
Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control

Prashanth L.A.

Institute for Systems Research, University of Maryland

PRASHLA@ISR.UMD.EDU

Cheng Jie

Department of Mathematics, University of Maryland

CJIE@MATH.UMD.EDU

Michael Fu

Robert H. Smith School of Business & Institute for Systems Research, University of Maryland

MFU@ISR.UMD.EDU

Steve Marcus

Department of Electrical and Computer Engineering & Institute for Systems Research, University of Maryland

MARCUS@UMD.EDU

Csaba Szepesvári

Department of Computing Science, University of Alberta

SZEPESVA@CS.UALBERTA.CA

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 a CPT-value optimization procedure that is based on the well-known simulation optimization idea of simultaneous perturbation stochastic approximation (SPSA). We provide theoretical convergence guarantees for all the proposed algorithms and also illustrate the usefulness of CPT-based criteria in a traffic signal control application.

1. Introduction

Since the beginning of its history, mankind has been deeply immersed in designing and improving systems to serve human needs. Policy makers are busy with designing systems that serve the education, transportation, economic, health and other needs of the public, while private sector enterprises work hard at creating and optimizing systems to better serve specialized needs of their customers. While it has been long recognized that understanding human behavior is a prerequisite to best serving human needs (Simon, 1959), it is only recently that this approach is gaining a wider recognition.¹

In this paper we consider *human-centered reinforcement learning problems* where the reinforcement learning agent controls a system to produce long term outcomes (“return”) that are maximally aligned with the preferences of one or possibly multiple humans, an arrangement shown in Figure 1. As a running example, consider traffic optimization where the goal is to maximize travelers’ satisfaction, a challenging problem in big cities. In this example, the outcomes (“return”) are travel times, or delays. To capture human preferences, the outcomes are mapped to a single numerical quantity. While preferences of rational agents facing decisions with stochastic outcomes can be modeled using expected utilities, i.e., the expectation of a nonlinear transformation, such as the exponential function, of the re-

¹ As evidence for this wider recognition in the public sector, we can mention a recent executive order of the White House calling for the use of behavioral science in public policy making, or the establishment of the “Committee on Traveler Behavior and Values” in the Transportation Research Board in the US.

wards or costs (Von Neumann & Morgenstern, 1944; Fishburn, 1970), humans are subject to various emotional and cognitive biases, and, as the psychology literature points out, human preferences are inconsistent with expected utilities regardless of what nonlinearities are used (Allais, 1953; Ellsberg, 1961; Kahneman & Tversky, 1979). An approach that gained strong support amongst psychologists, behavioral scientists and economists (e.g., Starmer, 2000; Quiggin, 2012) is based on (Kahneman & Tversky, 1979)’s celebrated *prospect theory* (PT), the theory that we will base our models of human preferences on in this work. More precisely, we will use *cumulative prospect theory* (CPT), a later, refined variant of prospect theory due to Tversky & Kahneman (1992), which superseded prospect theory (e.g., Barberis, 2013). CPT generalizes expected utility theory in that in addition to having a utility function transforming the outcomes, another function is introduced which distorts the probabilities in the cumulative distribution function. As compared to prospect theory, CPT is monotone with respect to stochastic dominance, a property that is thought to be useful and (mostly) consistent with human preferences.

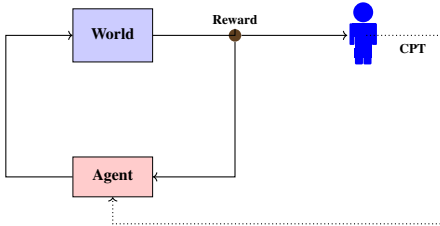


Figure 1. Operational flow of a human-based decision making system

Our contributions: To our best knowledge, we are the first to investigate (and define) human-centered RL, and, in particular, this is the first work to combine CPT with RL. Although on the surface the combination may seem straightforward, in fact there are many research challenges that arise from trying to apply a CPT objective in the RL framework, as we will soon see. We outline these challenges as well as our approach to addressing them below.

The first challenge stems from the fact that the CPT-value assigned to a random variable is defined through a non-linear transformation of the cumulative distribution functions associated with the random variable (cf. Section 2 for the definition). Hence, even the problem of estimating the CPT-value given a random sample requires quite an effort. In this paper, we consider a natural quantile-based estimator and analyze its behavior. Under certain technical assumptions, we prove consistency and give sample complexity bounds, the latter based on the Dvoretzky-Kiefer-Wolfowitz (DKW) theorem. As an example, we show that the sample complexity for estimating the CPT-value for

Lipschitz probability distortion (so-called “weight”) functions is $O\left(\frac{1}{\epsilon^2}\right)$, which coincides with the canonical rate for Monte Carlo-type schemes and is thus unimprovable. Since weight-functions that fit well to human preferences are only Hölder continuous, we also consider this case and find that (unsurprisingly) the sample complexity jumps to $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ where $\alpha \in (0, 1]$ is the weight function’s Hölder exponent.

Our results on estimating CPT-values form the basis of the algorithms that we propose to maximize CPT-values based on interacting either with a real environment, or a simulator. We set up this problem as an instance of policy search: We consider smoothly parameterized policies whose parameters are tuned via stochastic gradient ascent. For estimating gradients, we use two-point randomized gradient estimators, borrowed from simultaneous perturbation stochastic approximation (SPSA), a widely used algorithm in *simulation optimization* (Fu, 2015). Here a new challenge arises, which is that we can only feed the two-point randomized gradient estimator with *biased* estimates of the CPT-value. To guarantee convergence, we propose a particular way of controlling the arising bias-variance tradeoff.

To put things in context, risk-sensitive reinforcement learning problems are generally hard to solve. For a discounted MDP, Sobel (1982) showed that there exists a Bellman equation for the variance of the return, but the underlying Bellman operator is not necessarily monotone, thus ruling out policy iteration as a solution approach for variance-constrained MDPs. Further, even if the transition dynamics are known, Mannor & Tsitsiklis (2013) show that finding a globally mean-variance optimal policy in a discounted MDP is NP-hard. For average reward MDPs, Filar et al. (1989) motivate a different notion of variance and then provide NP-hardness results for finding a globally variance-optimal policy. Solving Conditional Value at Risk (CVaR) constrained MDPs is equally complicated (cf. Borkar & Jain 2010; Prashanth 2014; Tamar et al. 2014). Finally, we point out that the CPT-value is a generalization of all the risk measures above in the sense that one can recover these particular risk measures such as VaR and CVaR by appropriate choices of the distortions used in the definition of the CPT value.

The work closest to ours is by Lin (2013), who proposes a CPT-measure for an abstract MDP setting. We differ from this work in several ways: (i) We do not assume a nested structure for the CPT-value and this implies the lack of a Bellman equation for our CPT measure; (ii) we do not assume model information, i.e., we operate in a model-free RL setting. Moreover, we develop both estimation and control algorithms with convergence guarantees for the CPT-value function.

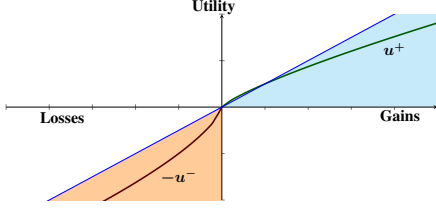


Figure 2. An example of a utility function. A reference point on the x axis serves as the point of separating gains and losses. For losses, the disutility $-u^-$ is typically convex, for gains, the utility u^+ is typically concave; they are always non-decreasing and both of them take on the value of zero at the reference point.

