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

Risk-sensitive reinforcement learning (RL) has received a lot of attention recently (cf. **????**). Previous works consider either an exponential utility formulation (cf. **?**) that implicitly controls the variance or a constrained formulation with explicit constraints on the variance of the cost-to-go (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. **??**).

P: Add refs for distorted weights

In this paper, we consider a risk measure based on *cumulative prospect theory* (CPT) **?**, which is a non-coherent and non-convex measure that is well known among psychologists and economists to be a good

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

model for human decision-making systems, with strong empirical support. In this paper, we incorporate CPT-based criteria into the classic objective *value function* in a reinforcement learning framework. Intuitively this combination is appealing because it taps into the notion of how humans evaluate outcomes and also, a CPT objective leads to a randomized policy, which although harder to estimate often leads to more intuitively appealing behavior, as illustrated via an example below and also the numerical experiments later.

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.

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 the CPT-value for a given policy is challenging, because it requires that the *entire* distribution of the total cost to be estimated.

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*.

In the following, we formalize the notion of CPT-value and then outline our schemes for estimation and control.

Setting: 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 $\pi = \left(\pi^1, \ldots, \pi^{\mathcal{L}\mathcal{M}}\right)^\intercal$ be a randomized policy, where π^i denotes the probability of choosing action $(i\%\mathcal{M})$ in state $\lceil i/\mathcal{M} \rceil$, with $\sum_{j=(i-1)\mathcal{M}+1}^{i\mathcal{M}} \pi^j = 1$, for $i = 1, \ldots, \mathcal{L}$. Let $D^\pi(x^0)$ be a random variable (r.v) that denotes the total cost from an episode simulated using policy π, i.e.,

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

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

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

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

where Π is the set of admissible policies that are *proper*[1].

In this paper, we adopt the CPT approach and aim to solve the following problem:

$$\min_{\pi \in \Pi} V^\pi(x^0),$$

where the CPT-value function $V^\pi(x^0)$ is defined as

$$V^\pi(x^0) = \int_0^{+\infty} w^+(P(u^+(D^\pi(x^0))) > z) dz - \int_0^{+\infty} w^-(P(u^-(D^\pi(x^0))) > z) dz. \tag{1}$$

Let us deconstruct the above definition:

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

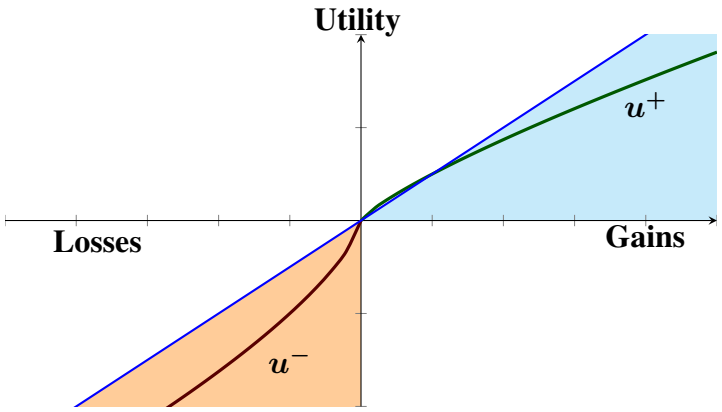


Figure 1: Utility function

Utility functions: u^+, u^- are utility functions corresponding to gains ($D^\pi(x^0)) \geq 0$) and losses ($D^\pi(x^0)) \leq 0$), 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 1. In contrast, the traditional value function makes no such distinction between gains and losses.

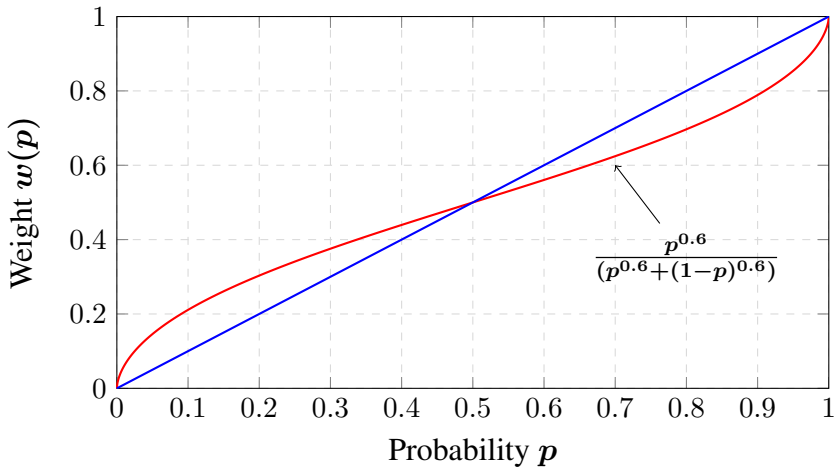


Figure 2: Weight function

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 \$10000 w.p. 0.001 over one that gives \$10 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 2, the weight functions are continuous,

non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. 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).[2]

[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).

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 (1) makes it the expectation of the total cost $D^\pi(x^0))$. It is also possible to get (1) to coincide with coherent risk measures (e.g. CVaR) by the appropriate choice of weight functions.

Sensitivity: Apart from the fact that CPT models human decisions well, it is also less 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, esp. considering model information is unavailable in practice. Note that there exists a deterministic policy that gives the optimal value function, while with a CPT-value objective, the optimal policy is not necessarily deterministic[3].

We summarize our contributions below.

Prediction: While one obtains samples of $D^\pi(x^0)$, the CPT-value integrals in (1) involve a distribution that is distorted using non-linear weight functions. Thus, one cannot just do sample means and hence, cannot employ classic stochastic approximation schemes (e.g. temporal difference learning).

Solution: We derive an estimate of the first integral in (1) as follows: first compute the empirical distribution function for $u^+(\cdot)$, then compose it with the weight function w^+ and finally, integrate the resulting composition to obtain the final estimate. The second integral in (1) is estimated in a similar fashion, and the CPT-value estimate is the difference in the estimates of the two integrals in (1). Assuming that the weight functions are Lipschitz, we establish convergence (asymptotic) of our CPT-value estimate to the true CPT-value. We also provide a sample complexity result that establishes that $O\left(\frac{1}{\epsilon^2}\right)$ samples are required to be ϵ-close to the CPT-value with high probability.

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

[3] See also the organ transplant example on pp. 75-81 of **?**,

Control: Optimizing the CPT-value is challenging owing to the following reasons:

(i) *Biased policy evaluation:* Since policy optimization is an iterative procedure, one can only simulate a finite number of episodes for a fixed policy, say π_n in iteration n and use the empirical distribution function (EDF) based scheme to provide an estimate of the CPT-value $V^{\pi_n}(x^0)$. But, such a finite sample estimate clearly results in a bias, which is bounded.

(ii) *Simulation optimization:* Given only an estimate of the CPT-value for any policy, it is necessary to devise an adaptive search scheme that improves the policy iteratively.

(iii) *Non-dynamic programming:* As mentioned earlier, dynamic programming approaches cannot be employed as there is no "Bellman equation" for CPT-value.

Solution: We increase the number of SSP episodes m_n simulated in each iteration n of the policy optimization algorithms such that the bias of CPT-value estimates vanishes asymptotically (See the technical condition on m_n in (A3)).

Using two well-known ideas from the *simulation optimization* literature **?**, we propose three policy optimization algorithms that overcome the second and third problems. Our proposed algorithms are summarized as follows:

Gradient-based methods: We propose two algorithms in this class. The first is a policy gradient algorithm that employs simultaneous perturbation stochastic approximation (SPSA)-based estimates for the gradient of the CPT-value, while the second is a policy Newton algorithm that also uses SPSA-based estimates of the gradient and also the Hessian. We remark here that, unlike traditional settings for SPSA, our estimates for CPT-value have a non-zero (albeit bounded) 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 optimizing policy update 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 policy (assuming it exists).

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 (1) 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 describe the empirical distribution based scheme for estimating the CPT-value of any random variable. In Sections 3–4, we present the gradient-based algorithms for optimizing the CPT-value. Next, in Section 5, 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 6. We present the results from numerical experiments for the CPT-value estimation scheme in Section **??** and finally, provide the concluding remarks in Section **??**.

2 CPT-value estimation

For the sake of notational simplicity, we let X denote the r.v. $D^\pi(x^0)$, where the policy π is assumed to be fixed.

On integrability Observe that the first integral in (1), i.e.,

$$\int_0^{+\infty} w^+(P(u^+(X) > z))dz \tag{2}$$

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 (2) 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 (1) 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 (1) 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.

2.1 Estimation scheme for Lipschitz continuous weights

As mentioned before, since the integrals in (1) require the distribution to be estimated over the entire domain, we use the EDF to approximate the distribution and then perform an integration of the weight-distorted EDF. Propositions 1 and 2 below establish that the resulting CPT-value estimate converges and also with the canonical Monte Carlo convergence rate.

