left\|\sum_{n=k}^{l}\gamma_n\xi_n\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{n=k}^{\infty}\gamma_n\xi_n\right\|^2 = \frac{1}{\epsilon^2}\sum_{n=k}^{\infty}\gamma_n^2\mathbb{E}\|\eta_n\|^2. \tag{44}$$

The last equality above follows by observing that, for $m < n$, $\mathbb{E}(\xi_m\xi_n) = \mathbb{E}(\xi_m\mathbb{E}(\xi_n \mid \mathcal{F}_n)) = 0$. We now bound $\mathbb{E}\|\xi_n\|^2$ as follows:

$$\begin{aligned}
\mathbb{E}\|\xi_n\|^2 \leq & \mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2 && (45)\\
\leq & \left(\left(\mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2} + \left(\mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2}\right)^2 && (46)\\
\leq & \frac{1}{4\delta_n^2}\left[\mathbb{E}\left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}}\right)\right]^{\frac{1}{1+\alpha 1}} \\
& \times\left(\left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) && (47)\\
\leq & \frac{1}{4\delta_n^2}\left(\left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) && (48)\\
\leq & \frac{C}{\delta_n^2}, \text{ for some } C < \infty. && (49)
\end{aligned}$$

The inequality in (45) 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 (46) 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 (47) uses Holder's inequality, with $\alpha_1, \alpha_2 > 0$ satisfying $\frac{1}{1+\alpha_1} + \frac{1}{1+\alpha_2} = 1$. The equality in (48) above follows owing to the fact that $\mathbb{E}\left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}}\right) = 1$ as Δ_n^i is Rademacher. The inequality in (49) follows by using the fact that, for any θ, the CPT-value estimate $\widehat{\mathbb{C}}(D^\theta) = \mathbb{C}(D^\theta) + \epsilon^\theta$. We assume a finite state-action spaced SSP (which implies that the costs $\max_{s,a} g(s,a) < \infty$) and consider only *proper* policies (which implies that the total cost D^θ is bounded for any policy θ) and finally, by (A1), the weight functions are Lipschitz - these together imply that $\mathbb{C}(D^\theta)$ is bounded for any policy θ. The bias ϵ^θ is bounded by Proposition 2 in the main paper.

P: Need to update the above arguments for the general case of X^θ, with θ in a compact set

Thus, $\mathbb{E}\left\|\xi_n\right\|^2 \leq \frac{C}{\delta_n^2}$ for some $C < \infty$. Plugging this in (44), we obtain

$$\lim_{k\to\infty} P\left(\sup_{l\geq k}\left\|\sum_{n=k}^{l}\gamma_n\xi_n\right\| \geq \epsilon\right) \leq \frac{dC}{\epsilon^2}\lim_{k\to\infty}\sum_{n=k}^{\infty}\frac{\gamma_n^2}{\delta_n^2} = 0.$$

The equality above follows from (A3) in the main paper.

- Observe that $\mathbb{C}(X^\theta)$ serves as a strict Lyapunov function for the ODE (42). This can be seen as follows:
$$\frac{d\mathbb{C}(X^\theta)}{dt} = \nabla\mathbb{C}(X^\theta)\dot{\theta} = \nabla\mathbb{C}(X^\theta)\check{\Gamma}\left(-\nabla\mathbb{C}(X^\theta\right) < 0.$$
Hence, the set $\mathcal{K} = \{\theta \mid \check{\Gamma}_i\left(-\nabla\mathbb{C}(X^\theta)\right) = 0, \forall i = 1,\ldots,d\}$ serves as the asymptotically stable attractor for the ODE (42).

The claim follows from the Kushner-Clark lemma. □

7.3 Proofs for CPT-SPSA-N

Before proving Theorem 3, we bound the bias in the SPSA based estimate of the Hessian in the following lemma.

Lemma 7. *For any $i, j = 1,\ldots,d$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)} - 2\widehat{V}_n^{\theta_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\middle|\mathcal{F}_n\right] - \nabla_{i,j}^2\mathbb{C}(X^{\theta_n})\right| \to 0 \text{ as } n\to\infty. \tag{50}$$

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

$$\begin{aligned}&\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)} - 2\widehat{V}_n^{\theta_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right]\\ &\xrightarrow{n\to\infty} \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}) + \mathbb{C}(X^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}) - 2\mathbb{C}(X^{\theta_n})}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right].\end{aligned} \tag{51}$$

Now, the RHS of (51) 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 **?**. We provide the proof here for the sake

of completeness. Using Taylor's expansion as in Lemma 6, we obtain

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

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

Lemma 8. *For any $i=1,\ldots,d$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n)}-\overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_i\mathbb{C}(X^{\theta_n})\right|\to 0 \text{ as } n\to\infty. \tag{52}$$

Proof. As in the proof of Lemma 6, 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 6. □

Proof of Theorem 3

Before we prove Theorem 3, we show that the Hessian recursion (17) converges to the true Hessian, for any policy θ.