2. CPT-value

For a real-valued random variable X , we introduce a “CPT-functional” that replaces the traditional expectation operator. The functional, denoted by $\mathbb{C}$, indexed by $u = (u^+, u^-)$, $w = (w^+, w^-)$, where $u^+, u^- : \mathbb{R} \rightarrow \mathbb{R}_+$ and $w^+, w^- : [0, 1] \rightarrow [0, 1]$ are continuous with $u^+(x) = 0$ when $x \leq 0$ and $u^-(x) = 0$ when $x \geq 0$ (see assumptions (A1)-(A2) in Section 3 for precise requirements on u and w), is defined as

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

For notational convenience, when u, w are fixed, we drop the dependence on them and use $\mathbb{C}(X)$ to denote the CPT-value of X . Note that when w^+, w^- and $u^+ (-u^-)$, when restricted to the positive (respectively, negative) half line, are the identity functions, and we let $(a)^+ = \max(a, 0)$, $(a)^- = \max(-a, 0)$, $\mathbb{C}(X) = \int_0^{+\infty} \mathbb{P}(X > z) dz - \int_0^{+\infty} \mathbb{P}(-X > z) dz = \mathbb{E}[(X)^+] - \mathbb{E}[(X)^-]$, showing the connection to expectations.

In the definition, u^+, u^- are utility functions corresponding to gains ($X \geq 0$) and losses ($X \leq 0$), respectively, where zero is chosen as a somewhat arbitrary “reference point” to separate gains and losses. Handling losses and gains separately is a salient feature of CPT, and this addresses the tendency of humans to play safe with gains and take risks with losses. To illustrate this tendency, consider a scenario where one can either earn \$500 with probability (w.p.) 1 or earn \$1000 w.p. 0.5 and nothing otherwise. The human tendency is to choose the former option of a certain gain. If we flip the situation, i.e., a certain loss of \$500 or a loss of \$1000 w.p. 0.5, then humans choose the latter option. This distinction of playing safe with gains and taking risks with losses is captured by a concave gain-utility u^+ and a convex disutility $-u^-$, cf. Fig. 2.

The functions w^+, w^- , called the weight functions, capture

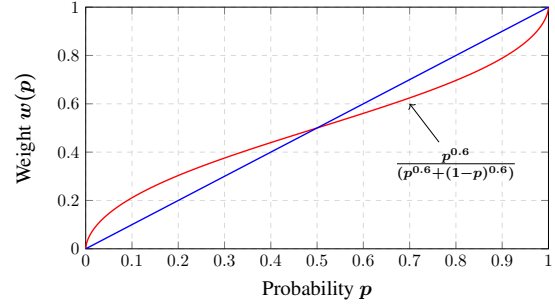


Figure 3. An example of a weight function. A typical CPT weight function inflates small, and deflates large probabilities, capturing the tendency of humans doing the same when faces with decisions of uncertain outcomes.

the idea that humans deflate high-probabilities and inflate low-probabilities. For example, humans usually choose a stock that gives a large reward, e.g., one million dollars w.p. $1/10^6$ over one that gives \$1 w.p. 1 and the reverse when signs are flipped. Thus the value seen by a human subject is non-linear in the underlying probabilities – an observation backed by strong empirical evidence (Tversky & Kahneman, 1992; Barberis, 2013). As illustrated with $w = w^+ = w^-$ in Fig 3, the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. Tversky & Kahneman (1992) recommend $w(p) = \frac{p^\eta}{(p^\eta + (1-p)^\eta)^{1/\eta}}$, while Prelec (1998) recommends $w(p) = \exp(-(-\ln p)^\eta)$, with $0 < \eta < 1$. In both cases, the weight function has an inverted-s shape.

Remark 1. (RL applications) For any RL problem setting, one can define the return for a given policy and then apply a CPT-functional on the return. For instance, with a fixed policy, the random variable (r.v.) X could be the total reward in a stochastic shortest path problem or the infinite horizon cumulative reward in a discounted MDP or the long-run average reward in an MDP.

Remark 2. (Generalization) As noted earlier, the CPT-value is a generalization of mathematical expectation. It is also possible to get (1) to coincide with risk measures (e.g. VaR and CVaR) by appropriate choice of weight functions.

3. CPT-value estimation

Before diving into the details of CPT-value estimation, let us discuss the conditions necessary for the CPT-value to be well-defined. Observe that the first integral in (1), i.e., $\int_0^{+\infty} w^+ (\mathbb{P}(u^+(X) > z)) dz$ may diverge even if the first moment of random variable $u^+(X)$ is finite. For example, suppose U has the tail distribution function $\mathbb{P}(U > z) = \frac{1}{z^2}$, $z \in [1, +\infty)$, and $w^+(z)$ takes the form $w(z) = z^{\frac{1}{3}}$. Then, the first integral in (1), i.e., $\int_1^{+\infty} z^{-\frac{2}{3}} dz$ does not

even exist. A similar argument applies to the second integral in (1) as well.

To overcome the above integrability issues, we impose additional 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.

3.1. Estimation scheme for Hölder continuous weights

Recall the Hölder continuity property first in definition 1:

Definition 1. (Hölder continuity) A function $f \in C([a, b])$ is said to satisfy a Hölder condition of order $\alpha \in (0, 1]$ (or to be Hölder continuous of order α) if there exists $H > 0$, s.t.

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

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

Assumption (A1). The weight functions w^+, w^- are Hölder continuous with common order α . Further, there exists $\gamma \leq \alpha$ such that (s.t.) $\int_0^{+\infty} P^\gamma(u^+(X) > z) dz < +\infty$ 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 5 in (Prashanth et al., 2015) for a formal proof.

Approximating CPT-value using quantiles: Let ξ_α^+ denote the α th quantile of the r.v. $u^+(X)$. Then, it can be seen that (see Proposition 6 in (Prashanth et al., 2015))

$$\begin{aligned} & \lim_{n \rightarrow \infty} \sum_{i=1}^n \xi_{\frac{i}{n}}^+ \left(w^+ \left(\frac{n+1-i}{n} \right) - w^+ \left(\frac{n-i}{n} \right) \right) \\ &= \int_0^{+\infty} w^+ (\mathbb{P}(u^+(X) > z)) dz. \end{aligned} \quad (2)$$

A similar claim holds with $u^-(X)$, ξ_α^- , w^- in place of $u^+(X)$, ξ_α^+ , w^+ , respectively. Here ξ_α^- denotes the α th quantile of $u^-(X)$.

However, we do not know the distribution of $u^+(X)$ or $u^-(X)$ and hence, we next present a procedure that uses order statistics for estimating quantiles and this in turn assists estimation of the CPT-value along the lines of (2). The estimation scheme is presented in Algorithm 1.

Algorithm 1 CPT-value estimation for Hölder continuous weights

- 1: Simulate n i.i.d. samples from the distribution of X .
- 2: Order the samples and label them as follows: $X_{[1]}, X_{[2]}, \dots, X_{[n]}$. Note that $u^+(X_{[1]}), \dots, u^+(X_{[n]})$ are also in ascending order.

3: Let

$$\bar{\mathbb{C}}_n^+ := \sum_{i=1}^n u^+(X_{[i]}) \left(w^+ \left(\frac{n+1-i}{n} \right) - w^+ \left(\frac{n-i}{n} \right) \right).$$

- 4: Apply u^- on the sequence $\{X_{[1]}, X_{[2]}, \dots, X_{[n]}\}$, notice that $u^-(X_{[i]})$ is in descending order since u^- is a decreasing function.