Let X_i, $i = 1, \ldots, n$ denote n samples of the random variable X. The EDF for $u^+(X)$ and $u^-(X)$, for any given real-valued functions u^+ and u^-, is defined as follows:

$$\hat{F}_n^+(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(u^+(X_i)\leq x)}, \hat{F}_n^-(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(u^-(X_i)\leq x)}. \tag{3}$$

Using EDFs, the CPT-value is estimated as follows:

$$\widehat{V}_n(X) = \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx - \int_0^{+\infty} w^-(1 - \hat{F}_n^-(x))dx. \tag{4}$$

Notice that we have substituted $1 - \hat{F}_n^+(x)$ (resp.$1 - \hat{F}_n^-(x)$) for $P(u^+(X) > x)$ (resp. $P(u^-(X) > x)$) in (1) and then performed an integration of the complementary EDF composed with the weight function.

Main results

As mentioned earlier, to overcome integrability issues, we make the following assumption:

Assumption (A1). The weight functions w^+, w^- are Lipschitz with common constant L.

For the convergence rate results below, we require the following assumption:

Assumption (A2). The utility functions $u^+(X)$ and $u^-(X)$ are bounded above by $M < \infty$.

The following result shows that the estimate (4) converges to the true CPT value almost surely and at the (nearly) canonical Monte Carlo asymptotic rate.

Proposition 1. ***(Asymptotic convergence and rate.)*** *Under (A1), we have*

$$\widehat{V}_n(X) \to V(X) \text{ a.s. as } n \to \infty. \tag{5}$$

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

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2\ln\ln n}} ||\hat{V}_n(X) - V(X)||_\infty \leq LM \quad \text{a.s.}$$

Proof. See Section 6.1. □

While the above result establishes that (4) is an unbiased estimate in the asymptotic sense, it is important to know the rate at which the estimate in (4) converges to the CPT-value. The following sample complexity result shows that $O\left(\frac{1}{\epsilon^2}\right)$ number of samples are required to be ϵ-close to the CPT-value in high probability.

Proposition 2. ***(Sample Complexity)*** *Under (A1) and (A2), for any $\epsilon, \delta > 0$, we have*

$$P(|\widehat{V}_n(X) - V(X)| \leq \epsilon) \geq 1 - \delta, \;\; \forall n \geq \frac{2L^2M^2}{\epsilon^2} ln\frac{4}{\delta}.$$

Proof. See Section 6.1. □

2.2 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*

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

In order to ensure integrability of the CPT-value (1), we make the following assumption:

Assumption (A1'). The weight functions w^+, w^- are Hölder continuous with common level α. Further, $\exists \gamma < \alpha$ 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 (1) is finite - see Proposition 7 in Section 6.1.2 for a formal proof.

Approximating CPT-value using quantiles: Let $\xi^+_{\frac{i}{n}}$ denote the $\frac{i}{n}$th quantile of the r.v. $u^+(X)$. Then, it can be seen that (see Proposition 8 in Section 6.1.2)

$$\lim_{n\to\infty} \sum_{0}^{n-1} \xi^+_{\frac{i}{n}} \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right)\right) = \int_0^{+\infty} w^+(P(u^+(X) > z))dz. \tag{6}$$

The identicalal property holds for the pairs $u^-(X)$, $\xi^-_{\frac{i}{n}}, w^-$, with $\xi^-_{\frac{i}{n}}$ denote the $\frac{i}{n}$th 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 (6). The estimation scheme is presented in Algorithm 1.

Algorithm 1 CPT-value estimation for Hölder continuous weights

1: Simulate n random samples with distribution X, sort them and denoted the ordered sample as $X_{[1]}, X_{[2]}, \dots X_{[n]}$.
2: Calculate $u^+(X_{[1]}), \dots u^+(X_{[n]})$
3: Order the simulated samples and label them as follows: $u^+(X_{[1]}), \dots, u^+(X_{[n]})$.
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: Return the statistic $\hat{V}^+_n(X) := \sum_1^{n-1} u^+(X_{[i]})(w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))$
6: Repeat the procedure on the sequence $X_{[1]}, X_{[2]}, \dots X_{[n]}$, with respect to the function u^-, and denote the statistic $\hat{V}^-_n(X) := \sum_1^{n-1} u^-(X_{[i]})(w^-(\frac{n-i}{n}) - w^-(\frac{n-i-1}{n}))$
7: Denote the statistic $\hat{V}_n(X) = \hat{V}^+_n(X) - \hat{V}^-_n(X)$.

Main results

Proposition 3. *(**Asymptotic convergence.**) Assume (A1') and also that $F^+(\cdot)$,$F^-(\cdot)$ - the distribution function of $u^+(X)$, and $u^-(X)$ are Lipschitz continuous with constant L^+ and L^-, respectively on the interval $(0, +\infty)$, and $(-\infty, 0)$. Then, we have*

$$\lim_{n\to+\infty} \hat{V}_n(X) = V(X) = \int_0^{+\infty} w^+(P(u^+(X) > z))dz - \int_0^{+\infty} w^-(P(u^-(X) > z))dz, \text{ a.s.} \tag{7}$$

holds, where $\hat{V}_n$ is as defined in Algorithm 1.

Proof. See Section 6.1.2. □

While the above proposition gives an asymptotic guarantee, in the following we provide a sample complexity result for Algorithm 1. For this result, we assume, as in the case of Lipschitz continuous weights, that the utility functions are bounded in addition to (A1').

Proposition 4. *(**Sample complexity.**) Assume (A1') and (A2). Then, $\forall \epsilon, \delta$, we have*

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

.

Proof. Notice the the following equivalence:

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

and also,

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

where $\hat{F_n^+}(x)$ and $\hat{F_n^-}(x)$ is the empirical distribution of $u^+(X)$ and $u^-(X)$. (see (3)). The above equality in conjunction with the well-known DKW inequality implies the following claim: See Section 6.1.2 for details. □

Corollary 1 (Lipschitz case)**.** *Assume the function $w^+(x)$ and $w^-(x)$ are Lipchitz continuous, then the asymptotic convergence property in proposition 3 holds, and in addition, we have*

$$P(\Big|\hat{V}_n(X) - V(X)\Big| \leq \epsilon) > \delta \,, \forall n \geq \ln(\frac{1}{\delta}) \cdot \frac{4L^2M^2}{\epsilon^2}$$

as the indicator of sample complexity of the estimation scheme defined in algorithm 1.

2.3 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. 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, ***Prospect theory*** (PT) value is defined as $V(X) = \sum_{i=1}^K u(x_i)w(p_i)$. As explained in the introduction, the idea is to take an utility function that is S-shaped, so that it satisfies the *diminishing sensitivity* property. If we take the weighting function w to be the identity, then one recovers the classic expected utility. A general weight function inflates low probabilities and deflates high probabilities and this has been shown to be close to the way humans make decisions (see **?**, **?** for a justification, in particular via empirical tests using human subjects). However, PT is lacking in some theoretical aspects as it violates first-order *stochastic dominance*.[4]

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

$$V(X) = (u^-(x_1)) \cdot w^-(p_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\sum_{j=1}^{i} p_j) - w^-(\sum_{j=1}^{i-1} p_j)\Big) \tag{8}$$

$$+ \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\sum_{j=i}^{K} p_j) - w^-(\sum_{j=i+1}^{K} p_j)\Big) + u^+(x_K) \cdot w^+(p_K), \tag{9}$$

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,

[4]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.

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.[5]

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

$$\hat{V}_n(X) = u^-(x_1)\cdot w^-(\hat{p}_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\sum_{j=1}^{i}\hat{p}_j) - w^-(\sum_{j=1}^{i-1}\hat{p}_j)\Big) \tag{10}$$

$$+ \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\sum_{j=i}^{K}\hat{p}_j) - w^-(\sum_{j=i+1}^{K}\hat{p}_j)\Big) + u^+(x_K)\cdot w^+(\hat{p}_K). \tag{11}$$

Owing to the fact that $\hat{p_k}$ converge a.e to $p_k = P(X_i = x_k)$, with X_i be the sample of X the above estimator obtains strong consistency property according to continuous mapping theorem.