Lemma 9. *For any $i,j=1,\ldots,d$, we have almost surely,*

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

Proof. Follows in a similar manner as in the proofs of Lemmas 7.10 and 7.11 of **?**. □

Proof. ***(Theorem 3)*** The proof follows in a similar manner as the proof of Theorem 7.1 in **?**; we provide a sketch below for the sake of completeness.

We first rewrite the recursion (16) as follows: For $i=1,\ldots,d$

$$\theta_{n+1}^i=\Gamma_i\left(\theta_n^i-\gamma_n\sum_{j=1}^{d}\bar{M}^{i,j}(\theta_n)\nabla_j\mathbb{C}(X_n^\theta)+\gamma_n\zeta_n+\chi_{n+1}-\chi_n\right), \tag{53}$$

where

$$\begin{aligned}
\bar{M}^{i,j}(\theta) =& \Upsilon(\nabla^2\mathbb{C}(X^\theta))^{-1} \\
\chi_n =& \sum_{m=0}^{n-1}\gamma_m\sum_{k=1}^{d}\bar{M}_{i,k}(\theta_m)\Bigg(\frac{\mathbb{C}(X^{\theta_m-\delta_m\Delta_m-\delta_m\widehat{\Delta}_m})-\mathbb{C}(X^{\theta_m+\delta_m\Delta_m+\delta_m\widehat{\Delta}_m})}{2\delta_m\Delta_m^k} \\
& - E\left[\frac{\mathbb{C}(X^{\theta_m-\delta_m\Delta_m-\delta_m\widehat{\Delta}_m})-\mathbb{C}(X^{\theta_m+\delta_m\Delta_m+\delta_m\widehat{\Delta}_m})}{2\delta_m\Delta_m^k}\mid\mathcal{F}_m\right]\Bigg) \text{ and} \\
\zeta_n =& \mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n)}-\overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_i\mathbb{C}(X^{\theta_n}).
\end{aligned}$$

In lieu of Lemmas 7–9, 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 (53) is a discretization of the ODE:

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

Further, $\mathcal{K} = \{\theta \mid \nabla\mathbb{C}(X^{\theta^i})\check{\Gamma}_i\left(\Upsilon(\nabla^2\mathbb{C}(X^\theta))^{-1}\nabla\mathbb{C}(X^{\theta^i})\right) = 0, \forall i = 1,\ldots,d\}$ serves as an asymptotically stable attractor set and the claim follows by applying Kusher-Clark lemma to (53). □

7.4 Proofs for gradient-free policy optimization algorithm

We begin by remarking that there is one crucial difference between our algorithm and MRAS$_2$ from **?**: MRAS$_2$ 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.

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 4 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 **?**, adapted to our setting.

Lemma 10. *The sequence of random variables $\{\theta_n^*, n = 0, 1, \ldots\}$ in Algorithm 4 converges w.p.1 as $n \to \infty$.*

Proof. Let $\mathcal{A}_n$ be the event that either the first if statement (see 16) is true or the second if statement in the else clause (see 21) is true within the Thresholding step of Algorithm 4. 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