5: Let

$$\bar{\mathbb{C}}_n^- := \sum_{i=1}^n u^-(X_{[i]}) \left(w^- \left(\frac{i}{n} \right) - w^- \left(\frac{i-1}{n} \right) \right).$$

- 6: Return $\bar{\mathbb{C}}_n = \bar{\mathbb{C}}_n^+ - \bar{\mathbb{C}}_n^-$.
-

MAIN RESULTS²

Proposition 1. (Asymptotic consistency) Assume (A1) and also that $F^+(\cdot)$, $F^-(\cdot)$ - the distribution functions of $u^+(X)$, and $u^-(X)$ are Lipschitz continuous on the respective intervals of $(0, +\infty)$, and $(-\infty, 0)$. Then, we have that

$$\bar{\mathbb{C}}_n \rightarrow \mathbb{C}(X) \text{ a.s. as } n \rightarrow \infty \quad (3)$$

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

Under (A2') stated below, our next result shows that $O(\frac{1}{\epsilon^{2/\alpha}})$ number of samples are sufficient to get a high-probability estimate of the CPT-value that is ϵ -accurate:

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

Assumption (A2'). In addition to (A2), the utilities $u^+(X)$ and $u^-(X)$ are bounded above by $M < \infty$ w.p. 1.

Proposition 2. (Sample complexity.) Assume (A1) and (A2'). Then, $\forall \epsilon > 0, \delta > 0$, we have

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

3.1.1. RESULTS FOR LIPSCHITZ CONTINUOUS WEIGHTS

In the previous section, it was shown that Hölder continuous weights incur a sample complexity of order $O(\frac{1}{\epsilon^{2/\alpha}})$

²All proofs can be found in the extended version of the paper (Prashanth et al., 2015)

and this is higher than the canonical Monte Carlo rate of $O(\frac{1}{\epsilon^2})$. In this section, we establish that one can achieve the canonical Monte Carlo rate if we consider Lipschitz continuous weights, i.e., the following assumption in place of (A1):

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

Setting $\alpha = 1$, one can make special cases of the claims regarding asymptotic convergence and sample complexity of Proposition 1–2. However, these results are under a restrictive Lipschitz assumption on the distribution functions of $u^+(X)$ and $u^-(X)$. Using a different proof technique that employs the dominated convergence theorem and DKW inequalities, one can obtain results similar to Proposition 1–2 with (A1') and (A2) only. The following claim makes this precise.

Proposition 3. Assume (A1') and (A2). Then, we have that

$$\bar{\mathbb{C}}_n \rightarrow \mathbb{C}(X) \text{ a.s. as } n \rightarrow \infty$$

In addition, if we assume (A2'), we have $\forall \epsilon > 0, \delta > 0$,

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

Note that according to this proposition, our estimation scheme is sample-efficient (choosing the weights to be the identity function, a folklore result of probability theory shows that the sample complexity cannot be improved).

3.2. Estimation scheme for locally Lipschitz weights and discrete X

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

$$F_k = \sum_{i=1}^k p_i \text{ if } k \leq l \text{ and } \sum_{i=k}^K p_i \text{ if } k > l. \quad (4)$$

Then, the CPT-value is defined as

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

where u^+, u^- are utility functions and w^+, w^- are weight functions corresponding to gains and losses, respectively. The utility functions u^+ and u^- are non-decreasing, while the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$.

Estimation scheme. Let $\hat{p}_k = \frac{1}{n} \sum_{i=1}^n I_{\{U=x_k\}}$ and

$$\hat{F}_k = \sum_{i=1}^k \hat{p}_i \text{ if } k \leq l \text{ and } \sum_{i=k}^K \hat{p}_i \text{ if } k > l. \quad (5)$$

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

$$\begin{aligned} \bar{\mathbb{C}}_n = & u^-(x_1)w^-(\hat{p}_1) + \sum_{i=2}^l u^-(x_i) \left(w^-(\hat{F}_i) - w^-(\hat{F}_{i-1}) \right) \\ & + \sum_{i=l+1}^{K-1} u^+(x_i) \left(w^+(\hat{F}_i) - w^+(\hat{F}_{i+1}) \right) + u^+(x_K)w^+(\hat{p}_K). \end{aligned}$$

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

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

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

Proposition 4. Assume (A3). Let $L = \max\{L_k, k = 2 \dots K\}$, where L_k is the local Lipschitz constant of function $w^-(x)$ at points F_k , where $k = 1, \dots, l$, and of function $w^+(x)$ at points $k = l + 1, \dots, K$. Let $M = \max\{u^-(x_k), k = 1, \dots, l\} \cup \{u^+(x_k), k = l + 1, \dots, K\}$ and $\rho = \min\{\rho_k\}$, where ρ_k is half the length of the interval centered at point F_k where (A3) holds with constant L_k . Then, $\forall \epsilon > 0, \delta > 0$, we have

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

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

In comparison to Propositions 2 and 3, observe that the sample complexity for discrete X scales with the local Lipschitz constant L and this can be much smaller than the global Lipschitz constant of the weight functions, or the weight functions may not be Lipschitz globally.

The detailed proofs of Propositions 1–4 are available in (Prashanth et al., 2015).

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

Optimization objective: Suppose the r.v. X in (1) is a function of a d -dimensional parameter θ . In this section we consider the problem

$$\text{Find } \theta^* = \arg \max_{\theta \in \Theta} \mathbb{C}(X^\theta), \quad (6)$$

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

4.1. Gradient estimation

Given that we operate in a learning setting and only have biased estimates of the CPT-value from Algorithm 1, we require a simulation scheme to estimate $\nabla \mathbb{C}(X^\theta)$. Simultaneous perturbation methods are a general class of stochastic gradient schemes that optimize a function given only noisy sample values - see (Bhatnagar et al., 2013) for a textbook introduction. SPSA is a well-known scheme that estimates the gradient using two sample values. In our context, at any iteration n of CPT-SPSA-G, with parameter θ_n , the gradient $\nabla \mathbb{C}(X^{\theta_n})$ is estimated as follows: For any $i = 1, \dots, d$,

$$\hat{\nabla}_i \mathbb{C}(X^\theta) = \frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i}, \quad (7)$$

where δ_n is a positive scalar that satisfies (A3) below, $\Delta_n = (\Delta_n^1, \dots, \Delta_n^d)^\top$, where $\{\Delta_n^i, i = 1, \dots, d\}$, $n = 1, 2, \dots$ are i.i.d. Rademacher, independent of $\theta_0, \dots, \theta_n$ and $\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n}$ (resp. $\bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}$) denotes the CPT-value estimate that uses m_n samples of the r.v. $X^{\theta_n + \delta_n \Delta_n}$ (resp. $X^{\theta_n - \delta_n \Delta_n}$). The (asymptotic) unbiasedness of the gradient estimate is proven in Lemma 5.

4.2. Update rule

We incrementally update the parameter θ in the ascent direction as follows: For $i = 1, \dots, d$,

$$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i + \gamma_n \hat{\nabla}_i \mathbb{C}(X^{\theta_n}) \right), \quad (8)$$

where γ_n is a step-size chosen to satisfy (A3) below and $\Gamma = (\Gamma_1, \dots, \Gamma_d)$ is an operator that ensures that the update (8) stays bounded within a compact and convex set Θ . Algorithm 2 presents the pseudocode.

On the number of samples m_n per iteration: The CPT-value estimation scheme is biased, i.e., providing samples with parameter θ_n at instant n , we obtain its CPT-value estimate as $\mathbb{C}(X^{\theta_n}) + \epsilon_n^\theta$, with ϵ_n^θ denoting the bias. The bias can be controlled by increasing the number of samples m_n in each iteration of CPT-SPSA (see Algorithm 2). This is unlike many simulation optimization settings where one only sees function evaluations with zero mean noise and there is no question of deciding on m_n to control the bias as we have in our setting.

To motivate the choice for m_n , we first rewrite the update rule (8) as follows:

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

Algorithm 2 Structure of CPT-SPSA-G algorithm.