Sample Complexity Before exploring on the convergence speed to the sample estimator, it is necessary to introduce Hoeffding's inequality:

Lemma 2. *Let $Y_1, ...Y_n$ be* independent random variables *satisfying $P(a \le Y_i \le b) = 1$, each i, where $a < b$.Then for $t > 0$,*

$$P\Big(\Big|\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} E(Y_i)\Big| \ge nt\Big) \le 2\exp\{-2nt^2/(b-a)^2\}$$

Notations: We will introduce

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

and $\hat{F}_k$ retains the same form as F_k by replacing p_k by $\hat{p}_k$

The Hoeffding inequality suggests the following proposition:

Proposition 5. *Let F_k and $\hat{F}_k$ as introduced above, Then for every $\epsilon > 0$,*

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

Proof. We focus on the case when $k > l$,and the case of $k \le l$ will be proved through the same fashion. Notice that when $k > l$, $\hat{F}_k = I_{(U_i \ge x_k)}$ and the random variables are independent to each other for each i, and it is bounded by 1. The probability $P(|F_k - F_k| > \epsilon)$ is equal to

$$\begin{aligned} & P\Big(\Big|\hat{F}_k - F_k\Big| > \epsilon\Big) \\ &= P\Big(\Big|\frac{1}{n}\sum_{i=1}^{n} I_{\{U_i \ge x_k\}} - \frac{1}{n}\sum_{i=1}^{n} E(I_{\{U_i \ge x_k\}})\Big| > \epsilon\Big) \\ &= P\Big(\Big|\sum_{i=1}^{n} I_{\{U_i \ge x_k\}} - \sum_{i=1}^{n} E(I_{\{U_i \ge x_k\}})\Big| > n\epsilon\Big) \\ &\le 2e^{-2n\epsilon^2} \end{aligned}$$

□

[5] 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.

Proposition 5 gives a convergence rate of $\hat{F}_k$ to the value F_k, regardless of what k is. Additionally, since w^+ and w^- are both locally Lipschitz as indicated in the paper, we can explore the sample complexity of the estimation algorithm through the following lemma:

Theorem 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 locally Lipschitz property with constant L_k holds. For any ϵ, ρ,let $M = \min(\delta^2, \epsilon^2/(KLA)^2)$, and we have*

$$P(\left|\hat{V}_n(X) - V(X)\right| \leq \epsilon) > 1 - \rho \,, \forall n > \frac{\ln(\frac{4K}{a})}{M} \tag{12}$$

Before proving the preceding theorem, we will introduce the following proposition :

Proposition 6. *Following the same notations and conditions introduced in theorem 2, and assume that w is Locally Lipschitz continuous with constants $L_1, ...L_K$ on the points $F_1, ...F_K$ as written in the statement of theorem 2, 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,

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

And as a result,

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

□

By giving the above proposition, we can prove theorem 2:

Proof of theorem 2: Since the functions w^- and w^+ all locally Lipschitz in the according points and with the according constants introduced in theorem 2, it is suggestive only to write w uniformly in place of w^- and w^+ in the separate cases $1 \leq k \leq l$ and $k > l$, in order to avoid unnecessary technicalities. The proof is equivalently to show that

$$
P(\left|\sum_{i=1}^{K} u(x_k) \cdot (w(\hat{F}_k) - w(\hat{F}_{k+1})) - \sum_{i=1}^{K} u(x_k) \cdot (w(F_k) - w(F_{k+1}))\right| \leq \epsilon) > 1 - \rho \, , \forall n > \frac{\ln(\frac{4K}{a})}{M} \tag{13}
$$

under which w is Locally Lipschitz continuous with constants $L_1, ...L_K$ on the points $F_1, ...F_K$ as written in the statement of theorem 2. Observe that by repeating the identical procedure in the proof of proposition 6 one can show that

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

Therefore,

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

And by introducing the notation $M = \min(\delta^2, \epsilon^2/(KLA)^2)$, one can conclude the sample complexity property stated in the theorem.

□

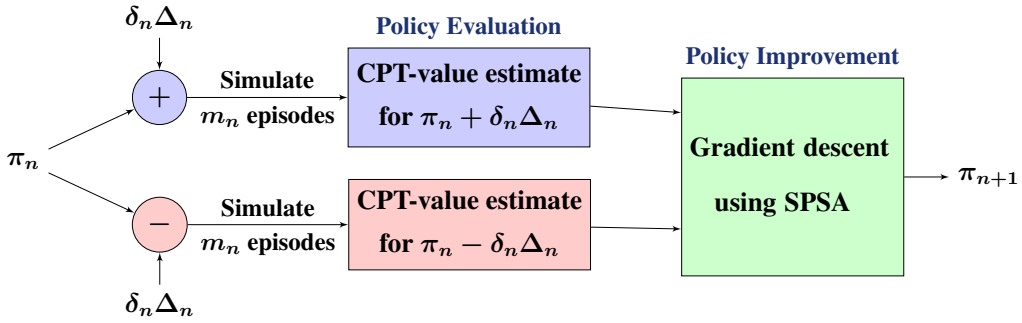


Figure 3: Overall flow of PG-CPT-SPSA.

3 Policy gradient algorithm (PG-CPT-SPSA)

3.1 Gradient estimation

Given that we operate in a learning setting and only have biased estimates of the CPT-value from (4), we require a simulation optimization scheme that estimates $\nabla V_n^{\pi}(x^0)$. 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 values. In our context, at any iteration n of PG-CPT-SPSA, with policy π_n, the gradient $\nabla V_n^{\pi_n}(x^0)$ is estimated as follows: For any state $i = 1, \ldots, \mathcal{LM}$,

$$\widehat{\nabla}_i V^{\pi}(x^0) = \frac{\widehat{V}_n^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}_n^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i},$$

where δ_n is a positive scalar that satisfies (A3) below and $\Delta_n = \left(\Delta_n^1, \ldots, \Delta_n^{\mathcal{LM}}\right)^{\intercal}$, where $\{\Delta_n^i, i = 1, \ldots, \mathcal{LM}\}$ are i.i.d. Rademacher, independent of $\pi_0, \ldots, \pi_n$. The (asymptotic) unbiasedness of the gradient estimate is proven in Lemma 10.

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. **?**, **?**.

3.2 Update rule

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

$$\pi_{n+1}^i = \Gamma_i\left(\pi_n^i - \gamma_n \widehat{\nabla}_i V_n^{\pi_n}(x^0)\right), \tag{14}$$

where γ_n is a step-size chosen to satisfy (A3) below and $\Gamma = (\Gamma_1, \ldots, \Gamma_{\mathcal{LM}})$ is an operator that ensures that the update (14) results in a probability distribution over actions for each state. Fig. 3 illustrates the overall flow of the policy gradient algorithm based on SPSA, while Algorithm 2 presents the pseudocode.

On the number of episodes m_n per iteration: Recall that the CPT-value estimation scheme is biased, i.e., providing samples with policy π_n at instant n, we obtain its CPT-value estimate as $V^{\pi}(x_0) + \epsilon_n^{\pi}$ with ϵ_n^{π} denoting the bias. We rewrite the update rule (14) as follows:

$$\pi_{n+1}^i = \Gamma_i\Bigg(\pi_n^i - \gamma_n\Bigg(\frac{(V^{\pi_n+\delta_n\Delta_n}(x_0) - V^{\pi_n-\delta_n\Delta_n}(x_0))}{2\delta_n\Delta_n^i}\Bigg) + \underbrace{\frac{(\epsilon_n^{\pi_n+\delta_n\Delta_n} - \epsilon_n^{\pi_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i}}_{\kappa_n}\Bigg).$$

Algorithm 2 Structure of PG-CPT-SPSA algorithm.

Input: initial parameter π_0, perturbation constants $\delta_n > 0$, trajectory lengths $\{m_n\}$, 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 $\{\Delta_m, m = 0, 1, \ldots, n-1\}$

Policy Evaluation (Trajectory 1)

Simulate m_n episodes using policy $(\pi_n + \delta_n \Delta_n)$

Obtain CPT-value estimate $\widehat{V}_n^{(\pi_n + \delta_n \Delta_n)}(x^0)$

Policy Evaluation (Trajectory 2)

Simulate m_n episodes using policy $\pi_n - \delta_n \Delta_n$

Obtain CPT-value estimate $\widehat{V}_n^{\pi_n - \delta_n \Delta_n}(x^0))$

Policy Improvement (Gradient descent)

$\pi_{n+1}^i = \Gamma_i \left(\pi_n^i - \gamma_n \widehat{\nabla}_i V_n^{\pi_n}(x^0) \right)$

end for

Return π_n

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^π is bounded above, to establish convergence of the policy gradient recursion (14), 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

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

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 $\gamma > \nu/2$.

3.3 Convergence result

Theorem 4. *Assume (A1)-(A3). Consider the ordinary differential equation (ODE):*

$$\dot{\pi}_t^i = \check{\Gamma}_i \left(\nabla V^{\pi_t^i}(x^0) \right), \textit{for } i = 1, \ldots, \mathcal{LM},$$

where

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

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

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

Proof. See Section 6.2. □

4 Policy Newton algorithm (PN-CPT-SPSA)

4.1 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 greather than $1/3\lambda_{min}(\nabla^2 V^{\pi^*}(x^0))$, then the optimum rate of convergence is not achieved. Here $\lambda_{\min}$ denotes the minimum eigenvalue, while $\pi^* \in \mathcal{K}$ (see Theorem 4). 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 $\gamma_n = 1/n^\alpha$, where $\alpha \in (1/2, 1)$, 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 $\left(\nabla^2 V^{\pi^*}(x^0)\right)^{-1}$, i.e., the inverse of the Hessian of the CPT-value at the optimum π^*. 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 policy optmization scheme.