$$
\begin{aligned}
P(\mathcal{A}_n \cap \mathcal{B}_n) \leq & P\left(\{\overline{\mathbb{C}}_n^{\theta_n^*}(x^0) - \overline{\mathbb{C}}_{n-1}^{\theta_{n-1}^*} \geq \varepsilon\} \cap \left\{\mathbb{C}(X^{\theta_n^*} - \mathbb{C}(X^{\theta_{n-1}^*} \leq \frac{\varepsilon}{2}\right\}\right) \\
\leq & P\Big(\bigcup_{\theta\in\Lambda_n, \theta'\in\Lambda_{n-1}} \left\{\{\overline{\mathbb{C}}_n^{\theta}(x^0) - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\} \cap \left\{\mathbb{C}(X^{\theta}) - V^{\theta'}(x^0) \leq \frac{\varepsilon}{2}\right\}\right\}\Big) \\
\leq & \sum_{\theta\in\Lambda_n, \theta'\in\Lambda_{k-1}} P\left(\{\overline{\mathbb{C}}_n^{\theta}(x^0) - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\} \cap \left\{\mathbb{C}(X^{\theta}) - V^{\theta'}(x^0) \leq \frac{\varepsilon}{2}\right\}\right) \\
\leq & |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} P\left(\{\overline{\mathbb{C}}_n^{\theta}(x^0) - \overline{\mathbb{C}}_{n-1}^{\theta'} \geq \varepsilon\} \cap \left\{\mathbb{C}(X^{\theta}) - V^{\theta'}(x^0) \leq \frac{\varepsilon}{2}\right\}\right) \\
\leq & |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} P\left(\overline{\mathbb{C}}_n^{\theta}(x^0) - \overline{\mathbb{C}}_{n-1}^{\theta'} - \mathbb{C}(X^{\theta}) + V^{\theta'}(x^0) \geq \frac{\varepsilon}{2}\right) \\
\leq & |\Lambda_n||\Lambda_{n-1}| \sup_{\theta,\theta'\in\Theta} \left(P\left(\overline{\mathbb{C}}_n^{\theta}(x^0) - \mathbb{C}(X^{\theta}) \geq \frac{\varepsilon}{4}\right) + P\left(\overline{\mathbb{C}}_{n-1}^{\theta'} - V^{\theta'}(x^0) \geq \frac{\varepsilon}{4}\right)\right) \\
\leq & 4|\Lambda_n||\Lambda_{k-1}| e^{-m_n \frac{\epsilon^2}{8L^2M^2}}.
\end{aligned}
$$

From the foregoing, we have $\sum_{n=1}^{\infty} P\left(\mathcal{A}_n \cap \mathcal{B}_n\right) < \infty$ since $m_n \to \infty$ as $n \to \infty$. Applying the Borel-Cantelli lemma, we obtain

$$
P\left(\mathcal{A}_n \cap \mathcal{B}_n \text{ i.o.}\right) = 0.
$$

From the above, it follows that if $\mathcal{A}_n$ happens infinitely often, then $\mathcal{B}_n^c$ will also happen infinitely often. Hence,

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

In the above, the first equality follows from the fact that if the else clause in the second if statement (see 23) in Algorithm 4 is hit, then $\theta_n^* = \theta_{n-1}^*$. From the last equality above, we conclude that it is a contradiction because, $\mathbb{C}(X^{\theta}) > V^{\theta^*}(x^0)$ for any θ (since θ^* is the global minimum). The main claim now follows since $\mathcal{A}_n$ can happen only a finite number of times. □

Proof of Theorem 4

Proof. Once we have established Lemma 10, the rest of the proof follows in an identical fashion as the proof of Corollary 4.18 of **?**. This is because our algorithm operates in a similar manner as MRAS$_2$ with respect to 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 **?**. □

8 Simulation Experiments

8.1 Simulation Setup

We consider a SSP version of an example[5] for buying a house at the optimal price. Suppose the house is priced at x_k at 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 $D^\theta(x^0) = \sum_{k=0}^{\tau} \left(I_{\{a_k=b\}} x_k + I_{\{a_k=w\}} h \right) + I_{\{\tau=T\}} x_T$, where $\tau = \{k | \theta(x_k) = 1\} \wedge T$. We set $T = 20, h = 0.1, C_{up} = 2, C_{down} = 0.5$, and $x_0 = 1$.

Implementation: On this example, we implement the first-order CPT-SPSA-G 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 CPT-value using simulated episodes. We implement the algorithm from **?** for the SSP example described in the numerical experiments of the main paper. The latter, henceforth referred to as NoCPT-SPSA-G, is an SPSA-based scheme that optimizes the traditional value function objective in a discounted MDP setting and we make a trivial adaptation of this algorithm for the SSP setting. For CPT-SPSA-G and NoCPT-SPSA-G, 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 θ_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-G and 333 iterations of CPT-SPSA-N. The results presented are averages over 500 independent simulations. For CPT-SPSA-G/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.