Input: initial parameter $\theta_0 \in \Theta$ where Θ is a compact and convex subset of $\mathbb{R}^d$, perturbation constants $\delta_n > 0$, sample sizes $\{m_n\}$, step-sizes $\{\gamma_n\}$, operator $\Gamma : \mathbb{R}^d \rightarrow \Theta$.

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

Generate $\{\Delta_n^i, i = 1, \dots, d\}$ using Rademacher distribution, independent of $\{\Delta_m, m = 0, 1, \dots, n-1\}$.

CPT-value Estimation (Trajectory 1)

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

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

CPT-value Estimation (Trajectory 2)

Simulate m_n samples using $(\theta_n - \delta_n \Delta_n)$.

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

Gradient Ascent

Update θ_n using (8).

end for

Return θ_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 (Borkar, 2008)):

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

While Theorems 1–2 show that the bias ϵ^θ is bounded above, to establish convergence of the policy gradient recursion (8), we increase the number of samples m_n so that the bias vanishes asymptotically. The assumption below provides a condition on the increase rate of m_n .

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

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

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

4.3. Convergence result³

Theorem 1. Assume (A1)-(A3) and also that $\mathbb{C}(X^\theta)$ is a continuously differentiable function of θ , for any $\theta \in \Theta^4$. Consider the ordinary differential equation (ODE):

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

³The detailed proof is available in (Prashanth et al., 2015).

⁴In a typical RL setting, it is sufficient to assume that the policy is continuously differentiable in θ .

where $\check{\Gamma}_i(f(\theta)) := \lim_{\alpha \downarrow 0} \frac{\Gamma_i(\theta + \alpha f(\theta)) - \theta}{\alpha}$, for any continuous $f(\cdot)$. Let $\mathcal{K} = \{\theta \mid \check{\Gamma}_i(\nabla_i \mathbb{C}(X^\theta)) = 0, \forall i = 1, \dots, d\}$. Then, for θ_n governed by (8), we have

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

5. Simulation Experiments

We consider a traffic signal control application where the aim is to improve the road user experience by an adaptive traffic light control (TLC) algorithm. We apply the CPT-functional to the delay experienced by road users, since CPT realistically captures the attitude of the road users towards delays. We then optimize the CPT-value of the delay and contrast this approach with traditional expected delay optimizing algorithms. It is assumed that the CPT functional's parameters (u, w) are given (usually, these are obtained by observing human behavior). The experiments are performed using the GLD traffic simulator (Wiering et al., 2004) and the implementation is available at <https://bitbucket.org/prashla/rl-gld>.

We consider a road network with $\mathcal{N}$ signalled lanes that are spread across junctions and $\mathcal{M}$ paths, where each path connects (uniquely) two edge nodes, from which the traffic is generated – cf. Fig. 4(a). At any instant n , let q_n^i and t_n^i denote the queue length and elapsed time since the lane turned red, for any lane $i = 1, \dots, \mathcal{N}$. Let $d_n^{i,j}$ denote the delay experienced by j th road user on i th path, for any $i = 1, \dots, \mathcal{M}$ and $j = 1, \dots, n_i$, where n_i denotes the number of road users on path i . We specify the various components of the traffic control MDP below. The state $s_n = (q_n^1, \dots, q_n^{\mathcal{N}}, t_n^1, \dots, t_n^{\mathcal{N}}, d_n^{1,1}, \dots, d_n^{\mathcal{M}, n_{\mathcal{M}}})^\top$ is a vector of lane-wise queue lengths, elapsed times and path-wise delays. The actions are the feasible traffic signal configurations.

We consider three different notions of return as follows:

CPT: Let μ^i be the proportion of road users along path i , for $i = 1, \dots, \mathcal{M}$. Any road user along path i , will evaluate the delay he experiences in a manner that is captured well by CPT. Let X_i be the delay r.v. for path i and let the corresponding CPT-value be $\mathbb{C}(X_i)$. With the objective of maximizing the experience of road users across paths, the overall return to be optimized is given by

$$\text{CPT}(X_1, \dots, X_{\mathcal{M}}) = \sum_{i=1}^{\mathcal{M}} \mu^i \mathbb{C}(X_i). \quad (9)$$

EUT: Here we only use the utility functions u^+ and u^- to handle gains and losses, but do not distort probabilities. Thus, the EUT objective is defined as

$$\text{EUT}(X_1, \dots, X_{\mathcal{M}}) = \sum_{i=1}^{\mathcal{M}} \mu^i (\mathbb{E}(u^+(X_i)) - \mathbb{E}(u^-(X_i))),$$

where $\mathbb{E}(u^+(X_i)) = \int_0^{+\infty} \mathbb{P}(u^+(X_i) > z) dz$ and $\mathbb{E}(u^-(X_i)) = \int_0^{+\infty} \mathbb{P}(u^-(X_i) > z) dz$, for $i = 1, \dots, \mathcal{M}$.

AVG: This is EUT without the distinction between gains and losses via utility functions, i.e.,

$$\text{AVG}(X_1, \dots, X_{\mathcal{M}}) = \sum_{i=1}^{\mathcal{M}} \mu^i \mathbb{E}(X_i).$$

An important component of CPT is to employ a reference point to calculate gains and losses. In our setting, we use path-wise delays obtained from a pre-timed TLC (cf. the Fixed TLCs in (Prashanth & Bhatnagar, 2011)) as the reference point. If the delay of any algorithm (say CPT-SPSA) is less than that of pre-timed TLC, then the (positive) difference in delays is perceived as a gain and in the complementary case, the delay difference is perceived as a loss. Thus, the CPT-value $\mathbb{C}(X_i)$ for any path i in (9) is to be understood as a *differential delay*.

Using a Boltzmann policy that has the form

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

with features $\phi_{s,a}$ as described in Section V-B of (Prashanth & Bhatnagar, 2012), we implement the following TLC algorithms:

CPT-SPSA: This is the first-order algorithm with SPSA-based gradient estimates, as described in Algorithm 2. In particular, the estimation scheme in Algorithm 1 is invoked to estimate $\mathbb{C}(X_i)$ for each path $i = 1, \dots, \mathcal{M}$, with $d_n^{i,j}, j = 1, \dots, n_i$ as the samples.

EUT-SPSA: This is similar to CPT-SPSA, except that weight functions $w^+(p) = w^-(p) = p$, for $p \in [0, 1]$.

AVG-SPSA: This is similar to CPT-SPSA, except that weight functions $w^+(p) = w^-(p) = p$, for $p \in [0, 1]$.

For both CPT-SPSA and EUT-SPSA, we set the utility functions (see (1)) as follows:

$$u^+(x) = |x|^\sigma, \text{ and } u^-(x) = \lambda |x|^\sigma,$$

where $\lambda = 2.25$ and $\sigma = 0.88$. For CPT-SPSA, we set the weights as follows:

$$w^+(p) = \frac{p^{\eta_1}}{(p^{\eta_1} + (1-p)^{\eta_1})^{\frac{1}{\eta_1}}}, \text{ and } w^-(p) = \frac{p^{\eta_2}}{(p^{\eta_2} + (1-p)^{\eta_2})^{\frac{1}{\eta_2}}},$$

where $\eta_1 = 0.61$ and $\eta_2 = 0.69$. The choices for λ, σ, η_1 and η_2 are based on median estimates given by (Tversky & Kahneman, 1992) and have been used earlier in a traffic application (see (Gao et al., 2010)). For all the algorithms, motivated by standard guidelines (see Spall 2005), we set $\delta_n = 1.9/n^{0.101}$ and $a_n = 1/(n+50)$. The initial point θ_0

is the d -dimensional vector of ones and $\forall i$, the operator Γ_i projects θ_i onto the set $[0.1, 10.0]$.