4.2 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 policy parameters $\pi_n + \delta_n(\Delta_n + \widehat{\Delta}_n)$, $\pi_n - \delta_n(\Delta_n + \widehat{\Delta}_n)$ and π_n, where $\{\Delta_n^i, \widehat{\Delta}_n^i, i = 1, \ldots, \mathcal{LM}\}$ are i.i.d. Rademacher and independent of $\pi_0, \ldots, \pi_n$. Using the CPT-value estimates for the aforementioned policy parameters, we estimate the Hessian and the gradient of the CPT-value function as follows: For $i, j = 1, \ldots, \mathcal{LM}$, set

$$\widehat{\nabla}_i V_n^{\pi_n}(x^0) = \frac{\widehat{V}_n^{\pi_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - \widehat{V}_n^{\pi_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)}{2\delta_n \Delta_n^i},$$

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

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

4.3 Update rule

We update the policy incrementally using a Newton decrement as follows: For $i = 1, \ldots, \mathcal{LM}$,

$$\pi_{n+1}^i = \Gamma_i \left(\pi_n^i - \gamma_n \sum_{j=1}^{\mathcal{LM}} M_n^{i,j} \widehat{\nabla}_j V_n^{\pi}(x^0) \right), \tag{15}$$

$$\overline{H}_n = (1 - \xi_n)\overline{H}_{n-1} + \xi_n \widehat{H}_n, \tag{16}$$

Algorithm 3 Structure of PN-CPT-SPSA algorithm.

Input: initial parameter π_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 $\{\Delta_m, \widehat{\Delta}_m, m = 0, 1, \ldots, n-1\}$
Policy Evaluation (Trajectory 1)
Simulate m_n episodes of the SSP using policy $(\pi_n + \delta_n(\Delta_n + \hat{\Delta}_n))$
Obtain CPT-value estimate $\widehat{V}_n^{(\pi_n+\delta_n(\Delta_n+\hat{\Delta}_n))}(x^0)$ using (4)
Policy Evaluation (Trajectory 2)
Simulate m_n episodes of the SSP using policy $(\pi_n - \delta_n(\Delta_n + \hat{\Delta}_n))$
Obtain CPT-value estimate $\widehat{V}_n^{\pi_n-\delta_n(\Delta_n+\hat{\Delta}_n)}(x^0))$ using (4)
Policy Evaluation (Trajectory 3)
Simulate m_n episodes of the SSP using policy π_n
Obtain CPT-value estimate $\widehat{V}_n^{\pi_n}(x^0))$ using (4)
Policy Improvement (Newton decrement)
Gradient estimate $\widehat{\nabla}_i V_n^{\pi}(x^0) = \dfrac{\widehat{V}_n^{\pi_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - \widehat{V}_n^{\pi_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)}{2\delta_n\Delta_n^i}$
Hessian estimate $\widehat{H}_n = \dfrac{\widehat{V}_n^{\pi_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) + \widehat{V}_n^{\pi_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - 2\widehat{V}_n^{\pi_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}$
Policy update: $\pi_{n+1}^i = \Gamma_i\left(\pi_n^{ij} - \gamma_n\Upsilon(\overline{H}_n)^{-1}\widehat{\nabla}_i V_n^{\pi}(x^0)\right)$
Hessian update: $\overline{H}_n = (1-\xi_n)\overline{H}_{n-1} + \xi_n\widehat{H}_n$
end for
Return π_n

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 π_n in (15) - see (**?**, Chapter 6). Further, Γ is a projection operator as in PG-CPT-SPSA 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 π-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 PN-CPT-SPSA is similar to Fig. 3, except that three system trajectories with a different perturbation sequence are used. Algorithm 3 presents the pseudocode.

4.4 Convergence result

Theorem 5. *Assume (A1)-(A4). Consider the ODE:*

$$\dot{\pi}_t^i = \check{\Gamma}_i \left(\nabla V^{\pi_t^i}(x^0) \Upsilon(\nabla^2 V^{\pi_t}(x^0))^{-1} \nabla V^{\pi_t^i}(x^0) \right), \text{ for } i = 1, \ldots, \mathcal{LM},$$

where $\bar{\Gamma}_i$ is as defined in Theorem 4. Let $\mathcal{K} = \{\pi \mid \nabla V^{\pi^i}(x^0) \check{\Gamma}_i \left(\Upsilon(\nabla^2 V^{\pi}(x^0))^{-1} \nabla V^{\pi^i}(x^0) \right) = 0, \forall i = 1, \ldots, \mathcal{LM}\}$. Then, we have

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

Proof. See Section **??**. □

5 Gradient-free model-based policy-search algorithm (GF-CPT-MPS

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 π^* for the problem $\min_{\pi \in \Pi} V^{\pi}(x^0)$.

5.1 Basic algorithm

To illustrate the main idea in the algorithm, assume we know the form of $V^{\pi}(x^0)$. Then, the idea is to generate a sequence of reference distributions $g_k(\pi)$ on the policy space Π, such that it eventually concentrates on the global optimum π^*. One simple way, suggested in Chapter 4 of **?** is

$$g_k(\pi) = \frac{\mathcal{H}(V^{\pi}(x^0)) g_{k-1}(\pi)}{\int_{\Pi} \mathcal{H}(V^{\pi'}(x^0)) g_{k-1}(\pi') \nu(d\pi')}, \ \forall \pi \in \Pi, \tag{17}$$

where ν is the Lebesgue/counting measure on Π 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 π^*.

Next, consider a setting where one can obtain the CPT-value $V^{\pi}(x^0)$ (without any noise) for any policy π. In this case, we consider a family of parameterized distributions, say $\{f(\cdot, \eta), \eta \in \mathcal{C}\}$ and incrementally update the distribution parameter η such that it minimizes the following KL divergence:

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

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

$$\eta_{n+1} \in \arg\min_{\eta \in \mathcal{C}} E_{\eta_n} \left[\frac{[\mathcal{H}(V^{\mathcal{R}(\Pi)}(x^0))]^n}{f(\mathcal{R}(\Pi), \eta_n)} \ln f(\mathcal{R}(\Pi), \eta) \right], \tag{18}$$

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

Finally, we get to our setting where we only obtain biased estimate of the CPT-value $V^{\pi}(x^0)$ for any policy π. Recall that the bias is due to a finite sample run followed by estimation scheme (4). 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):

[1] Here $\widehat{V}_n^{\pi_n^{(i)}}(x^0)$ 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. $f(\pi,\eta_0)>0\;\forall\,\pi\in\Pi$, 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 $\Lambda_n=\{\pi_n^1,\ldots,\pi^{N_n}\}$.

CPT-value Estimation

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

Simulate m_n episodes of the SSP using policy π_n^i

Obtain CPT-value estimate $\widehat{V}_n^{\pi_n^i}(x^0)$ using (4)

end for

Elite Sampling

Order the CPT-value estimates[1] $\{\widehat{V}_n^{\pi_n^{(1)}}(x^0),\ldots,\widehat{V}_n^{\pi_n^{(N_n)}}(x^0)\}$.

Compute the $(1-\rho_n)$-quantile from the above samples as follows:

$$\widetilde{\chi}_n(\rho_n,N_n)=\widehat{V}_n^{\pi_n^{\lceil(1-\rho_k)N_k\rceil}}(x^0). \tag{19}$$

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 $\pi_n^*=\pi_{1-\rho_n}$, where $\pi_{1-\rho_n}$ is the policy that corresponds to the $(1-\rho_n)$-quantile in (19).

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 $\pi_n^*=\pi_{1-\bar{\rho}}$

else

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

end if

end if

Sampling Distribution Update

Parameter update[2]:

$$\eta_{n+1}\in\arg\min_{\eta\in\mathcal{C}}\frac{1}{N_n}\sum_{i=1}^{N_n}\frac{[\mathcal{H}(\hat{V}^{\pi_n^i}(x^0))]^n}{\widetilde{f}(\pi,\eta_n)}\widetilde{I}\big(\hat{V}^{\pi_n^i}(x^0),\bar{\chi}_n\big)\ln f(\pi,\eta).$$

end for

Return π_n

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

Step 2 (CPT-value estimation): Run m_n SSP episodes for each of the policies in $\pi_n^i, i = 1, \ldots, N_n$ and return CPT-value estimates $\hat{V}^{\pi_n^i}(x^0)$.

Step 3 (Parameter update):

$$\eta_{n+1} \in \arg\min_{\eta \in \mathcal{C}} \frac{1}{N_n} \sum_{i=1}^{N_n} \frac{[\mathcal{H}(\hat{V}^{\pi_n^i}(x^0))]^n}{f(\pi_n^i, \eta_n)} \ln f(\pi_n^i, \eta). \tag{20}$$

A few remarks are in order.