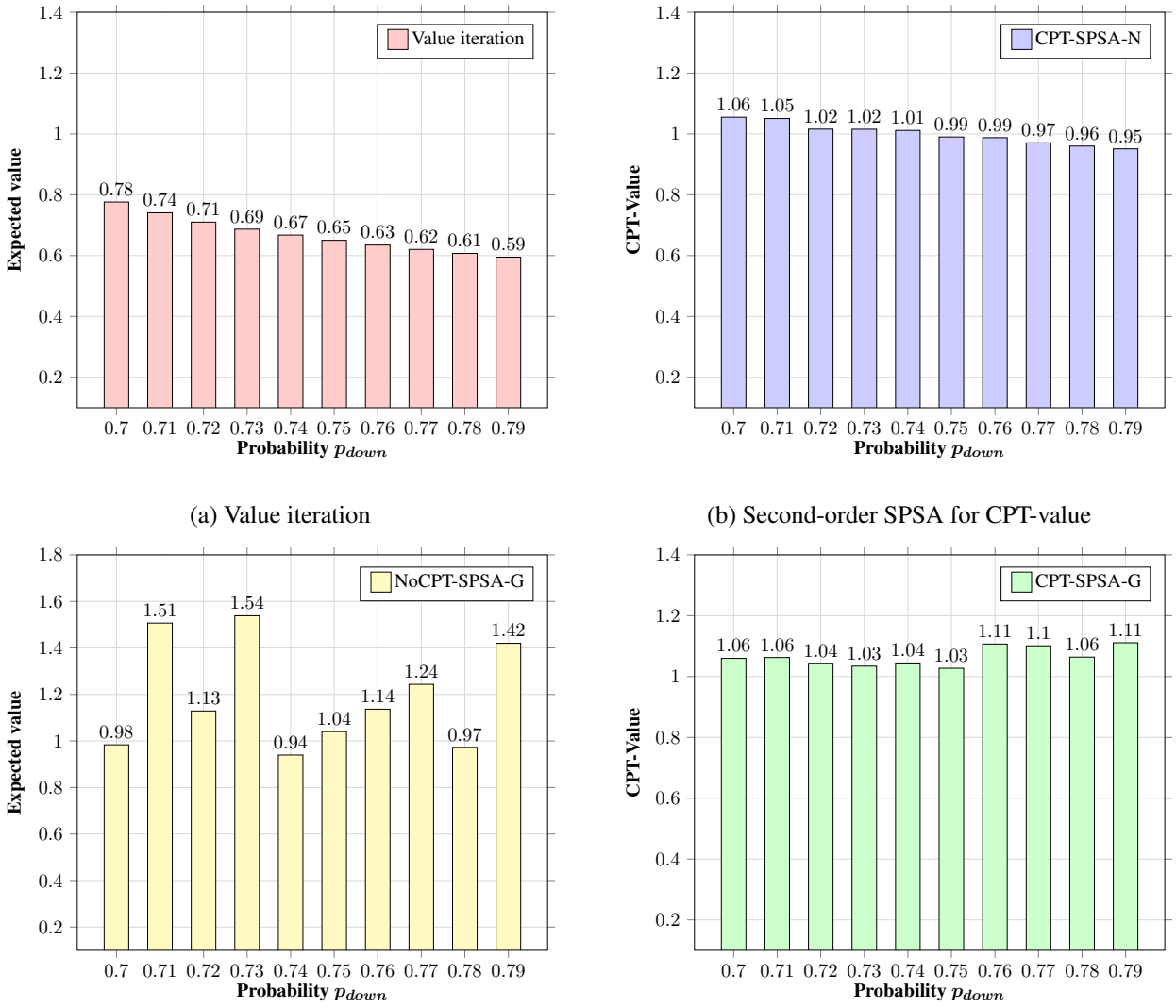
8.2 Results

Figures 6a–6c present the value function computed using value iteration and NoCPT-SPSA-G, while Figures 6d–6b present the CPT-value $V^{\theta_{end}}(x^0)$ for CPT-SPSA-G and CPT-SPSA-N, respectively. The performance plots are for various values of p_{down}, the probability of house price going down. From Figure 6a, we notice that the variations in expected total cost is larger in comparison to that in CPT-value. Figure 6c implies that a similar observation about variation of expected value holds true for NoCPT-SPSA-G algorithm from **?**. 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 PG/CPT-SPSA-N resulted in randomized policies that smoothly transitioned with changes in p_{down}. As motivated in the introduction, these plots 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-G. Finally, what is not shown is that the CPT-value obtained for PG/CPT-SPSA-N is much lower than that obtained for NoCPT-SPSA-G, thus making apparent the need for specialized algorithms that incorporate CPT-based criteria.

9 Conclusions and Future Work

CPT has been a very popular paradigm for modeling human decisions among psychologists/economists, but has escaped the radar of the AI community. This work is the first step in incorporating CPT-based criteria into an RL framework. However, both estimation and control of CPT-based value is challenging.

[5] A similar example has been considered in **?**.



(a) Value iteration

(b) Second-order SPSA for CPT-value

(c) SPSA for regular value function

(d) First-order SPSA for CPT-value

Figure 6: Performance of policy gradient algorithms with/without CPT for different down probabilities of the SSP

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 illustrated the usefulness of CPT-based criteria in a numerical example.