The experiments involve two phases: first, a training phase where we run each algorithm for 200 iterations, with each iteration involving two perturbed simulations, each of trajectory length 500. This is followed by a test phase where we fix the policy for each algorithm and 100 independent simulations of the MDP (each with a trajectory length of 1000) are performed. After each run in the test phase, the overall CPT-value (9) is estimated.

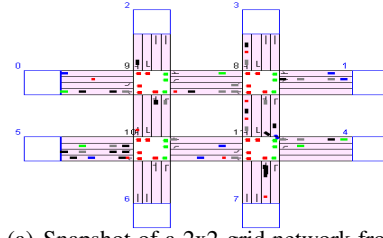
Figures 4(b)–4(d) present the histogram of the CPT-values from the test phase for AVG-SPSA, EUT-SPSA and CPT-SPSA, respectively. A similar exercise for pre-timed TLC resulted in a CPT-value of -46.14 . It is evident that each algorithm converges to a different policy. However, the CPT-value of the resulting policies is highest in the case of CPT-SPSA, followed by EUT-SPSA and AVG-SPSA in that order. Intuitively, this is expected because AVG-SPSA uses neither utilities nor probability distortions, while EUT-SPSA distinguishes between gains and losses using utilities while not using weights to distort probabilities. The results in Figure 4 argue for specialized algorithms that incorporate CPT-based criteria, esp. in the light of previous findings which show CPT matches human evaluation well and there is a need for algorithms that serve human needs well.

6. Conclusions

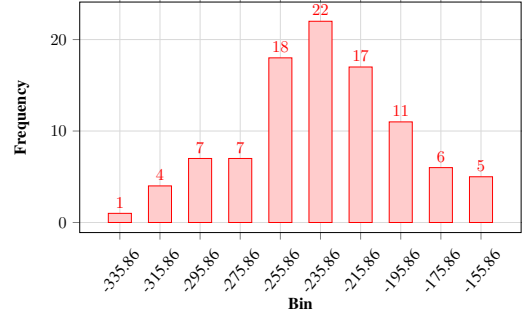
CPT has been a very popular paradigm for modeling human decisions among psychologists/economists, but has escaped the radar of the AI community. This work is the first step in incorporating CPT-based criteria into an RL framework. However, both prediction and control of CPT-based value is challenging. For prediction, we proposed a quantile-based estimation scheme. 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 a first-order algorithm for optimizing the CPT-value. We provided theoretical convergence guarantees for all the proposed algorithms and illustrated the usefulness of our algorithms for optimizing CPT-based criteria in a traffic signal control application.

ACKNOWLEDGMENTS

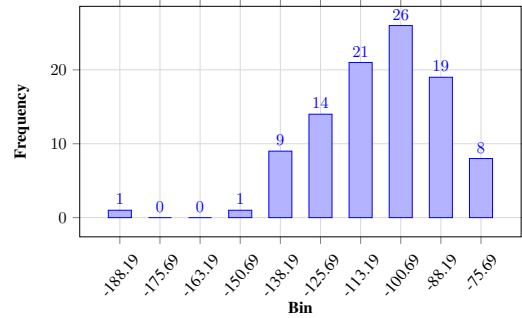
This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning, NSERC, the National Science Foundation (NSF) under Grants CMMI-1434419, CNS-1446665, and CMMI-1362303, and by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-15-10050.



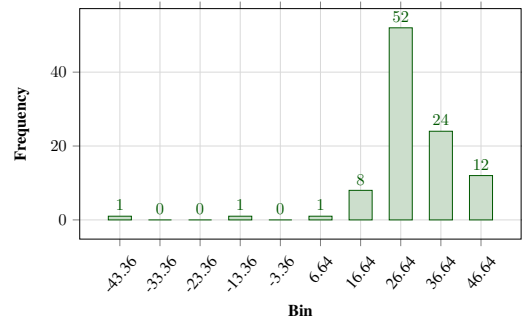
(a) Snapshot of a 2x2-grid network from GLD simulator. The figure shows eight edge nodes that generate traffic, four traffic lights and four-laned roads carrying cars.



(b) AVG-SPSA



(c) EUT-SPSA



(d) CPT-SPSA

Figure 4. Histogram of CPT-value of the differential delay (calculated with a pre-timed TLC as reference point) for three different algorithms (all based on SPSA): AVG uses plain sample means (no utility/weights), EUT uses utilities but no weights and CPT uses both utilities and weights. Note: larger values are better.

References

- Allais, M. Le comportement de l'homme rationel devant le risque: Critique des postulats et axiomes de l'école américaine. *Econometrica*, 21:503–546, 1953.
- Athreya, K. B. and Lahiri, S. N. *Measure Theory and Probability Theory*. Springer Science & Business Media, 2006.
- Barberis, Nicholas C. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27(1):173–196, 2013.
- Bertsekas, Dimitri P. *Dynamic Programming and Optimal Control*, vol. II, 3rd edition. Athena Scientific, 2007.
- Bhatnagar, S. and Prashanth, L. A. Simultaneous perturbation Newton algorithms for simulation optimization. *Journal of Optimization Theory and Applications*, 164(2):621–643, 2015.
- Bhatnagar, S., Prasad, H. L., and Prashanth, L. A. *Stochastic Recursive Algorithms for Optimization*, volume 434. Springer, 2013.
- Borkar, V. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.
- Borkar, V. and Jain, R. Risk-constrained Markov decision processes. In *IEEE Conference on Decision and Control (CDC)*, pp. 2664–2669, 2010.
- Ellsberg, D. Risk, ambiguity and the Savage's axioms. *The Quarterly Journal of Economics*, 75(4):643–669, 1961.
- Fathi, M. and Frikha, N. Transport-entropy inequalities and deviation estimates for stochastic approximation schemes. *Electronic Journal of Probability*, 18(67):1–36, 2013.
- Fennema, H. and Wakker, P. Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making*, 10:53–64, 1997.
- Filar, J., Kallenberg, L., and Lee, H. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14(1):147–161, 1989.
- Fishburn, P.C. *Utility Theory for Decision Making*. Wiley, New York, 1970.
- Fu, M. C. (ed.). *Handbook of Simulation Optimization*. Springer, 2015.
- Gao, Song, Frejinger, Emma, and Ben-Akiva, Moshe. Adaptive route choices in risky traffic networks: A prospect theory approach. *Transportation Research Part C: Emerging Technologies*, 18(5):727–740, 2010.
- Gill, P.E., Murray, W., and Wright, M.H. *Practical Optimization*. Academic Press, 1981.
- Kahneman, D. and Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica: Journal of the Econometric Society*, pp. 263–291, 1979.
- Kushner, H. and Clark, D. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.
- Lin, K. *Stochastic Systems with Cumulative Prospect Theory*. Ph.D. Thesis, University of Maryland, College Park, 2013.
- Mannor, Shie and Tsitsiklis, John N. Algorithmic aspects of mean–variance optimization in Markov decision processes. *European Journal of Operational Research*, 231(3):645–653, 2013.
- Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.
- Prashanth, L. A. Policy Gradients for CVaR-Constrained MDPs. In *Algorithmic Learning Theory*, pp. 155–169. Springer International Publishing, 2014.
- Prashanth, L.A. and Bhatnagar, S. Reinforcement Learning With Function Approximation for Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):412–421, 2011.
- Prashanth, L.A. and Bhatnagar, S. Threshold Tuning Using Stochastic Optimization for Graded Signal Control. *IEEE Transactions on Vehicular Technology*, 61(9):3865–3880, 2012.
- Prashanth, L.A., Jie, Cheng, Fu, M. C., Marcus, S. I., and Szepesvári, Csaba. Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control. *arXiv preprint arXiv:1506.02632v3*, 2015.
- Prelec, Drazen. The probability weighting function. *Econometrica*, pp. 497–527, 1998.
- Quiggin, John. *Generalized Expected Utility Theory: The Rank-dependent Model*. Springer Science & Business Media, 2012.
- Ruppert, D. Stochastic approximation. *Handbook of Sequential Analysis*, pp. 503–529, 1991.
- Simon, Herbert Alexander. Theories of decision-making in economics and behavioral science. *The American Economic Review*, 49:253–283, 1959.
- Sobel, M. The variance of discounted Markov decision processes. *Journal of Applied Probability*, pp. 794–802, 1982.