Remark 1. ***(Choice of sampling distribution)*** *A natural question is how to compute the KL-distance* (18) *in order to update the policy. A related question is how to choose the family of distributions $f(\cdot, \pi)$, so that the update* (18) *can be done efficiently. One choice is to employ the natural exponential family (NEF) since it ensures that the KL distance in* (18) *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.*

5.2 Convergence result

Theorem 6. *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,*

$$\lim_{n\to\infty} \mu_n = \pi^* \text{ and } \lim_{n\to\infty} \Sigma_n = 0_{d\times d} \ \ a.s. \tag{21}$$

Proof. See Section **??**. □

6 Convergence Proofs

6.1 Proofs for CPT-value estimator

6.1.1 Lipschitz continuous weights

In order to prove Proposition 1, we require the dominated convergence theorem in its generalized form, which is provided below.

Theorem 7. ***(Generalized Dominated Convergence theorem)*** *Let $\{f_n\}_{n=1}^{\infty}$ be a sequence of measurable functions on E that converge pointwise a.e. on a measurable space E to f. Suppose there is a sequence $\{g_n\}$ of integrable functions on E that converge pointwise a.e. on E to g such that $|f_n| \leq g_n$ for all $n \in \mathbb{N}$. If $\lim_{n\to\infty} \int_E g_n = \int_E g$, then $\lim_{n\to\infty} \int_E f_n = \int_E f$.*

Proof. This is a standard result that can be found in any textbook on measure theory. For instance, see Theorem 2.3.11 in **?**. □

In addition to the notational convention in section 2, we shall henceforth denote $u^+(X)$ and $u^-(X)$ by U^+ and U^-, respectively, for notational convenience. in the whole section 6.1.

Proof of Proposition 1: Asymptotic convergence

Proof. Recall that the CPT-value for any r.v. X is defined as

$$V(X) = \int_0^{+\infty} w^+(P(U^+ > x)dx - \int_0^{+\infty} w^-(P(U^- > x)dx.$$

Also, recall that we estimate $V(X)$ using the empirical distribution as follows:

$$\widehat{V}_n(X) = \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx - \int_0^{+\infty} w^-(1 - \hat{F}_n^-(x))dx, \tag{22}$$

where

$$\hat{F}_n^+(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(U^+ \leq x)}, \quad \text{and} \quad \hat{F}_n^-(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{(U^- \leq x)}.$$

We first prove the claim for the first integral in (22), i.e., we show

$$\int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx \to \int_0^{+\infty} w^+(P(U^+ > x)dx. \tag{23}$$

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

Notice that the empirical distribution function $\hat{F}_n^+(x)$ generates a Stieltjes measure which takes mass $1/n$ on each of the sample points U_i^+.

We have

$$\int_0^{+\infty} (P(U^+ > x))dx = E(U^+)$$

and

$$\int_0^{+\infty} (1 - \hat{F}_n^+(x))dx = \int_0^{+\infty} \int_x^{\infty} d\hat{F}_n(t)dx. \tag{24}$$

Since $\hat{F}_n^+(x)$ has bounded support on $\mathbb{R}$ $\forall n$, the integral in (24) is finite. Applying Fubini's theorem to the RHS of (24), we obtain

$$\int_0^{+\infty} \int_x^{\infty} d\hat{F}_n(t)dx = \int_0^{+\infty} \int_0^{t} dx d\hat{F}_n(t) = \int_0^{+\infty} t d\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} U_{[i]}^+, \tag{25}$$

where $U_{[i]}^+, i = 1, \ldots, n$ denote the order statistics, i.e., $U_{[1]}^+ \leq \ldots \leq U_{[n]}^+$.

Now, notice that

$$\frac{1}{n}\sum_{i=1}^{n} U_{[i]}^+ = \frac{1}{n}\sum_{i=1}^{n} U_i^+ \xrightarrow{a.s} E(U^+),$$

From the foregoing,

$$\lim_{n\to\infty} \int_0^{+\infty} L \cdot (1 - \hat{F}_n(x))dx \xrightarrow{a.s} \int_0^{+\infty} L \cdot (P(U^+ > x))dx.$$

Hence, we have

$$\int_0^{\infty} w^{(+)}(1 - \widehat{F_n}(x))dx \xrightarrow{a.s.} \int_0^{\infty} w^{(+)}(P(U^+) > x)dx.$$

The claim in (23) now follows by invoking the generalized dominated convergence theorem by setting $f_n = w^+(1 - \hat{F}_n^+(x))$ and $g_n = L \cdot (1 - \hat{F}_n(x))$, and noticing that $L \cdot (1 - \hat{F}_n(x)) \xrightarrow{a.s.} L(P(U^+ > x))$ uniformly $\forall x$. The latter fact is implied by the Glivenko-Cantelli theorem (cf. Chapter 2 of **?**).

Following similar arguments, it is easy to show that

$$w^-(1 - \hat{F}_n^-(x))dx \to \int_0^{+\infty} w^-(P(U^-) > x)dx.$$

The final claim regarding convergence of $\hat{V}_n(X)$ to $V(X)$ now follows. □

Proof of Proposition 1: Asymptotic convergence rate

In order to prove the convergence rate of the policy optimization algorithms, we need a uniform bound on the distance between $V_n(X)$ and the CPT-value $V(X)$, i.e., $||\widehat{V_n}(X) - V(X)||_\infty$. For this purpose, Proposition 1 had a law of the iterated logarithm type result, which states that $||\widehat{V_n}(X) - V(X)||_\infty$ is of the order $O(n^{-1/2})$ (ignoring log-factors) for sufficiently large n. Before proving this result, we recall the law of the iterated logarithm when empirical distribution function is in one dimension (see **?** for a detailed description):

Theorem 8. ***(Law of the iterated logarithm.)***
Let $\hat{F}_n$ denote the empirical distribution and F the true distribution. Then, for sufficiently large n, we have

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2loglogn}} ||\hat{F}_n - F||_\infty \leq \frac{1}{2}, a.s.$$

Proof. (***Proof of Proposition 1: Asymptotic convergence rate***)
Let's focus on the difference $\left|\int_0^{+\infty} w^+(P(u^+((X))) > x)dx - \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx\right|$. We have

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

Using the law of iterated logarithm, we obtain

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2loglogn}} \left\|\int_0^{+\infty} w^+(1 - \hat{F}_n^+(x)) - \int_0^{+\infty} P(U^+ > x)dx\right\|_\infty \leq \frac{1}{2}LM, \text{a.s.}$$

Along similar lines, we obtain

$$\limsup_{n\to\infty} \sqrt{\frac{n}{2loglogn}} \left\|\int_0^{+\infty} w^-(1 - \hat{F}_n^-(x)) - \int_0^{+\infty} P(U^- > x)dx\right\|_\infty \leq \frac{1}{2}LM, \text{a.s.}$$

The main claim follows by summing the above two inequalities. □

For proving Proposition 2, we require the following well-known inequality that provide a finite-time bound on the distance between empirical distribution and the true distribution:

Lemma 9. ***(Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)***
Let $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{((X_i)\leq x)}$ denote the empirical distribution of a r.v. X, with $X_1, \ldots, X_n$ being sampled from the true distribution $F(X)$. The, for any n and $\epsilon > 0$, we have

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

The reader is referred to Chapter 2 of **?** for more on empirical distributions in general and DKW inequality in particular.

Proof. (***Theorem 2***)
Since U^+ is bounded above by M and w^+ is Lipschitz with constant L, we have

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

Now, plugging in the DKW inequality, we obtain

$$\begin{aligned}
&P\left(\left|\int_0^{+\infty} w^+(P(U^+) > x)dx - \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx\right| > \epsilon/2\right) \\
&\leq P\left(LM \sup_{x\in\mathbb{R}} \left|(P(U^+ < x) - \hat{F}_n^+(x)\right| > \epsilon/2\right) \leq 2e^{-n\frac{\epsilon^2}{2L^2M^2}}. \qquad (26)
\end{aligned}$$

Along similar lines, we obtain

$$P\left(\left|\int_0^{+\infty} w^-(P(U^-) > x)dx - \int_0^{+\infty} w^-(1 - \hat{F}_n^-(x))dx\right| > \epsilon/2\right) \leq 2e^{-n\frac{\epsilon^2}{2L^2M^2}}. \qquad (27)$$

Combining (26) and (**??**), we obtain

$$\begin{aligned}
P(|\widehat{V}_n(X) - V(X)| > \epsilon) \leq& P\left(\left|\int_0^{+\infty} w^+(P(U^+) > x)dx - \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x))dx\right| > \epsilon/2\right) \\
&+ P\left(\left|\int_0^{+\infty} w^-(P(U^-) > x)dx - \int_0^{+\infty} w^-(1 - \hat{F}_n^-(x))dx\right| > \epsilon/2\right) \\
&\leq 4e^{-n\frac{\epsilon^2}{2L^2M^2}}.
\end{aligned}$$

And the claim follows. □

6.1.2 Hölder continuous weights

Proposition 7. *Under (A1'), the CPT-value $V(X)$ as defined by* (1) *is finite.*

Proof. Hölder continuity of w^+ implies that

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

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

Proposition 8. *Assume (A1'). Then, we have*

$$\begin{aligned} \lim_{n\to\infty} \sum_0^{n-1} \xi^+_{\frac{i}{n}}(w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) &= \int_0^{+\infty} w^+(P(U^+ > z))dz < +\infty, \\ \lim_{n\to\infty} \sum_0^{n-1} \xi^-_{\frac{i}{n}}(w^-(\frac{n-i}{n}) - w^-(\frac{n-i-1}{n})) &= \int_0^{+\infty} w^-(P(U^- > z))dz < +\infty \end{aligned} \tag{28}$$

and also

with the variable $\xi^+_{\frac{i}{n}}$ and $\xi^-_{\frac{i}{n}}$ denote respectively the $\frac{i}{n}$th quantile of the r.v. U^+ and U^-.

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:

$$\sum_{i=0}^{n-1} w^+_{\frac{i}{n}} \cdot I_{[\xi^+_{\frac{n-i-1}{n}}, \xi_{\frac{n-i}{n}}]}(t), \tag{29}$$

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

$$\sum_{i=0}^{n-1} w^+_{\frac{i}{n}} \cdot I_{[\xi^+_{\frac{n-i-1}{n}}, \xi^+_{\frac{n-i}{n}}]}(t) < w(P(U > t)),\ \forall t \in [0, +\infty) \tag{30}$$

The integral of (4) equals to

$$\int_0^{+\infty} \sum_{i=0}^{n-1} w^+_{\frac{i}{n}} \cdot I_{[\xi^+_{\frac{n-i-1}{n}}, \xi^+_{\frac{n-i}{n}}]}(t) \tag{31}$$

$$= \sum_{i=0}^{n-1} w^+_{\frac{i}{n}}(t) \cdot (\xi^+_{\frac{n-i}{n}} - \xi^+_{\frac{n-i-1}{n}}) \tag{32}$$

$$= \sum_{i=0}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+_{\frac{n-i}{n}} - w^+_{\frac{n-i-1}{n}}) \tag{33}$$

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 dominated convergence theorem. The second part of equation (28) can be justified in a similar fashion. □

Proof of Proposition 3

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

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

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

$$P(\left|\sum_{i=1}^{n-1} U^+_{[i]} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon), \quad (34)$$

with ϵ being any given constant $\epsilon > 0$. Observe the fact 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 are turning to 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 $P(U^+_{[i]} - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}})$. We denote $W_t = I_{(U^+_t > \xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})}$,t as a dummy value. According to the non-decreasing property of the probability distribution function, we know that

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

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

$$P(\sum_{i=1}^{n} W_t - n \cdot [1 - F(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}})] > n \cdot [F(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}) - \frac{i}{n}]) < e^{-2n\cdot\delta^{'}_t}, \tag{35}$$

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

$$P(U^+_{[i]} - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}) < e^{-2n\cdot L\frac{\epsilon}{n^{(1-\alpha)}}} = e^{-2n^{\alpha}\cdot L} \tag{36}$$

Through the similar fashion, one can show that

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

Subject to the assumption that $F(x)$ is Lipschitz continuous, one can state that

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

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

$$P(\left|\sum_{i=1}^{n-1} U^+_{[i]} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})) - \sum_{i=1}^{n-1} \xi^+_{\frac{i}{n}} \cdot (w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n}))\right| > \epsilon) \leq 2n \cdot e^{-2n^{\alpha}\cdot L}. \tag{37}$$

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

By applying 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 constitute the proof of theorem's statement, that

$$\lim_{n\to+\infty} \sum_{i=1}^{n-1} U^+_{[i]}(w^+(\frac{n-i+1}{n}) - w^+(\frac{n-i}{n})) \xrightarrow{n\to\infty} \int_0^{+\infty} w^+(P(U > t))dt, \text{w.p1} \tag{38}$$

□

Proof. Follows from (36) in the proof of Proposition 3. □

Proof of Proposition 4

Proof. We would prove the w^+ part, and the w^- part will be proved through 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

$$\begin{aligned}
&P\left(\left|\int_0^{+\infty} w^+(P(U^+)>t)dt - \int_0^{+\infty} w^+(1-\hat{F}_n^+(t))dt\right|>\epsilon\right) \\
&\leq P\left(LM\sup_{t\in\mathbb{R}}\left|(P(U^+<t)-\hat{F}_n^+(t)\right|^\alpha>\epsilon\right)\leq e^{-n\frac{\epsilon^{(2/\alpha)}}{2L^2M^2}}. \qquad (39)
\end{aligned}$$

□

6.2 Proofs for PG-CPT-SPSA

To prove the main result in Theorem 4, 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 noise is zero-mean. Following this lemma, we complete the proof of Theorem 4 by invoking the well-known Kushner-Clark lemma **?**.

Lemma 10. *Let $\mathcal{F}_n=\sigma(\pi_m, m\leq n)$, $n\geq 1$. Then, for any $i=1,\ldots,\mathcal{LM}$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\widehat{V}_n^{\pi_n+\delta_n\Delta_n}(x^0)-\widehat{V}_n^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_i V^{\pi_n}(x^0)\right|\to 0 \text{ as } n\to\infty. \qquad (40)$$

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

We claim

$$\mathbb{E}\left[\frac{\widehat{V}_n^{\pi_n+\delta_n\Delta_n}(x^0)-\widehat{V}_n^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\mid\mathcal{F}_n\right]= \mathbb{E}\left[\frac{V_n^{\pi_n+\delta_n\Delta_n}(x^0)-V_n^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\mid\mathcal{F}_n\right]+\mathbb{E}\left[\eta_n\mid\mathcal{F}_n\right], \qquad (41)$$

where $\eta_n=\left(\frac{\epsilon^{\pi_n+\delta_n\Delta}-\epsilon^{\pi_n-\delta_n\Delta}}{2\delta_n\Delta_n^i}\right)$ is the bias arising out of the empirical distribution based CPT-value estimation scheme. From Proposition 1, we see that, $\epsilon^\pi=LM\sqrt{\frac{2\log\log m_n}{m_n}}$, provided m_n is sufficiently

large. Thus, $\eta_n = O\left(\sqrt{\frac{2\log\log m_n}{m_n}}\frac{1}{\delta_n}\right)$ and since $\frac{1}{\sqrt{m_n}\delta_n} \to 0$ by assumption (A3) in the main paper, we have that η_n goes to zero asymptotically. In other words,

$$\mathbb{E}\left[\frac{\widehat{V}_n^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}_n^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] \xrightarrow{n\to\infty} \mathbb{E}\left[\frac{V^{\pi_n+\delta_n\Delta_n}(x^0) - V^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right]. \quad (42)$$

We now analyse the RHS of the above. By using suitable Taylor's expansions,

$$V^{\pi_n+\delta_n\Delta_n}(x^0) = V^{\pi_n}(x^0) + \delta_n\Delta_n^\intercal\nabla V^{\pi_n}(x^0) + \frac{\delta^2}{2}\Delta_n^\intercal\nabla^2 V^{\pi_n}(x^0)\Delta_n + O(\delta_n^3),$$
$$V^{\pi_n-\delta_n\Delta_n}(x^0) = V^{\pi_n}(x^0) - \delta_n\Delta_n^\intercal\nabla V^{\pi_n}(x^0) + \frac{\delta^2}{2}\Delta_n^\intercal\nabla^2 V^{\pi_n}(x^0)\Delta_n + O(\delta_n^3).$$

From the above, it is easy to see that

$$\frac{V^{\theta_n+\delta_n\Delta_n}(x^0) - V^{\theta_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i} - \nabla_i V^{\theta_n}(x^0) = \underbrace{\sum_{j=1,j\neq i}^{N}\frac{\Delta_n^j}{\Delta_n^i}\nabla_j V^{\theta_n}(x^0)}_{(I)} + O(\delta_n^2).$$

Taking conditional expectation on both sides, we obtain

$$\begin{aligned}\mathbb{E}\left[\frac{V^{\theta_n+\delta_n\Delta_n}(x^0) - V^{\theta_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right] =& \nabla_i V^{\theta_n}(x^0) + \mathbb{E}\left[\sum_{j=1,j\neq i}^{N}\frac{\Delta_n^j}{\Delta_n^i}\right]\nabla_j V^{\theta_n}(x^0) + O(\delta_n^2)\\ =& \nabla_i V^{\theta_n}(x^0) + O(\delta_n^2). \quad (43)\end{aligned}$$

The first equality above follows from the fact that Δ_n is distributed according to a $\mathcal{LM}$-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 $i, j = 1, \ldots, \mathcal{LM}$, $j \neq i$.