- Spall, J. C. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Auto. Cont.*, 37(3):332–341, 1992.
- Spall, J. C. Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Trans. Autom. Contr.*, 45:1839–1853, 2000.
- Spall, J. C. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*, volume 65. John Wiley & Sons, 2005.
- Starmer, Chris. Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of economic literature*, pp. 332–382, 2000.
- Tamar, Aviv, Glassner, Yonatan, and Mannor, Shie. Optimizing the CVaR via sampling. *arXiv preprint arXiv:1404.3862*, 2014.
- Tversky, A. and Kahneman, D. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.
- Von Neumann, J. and Morgenstern, O. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, 1944.
- Wasserman, L. A. *All of Nonparametric Statistics*. Springer, 2015.
- Wiering, M., Vreeken, J., van Veenen, J., and Koopman, A. Simulation and optimization of traffic in a city. In *IEEE Intelligent Vehicles Symposium*, pp. 453–458, June 2004.

Appendix

A. Background on CPT

Cs: Is this section referred to from the main text?

The purpose of this appendix is to explain the ideas underlying CPT. To avoid technicalities we will do this in the context of discrete random variables. Let X be a discrete valued random variables, taking values in $\{x_1, \dots, x_K\}$ (the set of possible “prospects”) and let $p_i = \mathbb{P}(X = x_i)$. We shall think of X as a random loss or gain incurred by a human decision maker. Given a utility function $u : \mathbb{R} \rightarrow \mathbb{R}_+ : [0, 1] \rightarrow [0, 1]$ and weighting function w , the **prospect theory** (PT) value of X is defined as

Cs: What is the domain and range of these functions?

$$\mathbb{C}(X) = \sum_{i=1}^K u(x_i)w(p_i). \quad (10)$$

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. Kahneman & Tversky (1979) and Fennema & Wakker (1997) present justification of PT in the form of experimental results using human subjects.

However, the weight function in PT lacking in some theoretical aspects as it violates first-order *stochastic dominance*. We illustrate this via the following example: Consider a prospect, say P1, where outcomes 10 and 0 occur with probabilities 0.1, 0.9, respectively. Let P2 be another prospect where the outcomes are 10, $10 + \epsilon$ and 0, with respective probabilities 0.05, 0.05 and 0.9.

The weight function w is non-additive as it is non-linear and since 0.05 and 0.1 are in the low-probability regime, we can assume that $w(0.1) > 2w(0.05)$. Suppose $\epsilon > 0$ be small enough such that $w(0.1) > (2 + \epsilon)w(0.05)$. Then, the PT-value, as defined in (10), of P1 is higher than that of P2 and hence, P1 is preferred.

On the other hand, P2 stochastically dominates P1, where stochastic dominance is defined as follows: A prospect $\mathcal{B}$ stochastically dominates prospect $\mathcal{A}$ if the probability of receiving a value x or greater is at least as high under prospect $\mathcal{B}$ as it is under prospect $\mathcal{A}$ for all values of x , and is strictly greater for some value of x .

Cumulative prospect theory (CPT) (Tversky & Kahneman, 1992) uses a similar measure as PT, except that the weights are a function of cumulative probabilities. First, separate the gains and losses as $x_1 \leq \dots \leq x_l \leq 0 \leq x_{l+1} \leq \dots \leq x_K$, where zero is used as an arbitrary reference point. Then, the CPT-value of X is defined as

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

Cs: What does “it” refer to here? Consider rephrasing this

Cs: Why is this true??

Cs: What does “this” refer to here?

Cs: This is hard to understand; consider rephrasing it more clearly.

where u^+, u^- are utility functions and w^+, w^- are weight functions to distort gains and losses, respectively. The utility functions u^+ and u^- are non-decreasing, while the weight functions are continuous, non-decreasing and have the range $[0, 1]$ with $w^+(0) = w^-(0) = 0$ and $w^+(1) = w^-(1) = 1$. Unlike PT, the CPT-value does not violate first-order stochastic dominance, as the weight functions act on cumulative probabilities.

Allais paradox

Allais paradox is not as much as a paradox, but an argument that shows that expected utility theory is inconsistent with human preferences. Suppose we have the following two traffic light switching policies:

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

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

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

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

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

Humans usually choose Policy 4 over Policy 3.

We can now argue against using expected utility (EU) as an objective as follows accepting the above preferences: Let u be the utility function in EU. Since Policy 1 is preferred over Policy 2, $u(1000) > 0.1u(10000) + 0.89u(1000) + 0.01u(100)$, which is equivalent to

$$0.11u(1000) > 0.1u(10000) + 0.01u(100). \quad (11)$$

On the other hand, since Policy 4 is preferred over Policy 3, $0.89u(100) + 0.11u(1000) < 0.9u(100) + 0.1u(10000)$, which is equivalent to

$$0.11u(1000) < 0.1u(10000) + 0.01u(100). \quad (12)$$

Now note that (11) contradicts (12). Hence, no utility function exist that is consistent with human preferences.

B. CPT-value in a Stochastic Shortest Path Setting

We consider a stochastic shortest path (SSP) problem with states $\mathcal{S} = \{0, \dots, \mathcal{L}\}$, where 0 is a special reward-free absorbing state. A randomized policy π is a function that maps any state $s \in \mathcal{S}$ onto a probability distribution over the actions $\mathcal{A}(s)$ in state s . As is standard in policy gradient algorithms, we parameterize π and assume it is continuously differentiable in its parameter $\theta \in \mathbb{R}^d$. An *episode* is a simulated sample path using policy θ that starts in state $s^0 \in \mathcal{S}$, visits $\{s_1, \dots, s_{\tau-1}\}$ before ending in the absorbing state 0, where τ is the first passage time to state 0. Let $D^\theta(s^0)$ be a random variable (r.v) that denote the total reward from an episode, defined by

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

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

Instead of the traditional RL objective for an SSP of maximizing the expected value $\mathbb{E}(D^\theta(s^0))$, we adopt the CPT approach and aim to solve the following problem:

$$\max_{\theta \in \Theta} \mathbb{C}(D^\theta(s^0)),$$

where Θ is the set of admissible policies that are *proper*⁵ and the CPT-value function $\mathbb{C}(D^\theta(s^0))$ is defined as

$$\begin{aligned} \mathbb{C}(D^\theta(s^0)) &= \int_0^{+\infty} w^+ (\mathbb{P}(u^+(D^\theta(s^0)) > z)) dz \\ &\quad - \int_0^{+\infty} w^- (\mathbb{P}(u^-(D^\theta(s^0)) > z)) dz. \end{aligned} \quad (13)$$

C. Proofs for CPT-value estimator

C.1. Hölder continuous weights

For proving Proposition 1 and 4, we require Hoeffding's inequality, which is given below.

⁵A policy θ is proper if 0 is recurrent and all other states are transient for the Markov chain underlying θ . It is standard to assume that policies are proper in an SSP setting - cf. (Bertsekas, 2007).

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

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

Proposition 5. Under (A1'), the CPT-value $\mathbb{C}(X)$ as defined by (13) is finite.