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

Proof of Theorem 4

Proof. We first rewrite the update rule (14) as follows: For $i = 1, \ldots, \mathcal{LM}$,

$$\pi_{n+1}^i = \pi_n^i - \gamma_n(\nabla_i V^{\pi_n}(x_0) + \beta_n + \xi_n), \quad (44)$$

where

$$\begin{aligned}\beta_n &= \mathbb{E}\left(\frac{(\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0))}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right) - \nabla V^{\pi}(x^0), \text{ and}\\ \xi_n &= \left(\frac{\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\right) - \mathbb{E}\left(\frac{(\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0))}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n\right).\end{aligned}$$

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 (43) 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) $\nabla V^{\pi}(x^0)$ is a continuous $\mathbb{R}^{\mathcal{LK}}$-valued function.

(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:

$$\dot{\pi}_t^i = \check{\Gamma}_i\left(\nabla V^{\pi_t^i}(x^0)\right), \text{ for} i = 1, \ldots, \mathcal{LM}, \tag{45}$$

In the following, we verify the above assumptions for the recursion (14):

- (B1) holds by assumption in our setting.
- Lemma 10 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) in the main paper.
- 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 **?**):

$$P\left(\sup_{m\geq 0}\|W_l\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\lim_{l\to\infty}\mathbb{E}\|W_l\|^2. \tag{46}$$

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

$$P\left(\sup_{l\geq k}\left\|\sum_{n=k}^{l}\gamma_n\xi_n\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{n=k}^{\infty}\gamma_n\xi_n\right\|^2 = \frac{1}{\epsilon^2}\sum_{n=k}^{\infty}\gamma_n^2\mathbb{E}\|\eta_n\|^2. \tag{47}$$

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:

$$\begin{aligned}
\mathbb{E}\|\xi_n\|^2 \leq & \mathbb{E}\left(\frac{\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0) - \widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\right)^2 && (48)\\
\leq & \left(\left(\mathbb{E}\left(\frac{\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2} + \left(\mathbb{E}\left(\frac{\widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2}\right)^2 && (49)\\
\leq & \frac{1}{4\delta_n^2}\left[\mathbb{E}\left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}}\right)\right]^{\frac{1}{1+\alpha_1}} \\
& \times\left(\left[\mathbb{E}\left[(\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0))\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0))\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) && (50)\\
\leq & \frac{1}{4\delta_n^2}\left(\left[\mathbb{E}\left[(\widehat{V}^{\pi_n+\delta_n\Delta_n}(x^0))\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\widehat{V}^{\pi_n-\delta_n\Delta_n}(x^0))\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) && (51)\\
\leq & \frac{C}{\delta_n^2}, \text{ for some } C < \infty. && (52)
\end{aligned}$$

The inequality in (47) 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 (48) 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 (49) 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 (50) 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 (51) follows by using the fact that, for any π, the CPT-value estimate $\widehat{V}^\pi(x^0) = V^\pi(x^0) + \epsilon^\pi$. 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 $D^\pi(x^0)$ is bounded for any policy π) and finally, by (A1), the weight functions are Lipschitz - these together imply that $V^\pi(x^0)$ is bounded for any policy π. The bias ϵ^π is bounded by Proposition 1 in the main paper.

Thus, $\mathbb{E}\left\|\xi_n\right\|^2 \leq \frac{C}{\delta_n^2}$ for some $C < \infty$. Plugging this in (46), 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.

- The set $\mathcal{K} = \{\pi \mid \check{\Gamma}_i\left(\nabla V^\pi(x^0)\right) = 0, \forall i = 1, \ldots, \mathcal{LM}\}$ serves as the asymptotically stable attractor for the ODE (44).

The claim follows from Kushner-Clark lemma. □

6.3 Proofs for PN-CPT-SPSA

Before proving Theorem 5, we bound the bias in the SPSA based estimate of the Hessian in the following lemma.

Lemma 11. *For any $i, j = 1, \ldots, \mathcal{LM}$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\widehat{V}_n^{\pi_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) + \widehat{V}_n^{\pi_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - 2\widehat{V}_n^{\pi_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\middle|\mathcal{F}_n\right] - \nabla^2_{i,j}V^{\pi_n}(x^0)\right| \to 0 \text{ as } n\to\infty. \tag{53}$$

Proof. As in the proof of Lemma 10, we can ignore the bias from the CPT-value estimation scheme and conclude that,

$$\begin{aligned}&\mathbb{E}\left[\frac{\widehat{V}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) + \widehat{V}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - 2\widehat{V}_n^{\theta_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right]\\ &\xrightarrow{n\to\infty} \mathbb{E}\left[\frac{V^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) + V^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0) - 2V^{\theta_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right].\end{aligned} \tag{54}$$

Now, the RHS of the above approximates the true gradient with only an $O(\delta_n^2)$ error and this can be inferred using arguments similar to that 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 10, we obtain

$$\begin{aligned}
&\frac{V^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)+V^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)-2V^{\theta_n}(x^0)}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j}\\
=&\frac{(\Delta_n+\hat{\Delta_n})^\intercal\nabla^2V^{\theta_n}(x^0)(\Delta_n+\hat{\Delta_n})}{\triangle_i(n)\hat{\triangle}_j(n)}+O(\delta_n^2)\\
=&\sum_{l=1}^{\mathcal{LM}}\sum_{m=1}^{\mathcal{LM}}\frac{\Delta_n^l\nabla^2_{l,m}V^{\theta_n}(x^0)\Delta_n^m}{\Delta_n^i\hat{\Delta}_n^j}+2\sum_{l=1}^{\mathcal{LM}}\sum_{m=1}^{\mathcal{LM}}\frac{\Delta_n^l\nabla^2_{l,m}V^{\theta_n}(x^0)\hat{\Delta}_n^m}{\Delta_n^i\hat{\Delta}_n^j}+\sum_{l=1}^{\mathcal{LM}}\sum_{m=1}^{\mathcal{LM}}\frac{\hat{\Delta}_n^l\nabla^2_{l,m}V^{\theta_n}(x^0)\hat{\Delta}_n^m}{\Delta_n^i\hat{\Delta}_n^j}+O(\delta_n^2).
\end{aligned}$$

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

Lemma 12. *For any $i=1,\ldots,\mathcal{LM}$, we have almost surely,*

$$\left|\mathbb{E}\left[\frac{\widehat{V}_n^{\theta_n+\delta_n(\Delta_n+\hat{\Delta}_n)}(x^0)-\widehat{V}_n^{\theta_n-\delta_n(\Delta_n+\hat{\Delta}_n)}(x^0)}{2\delta_n\Delta_n^i}\middle|\mathcal{F}_n\right]-\nabla_iV^{\theta_n}(x^0)\right|\to 0 \text{ as } n\to\infty. \tag{55}$$

Proof. As in the proof of Lemma 10, we can ignore the bias from the CPT-value estimation scheme and conclude that,

$$\mathbb{E}\left[\frac{\widehat{V}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)-\widehat{V}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)}(x^0)}{2\delta_n\Delta_n^i}\mid\mathcal{F}_n\right]\xrightarrow{n\to\infty}\mathbb{E}\left[\frac{V^{\theta_n+\delta_n\Delta_n}(x^0)-V^{\theta_n-\delta_n\Delta_n}(x^0)}{2\delta_n\Delta_n^i}\mid\mathcal{F}_n\right].$$

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 and this can be done in a similar manner as the proof of Lemma 10. □

Proof of Theorem 5

Before we prove Theorem 5, we show that the Hessian recursion (16) in the main paper converges to the true Hessian, for any policy π.

Lemma 13. *For any $i,j=1,\ldots,\mathcal{LM}$, we have almost surely,*

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

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

Proof. ***(Theorem 5)*** The proof follows in a similar manner as the proof of Theorem 7.1 in **?** and we provide a sketch below for the sake of completeness.

We first rewrite the recursion (15) in the main paper as follows: For $i=1,\ldots,\mathcal{LM}$

$$\pi_{n+1}^i=\Gamma_i\left(\pi_n^i-\gamma_n\sum_{j=1}^{\mathcal{LM}}\bar{M}^{i,j}(\pi_n)\nabla_jV_n^{\pi}(x^0)+\gamma_n\zeta_n+\chi_{n+1}-\chi_n\right), \tag{56}$$

where

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

In lieu of Lemmas **??**–**??**, 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 (**??**) is a discretization of the ODE:

$$\dot{\pi}_t^i = \check{\Gamma}_i \left(\nabla V^{\pi_t^i}(x^0) \Upsilon(\nabla^2 V^{\pi_t}(x^0))^{-1} \nabla V^{\pi_t^i}(x^0) \right).$$

Further, $\mathcal{K} = \{\pi \mid \nabla V^{\pi^i}(x^0) \check{\Gamma}_i \left(\Upsilon(\nabla^2 V^{\pi}(x^0))^{-1} \nabla V^{\pi^i}(x^0) \right) = 0, \forall i = 1, \ldots, \mathcal{LM}\}$ serves as an asymptotically stable attractor set and the claim follows by applying Kusher-Clark lemma to (**??**). □

6.4 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 14. *The sequence of random variables* $\{\pi_n^*, n = 0, 1, \ldots\}$ *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 $\mathcal{B}_n := \{V^{\pi_n^*}(x^0) - V^{\pi_{n-1}^*}(x^0) \leq \frac{\varepsilon}{2}\}$. Whenever $\mathcal{A}_n$ holds, we have $\widehat{V}_n^{\pi_n^*}(x^0) - \widehat{V}_n^{\pi_{n-1}^*}(x^0) \geq \varepsilon$ and hence, we obtain

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

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 is implied that if $\mathcal{A}_n$ happens infinitely often, then $\mathcal{B}_n^c$ will also happen infinitely often. Hence,

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

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 $\pi_n^* = \pi_{n-1}^*$. From the last equality above, we conclude that it is a contradiction because, $V^{\pi}(x^0) > V^{\pi^*}(x^0)$ for any π (since π^* is the global minimum). The main claim now follows since $\mathcal{A}_n$ can happen only a finite number of times. □

Proof of Theorem 6

Proof. Once we have established Lemma **??**, 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$_2$ w.r.t. generating the candidate solution using a parameterized family $f(\cdot, \eta)$ and updating the distribution parameter η. The difference, as mentioned earlier, is the manner in which the samples are generated and the objective (CPT-value) function is estimated. The aforementioned lemma established that the elite sampling and thresholding achieve the same effect as that in MRAS$_2$ and hence the rest of the proof follows from **?**. □

7 Simulation Experiments

7.1 Simulation Setup

We consider a SSP version of an example[6] for buying a house at the optimal price. Suppose the house is priced at x_k any instant k and at the next instant, the price either goes down to $(x_k \times C_{down})$ w.p. p_{down} or goes up to $(x_k \times C_{up})$ w.p. $1 - p_{down}$. The actions are to either wait (denoted w), which results in a holding cost h or to buy (denoted b) at the current price. The horizon is capped at T, with a terminal cost x_T. The goal is to minimize the total cost defined as $D^{\pi}(x^0) = \sum_{k=0}^{\tau} \left(I_{\{a_k=b\}} x_k + I_{\{a_k=w\}} h \right) + I_{\{\tau=T\}} x_T$, where $\tau = \{k | \pi(x_k) = 1\} \wedge T$. We set $T = 20, h = 0.1, C_{up} = 2, C_{down} = 0.5$, and $x_0 = 1$.

Implementation: On this example, we implement the first-order PG-CPT-SPSA and the second-order PN-CPT-SPSA 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 PG-NoCPT-SPSA, 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 PG-CPT-SPSA and PG-NoCPT-SPSA, we set $\delta_n = 1.9/n^{0.101}$ and $\gamma_n = 1/n$, while for PN-CPT-SPSA, we set $\delta_n = 3.8/n^{0.166}$ and $\gamma_n = 1/n^{0.6}$. For all algorithms, we set each entry of the initial policy π_0 to 0.1. For CPT-value estimation, we simulate 1000 SSP episodes, with the SSP horizon T set to 20. All algorithms are run with a budget of 1000 samples, which implies 500 iterations of PG-CPT-SPSA and 333 iterations of PN-CPT-SPSA. The results presented are averages over 500 independent simulations. For PG-CPT-SPSA/PN-CPT-SPSA, 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.

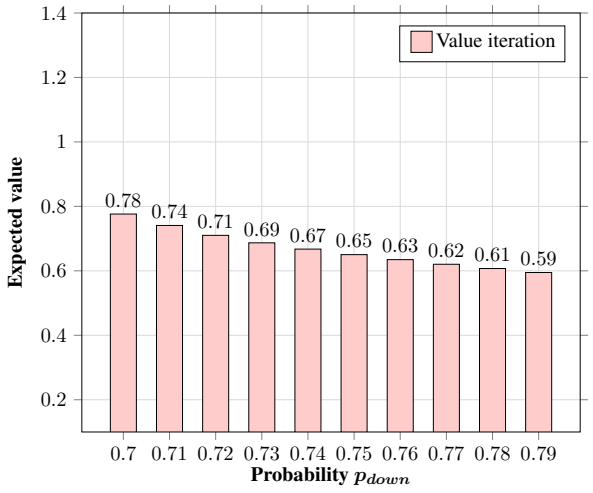
7.2 Results

Figures **??–??** present the value function computed using value iteration and PG-NoCPT-SPSA, while Figures **??–??** present the CPT-value $V^{\pi_{end}}(x^0)$ for PG-CPT-SPSA and PN-CPT-SPSA, respectively. The performance plots are for various values of p_{down}, the probability of house price going down. From Figure **??**, we notice that the variations in expected total cost is larger in comparison to that in CPT-value. Figure **??** implies that a similar observation about variation of expected value holds true for PG-NoCPT-SPSA 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/PN-CPT-SPSA 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 PN-CPT-SPSA gives marginally better results than its first-order counterpart PG-CPT-SPSA. Finally, what isn't shown is that the CPT-value obtained for PG/PN-CPT-SPSA is much lower than that obtained for PG-NoCPT-SPSA, thus making apparent the need for specialized algorithms that incorporate CPT-based criteria.

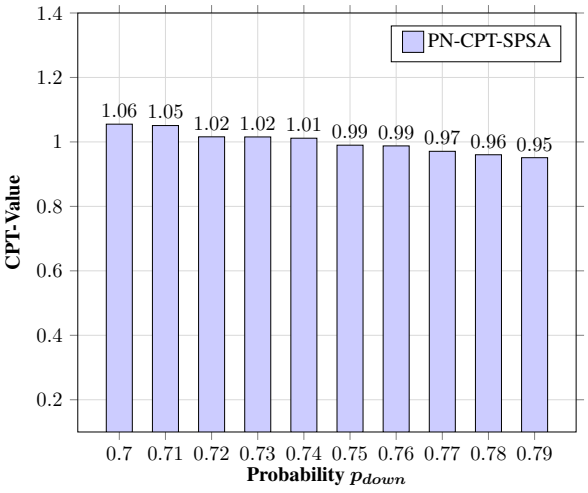
8 Conclusions and Future Work

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

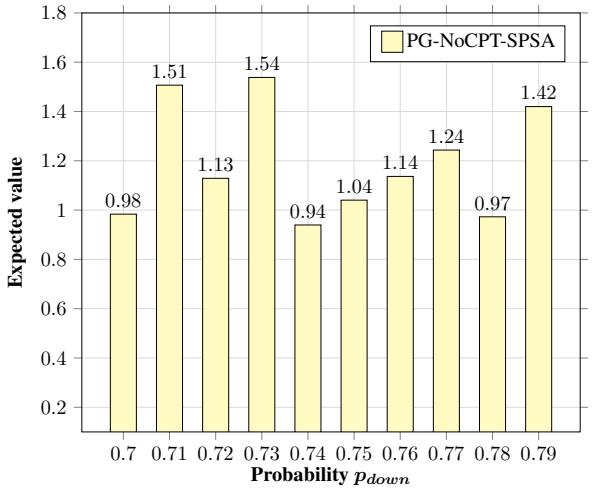
[6] A similar example has been considered in **?**.



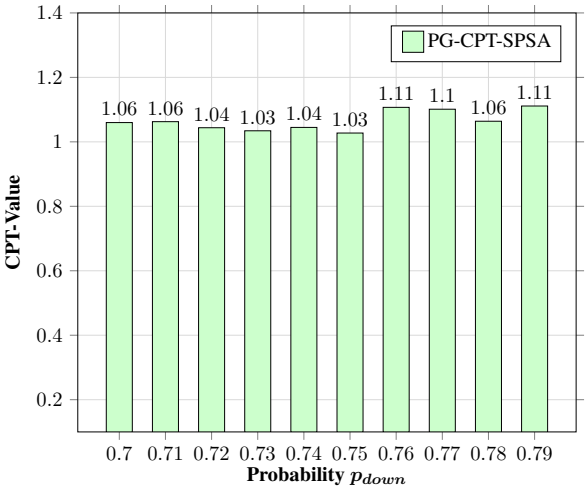
(a) Value iteration



(b) Second-order SPSA for CPT-value



(c) SPSA for regular value function



(d) First-order SPSA for CPT-value

Figure 4: Performance of policy gradient algorithms with/without CPT for different down probabilities of the SSP

criteria into an RL framework. However, both estimation and control of CPT-based value is challenging. Using temporal-difference learning type algorithms for estimation was ruled out for CPT-value since the underlying probabilities get (non-linearly) distorted by a weight function. Using empirical distributions, we proposed an estimation scheme that converges at the optimal rate. Next, for the problem of control, since CPT-value does not conform to any Bellman equation, we employed SPSA - a popular simulation optimization scheme and designed both first and second-order algorithms for optimizing the CPT-value function. We provided theoretical convergence guarantees for all the proposed algorithms. We illustrated the usefulness of CPT-based criteria in a numerical example.