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

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

The second inequality is valid since $\mathbb{P}(u^+(X) > z) \leq 1$. The claim follows for the first integral in (13) and the finiteness of the second integral in (13) can be argued in an analogous fashion. $\square$

Proposition 6. Assume (A1'). Let $\xi_{\frac{i}{n}}^+$ and $\xi_{\frac{i}{n}}^-$ denote the $\frac{i}{n}$ th quantile of $u^+(X)$ and $u^-(X)$, respectively. Then, we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} \xi_{\frac{i}{n}}^+ \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) &= \int_0^{+\infty} w^+(\mathbb{P}(u^+(X) > z)) dz < +\infty, \\ \lim_{n \rightarrow \infty} \sum_{i=0}^{n-1} \xi_{\frac{i}{n}}^- \left(w^- \left(\frac{n-i}{n} \right) - w^- \left(\frac{n-i-1}{n} \right) \right) &= \int_0^{+\infty} w^-(\mathbb{P}(u^-(X) > z)) dz < +\infty \end{aligned} \quad (14)$$

Cs: Fix the ugly parenthesis! Also, why do we suddenly number all equations as if they all needed a number??

Ch: Fix the upper limit..I think it should run up to n

Proof. We shall focus on proving the first part of equation (14). Consider the following linear combination of simple functions:

$$\sum_{i=0}^{n-1} w^+ \left(\frac{i}{n} \right) \cdot I_{\left[\xi_{\frac{n-i-1}{n}}^+, \xi_{\frac{n-i}{n}}^+ \right]}(t), \quad (15)$$

which will converge almost everywhere to the function $w^+(\mathbb{P}(u^+(X) > t))$ in the interval $[0, +\infty)$, and also notice that

$$\sum_{i=0}^{n-1} w^+ \left(\frac{i}{n} \right) \cdot I_{\left[\xi_{\frac{n-i-1}{n}}^+, \xi_{\frac{n-i}{n}}^+ \right]}(t) < w^+(\mathbb{P}(u^+(X) > t)), \quad \forall t \in [0, +\infty).$$

The integral of (15) can be simplified as follows:

$$\begin{aligned} \int_0^{+\infty} \sum_{i=0}^{n-1} w^+ \left(\frac{i}{n} \right) \cdot I_{\left[\xi_{\frac{n-i-1}{n}}^+, \xi_{\frac{n-i}{n}}^+ \right]}(t) &= \sum_{i=0}^{n-1} w_{\frac{i}{n}}^+(t) \cdot \left(\xi_{\frac{n-i}{n}}^+ - \xi_{\frac{n-i-1}{n}}^+ \right) \\ &= \sum_{i=0}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right). \end{aligned}$$

The Hölder continuity property assures the fact that $\lim_{n \rightarrow \infty} |w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})| = 0$, and the limit in (14) holds through a typical application of the dominated convergence theorem.

The second part of (14) can be justified in a similar fashion. $\square$

Proof of Proposition 1

Proof. Without loss of generality, assume that Hölder constant H of the weight functions w^+ and w^- are both 1. We first prove that

$$\bar{\mathbb{C}}_n^+ \rightarrow \mathbb{C}^+(X) \text{ a.s. as } n \rightarrow \infty.$$

The statement above is equivalent to the following:

CS: Don't start sentences with "and" or "or".

$$\lim_{n \rightarrow +\infty} \sum_{i=1}^{n-1} u^+(X_{[i]}) \left(w^+\left(\frac{n-i+1}{n}\right) - w^+\left(\frac{n-i}{n}\right) \right) \xrightarrow{n \rightarrow \infty} \int_0^{+\infty} w^+(P(U > t)) dt, \text{ w.p. 1} \quad (16)$$

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

$$P \left(\left| \sum_{i=1}^{n-1} u^+(X_{[i]}) \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) \right| > \epsilon \right), \quad (17)$$

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

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i=1}^{n-1} u^+(X_{[i]}) \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) \right| > \epsilon \right) \\ & \leq \mathbb{P} \left(\bigcup_{i=1}^{n-1} \left\{ \left| u^+(X_{[i]}) \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) - \xi_{\frac{i}{n}}^+ \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) \right| > \frac{\epsilon}{n} \right\} \right) \\ & \leq \sum_{i=1}^{n-1} \mathbb{P} \left(\left| u^+(X_{[i]}) \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) - \xi_{\frac{i}{n}}^+ \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) \right| > \frac{\epsilon}{n} \right) \end{aligned} \quad (18)$$

$$\begin{aligned} & = \sum_{i=1}^{n-1} \mathbb{P} \left(\left| \left(u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ \right) \cdot \left(w^+\left(\frac{n-i}{n}\right) - w^+\left(\frac{n-i-1}{n}\right) \right) \right| > \frac{\epsilon}{n} \right) \\ & \leq \sum_{i=1}^{n-1} \mathbb{P} \left(\left| \left(u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ \right) \cdot \left(\frac{1}{n} \right)^\alpha \right| > \frac{\epsilon}{n} \right) \\ & = \sum_{i=1}^{n-1} \mathbb{P} \left(\left| u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ \right| > \frac{\epsilon}{n^{1-\alpha}} \right). \end{aligned} \quad (19)$$

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

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

We focus on the term $\mathbb{P} \left(u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ > \frac{\epsilon}{n^{(1-\alpha)}} \right)$. Let $W_t = I_{\left(u^+(X_t) > \xi_{\frac{i}{n}}^+ + \frac{\epsilon}{n^{(1-\alpha)}} \right)}$, $t = 1, \dots, n$. Using the fact that a probability distribution function is non-decreasing, we obtain

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

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

$$\mathbb{P} \left(\sum_{i=1}^n W_t - n \cdot \left[1 - F^+ \left(\xi_{\frac{i}{n}}^+ + \frac{\epsilon}{n^{(1-\alpha)}} \right) \right] > n \cdot \left[F^+ \left(\xi_{\frac{i}{n}}^+ + \frac{\epsilon}{n^{(1-\alpha)}} \right) - \frac{i}{n} \right] \right) < e^{-2n \cdot \delta'_i}, \quad (20)$$

where $\delta'_i = F^+ \left(\xi_{\frac{i}{n}}^+ + \frac{\epsilon}{n^{(1-\alpha)}} \right) - \frac{i}{n}$. Since F^+ is Lipschitz, we have that $\delta'_i \leq L^+ \cdot \left(\frac{\epsilon}{n^{(1-\alpha)}} \right)$. Hence, we obtain

$$\mathbb{P} \left(u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ > \frac{\epsilon}{n^{(1-\alpha)}} \right) < e^{-2n \cdot L^+ \cdot \frac{\epsilon}{n^{(1-\alpha)}}} = e^{-2n^\alpha \cdot L^+ \epsilon} \quad (21)$$

In a similar fashion, one can show that

$$\mathbb{P} \left(u^+(X_{[i]}) - \xi_{\frac{i}{n}}^+ < -\frac{\epsilon}{n^{(1-\alpha)}} \right) \leq e^{-2n^\alpha \cdot L^+ \epsilon}. \quad (22)$$

Combining (21) and (22), we obtain

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

Plugging the above in (19), we obtain

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i=1}^{n-1} u^+(X_{[i]}) \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) \right| > \epsilon \right) \\ & \leq 2n \cdot e^{-2n^\alpha \cdot L^+ \epsilon}. \end{aligned} \quad (23)$$

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

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

$$\begin{aligned} & \mathbb{P} \left(\left| \sum_{i=1}^{n-1} u^+(X_{[i]}) \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) \right| > \epsilon, i.o. \right) \\ & = 0, \end{aligned}$$

which implies

$$\begin{aligned} & \sum_{i=1}^{n-1} u^+(X_{[i]}) \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) - \sum_{i=1}^{n-1} \xi_{\frac{i}{n}}^+ \cdot \left(w^+ \left(\frac{n-i}{n} \right) - w^+ \left(\frac{n-i-1}{n} \right) \right) \\ & \xrightarrow{n \rightarrow +\infty} 0 \text{ w.p } 1, \end{aligned}$$

which proves (16).

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

Proof of Proposition 2

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

Let $\hat{F}_n(u) = \frac{1}{n} \sum_{i=1}^n I_{[(u(X_i)) \leq u]}$ denote the empirical distribution of a r.v. U , with $u(X_1), \dots, u(X_n)$ being sampled from the r.v. $u(X)$. The, for any n and $\epsilon > 0$, we have

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

Cs: Is not $\mathbb{N} \cap (0, 1) = \emptyset$???

The reader is referred to Chapter 2 of (Wasserman, 2015) for detailed description of empirical distributions in general and DKW inequality in particular.

Proof. We prove the w^+ part, and the w^- part follows in a similar fashion. Since $u^+(X)$ is bounded above by M and w^+ is Hölder-continuous, we have

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

Now, plugging in the DKW inequality, we obtain

$$\begin{aligned} & \mathbb{P} \left(\left| \int_0^{+\infty} w^+ (\mathbb{P}(u^+(X) > t)) dt - \int_0^{+\infty} w^+ (1 - \hat{F}_n^+(t)) dt \right| > \epsilon \right) \\ &\leq \mathbb{P} \left(HM \sup_{t \in \mathbb{R}} |\mathbb{P}(u^+(X) < t) - \hat{F}_n^+(t)|^\alpha > \epsilon \right) \leq e^{-n \frac{\epsilon(2/\alpha)}{2H^2 M^2}}. \end{aligned} \quad (24)$$

The claim follows. $\square$

C.2. Lipschitz continuous weights

Setting $\alpha = \gamma = 1$ in the proof of Proposition 3, it is easy to see that the CPT-value (13) is finite.

Next, in order to prove the asymptotic convergence claim in Proposition 3, we require the dominated convergence theorem in its generalized form, which is provided below.

Theorem 4. (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 $\lim_{n \rightarrow \infty} \int_E g_n = \int_E g$, then $\lim_{n \rightarrow \infty} \int_E f_n = \int_E f$.

Proof. This is a standard result that can be found in any textbook on measure theory. For instance, see Theorem 2.3.11 in (Athreya & Lahiri, 2006). $\square$

Proof of Proposition 3: Asymptotic convergence

Proof. Notice the the following equivalence:

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

and also,

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

where $\hat{F}_n^+(x)$ and $\hat{F}_n^-(x)$ is the empirical distribution of $u^+(X)$ and $u^-(X)$.

Thus, the CPT estimator $\bar{\mathbb{C}}_n$ in Algorithm 1 can be written equivalently as follows:

$$\bar{\mathbb{C}}_n = \int_0^{+\infty} w^+ (1 - \hat{F}_n^+(x)) dx - \int_0^{+\infty} w^- (1 - \hat{F}_n^-(x)) dx. \quad (25)$$

Cs: The hats should be on F only, as in $\hat{F}^+$ as opposed to F^+ .

We first prove the asymptotic convergence claim for the first integral in (25), i.e., we show

$$\int_0^{+\infty} w^+ \left(1 - \hat{F}_n^+(x)\right) dx \rightarrow \int_0^{+\infty} w^+ \left(P(u^+(X) > x)\right) dx. \quad (26)$$

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

We have

$$\int_0^{+\infty} \left(P(u^+(X) > x)\right) dx = E(u^+(X))$$

and

$$\int_0^{+\infty} \left(1 - \hat{F}_n^+(x)\right) dx = \int_0^{+\infty} \int_x^\infty d\hat{F}_n(t) dx. \quad (27)$$

Since $\hat{F}_n^+(x)$ has bounded support on $\mathbb{R} \forall n$, the integral in (27) is finite. Applying Fubini's theorem to the RHS of (27), we obtain

$$\int_0^{+\infty} \int_x^\infty d\hat{F}_n(t) dx = \int_0^{+\infty} \int_0^t dx d\hat{F}_n(t) = \int_0^{+\infty} t d\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n u^+(X_{[i]}), \quad (28)$$

where $u^+(X_{[i]})$, $i = 1, \dots, n$ denote the order statistics, i.e., $u^+(X_{[1]}) \leq \dots \leq u^+(X_{[n]})$.

Now, notice that

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

From the foregoing,

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

Hence, we have

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

The claim in (26) 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 $L \cdot (1 - \hat{F}_n^+(x)) \xrightarrow{a.s.} L(P(u^+(X) > x))$ uniformly $\forall x$. The latter fact is implied by the Glivenko-Cantelli theorem (cf. Chapter 2 of (Wasserman, 2015)).

Following similar arguments, it is easy to show that

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

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

Proof of Proposition 3: Sample complexity

Proof. Since $u^+(X)$ is bounded above by M and w^+ is Lipschitz with constant L , we have

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

Now, plugging in the DKW inequality, we obtain

$$\begin{aligned}
 & P \left(\left| \int_0^{+\infty} w^+(P(u^+(X)) > x) dx - \int_0^{+\infty} w^+(1 - \hat{F}_n^+(x)) dx \right| > \epsilon/2 \right) \\
 &\leq P \left(LM \sup_{x \in \mathbb{R}} |P(u^+(X) < x) - \hat{F}_n^+(x)| > \epsilon/2 \right) \leq 2e^{-n \frac{\epsilon^2}{2L^2 M^2}}.
 \end{aligned} \tag{29}$$

Along similar lines, we obtain

$$P \left(\left| \int_0^{+\infty} w^-(P(u^-(X)) > x) dx - \int_0^{+\infty} w^-(1 - \hat{F}_n^-(x)) dx \right| > \epsilon/2 \right) \leq 2e^{-n \frac{\epsilon^2}{2L^2 M^2}}. \tag{30}$$

Combining (29) and (30), we obtain

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

And the claim follows. $\square$

C.3. Proofs for discrete valued X

Without loss of generality, assume $w^+ = w^- = w$, and let

$$\hat{F}_k = \begin{cases} \sum_{i=1}^k \hat{p}_i & \text{if } k \leq l \\ \sum_{i=k}^K \hat{p}_i & \text{if } k > l. \end{cases} \tag{31}$$

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

Proposition 7. *Let F_k and $\hat{F}_k$ be as defined in (4), (31), Then, we have that, for every $\epsilon > 0$,*

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

Proof. We focus on the case when $k > l$, while the case of $k \leq l$ is proved in a similar fashion. Notice that when $k > l$, $\hat{F}_k = I_{(X_i \geq x_k)}$. Since the random variables X_i are independent of each other and for each i , are bounded above by 1, we

can apply Hoeffding's inequality to obtain

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

□

The proof of Proposition 4 requires the following claim which gives the convergence rate under local Lipschitz weights.

Proposition 8. *Under conditions of Proposition 4, with F_k and $\hat{F}_k$ as defined in (4) and (31), we have*

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

$$u_k = \begin{cases} u^-(x_k) & \text{if } k \leq l \\ u^+(x_k) & \text{if } k > l. \end{cases} \quad (32)$$

Proof. Observe that

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

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

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

According to the property of locally Lipschitz continuous, we have

$$\begin{aligned}
 &P\left(|F_k - \hat{F}_k| \leq \delta \cap |u_k w(\hat{F}_k) - u_k w(F_k)| > \frac{\epsilon}{K}\right) \\
 &\leq P(u_k L |F_k - \hat{F}_k| > \frac{\epsilon}{K}) \leq e^{-\epsilon \cdot 2n / (KLu_k)^2} \leq e^{-\epsilon \cdot 2n / (KLA)^2}, \forall k.
 \end{aligned}$$

And similarly,

$$P(|F_k - \hat{F}_k| > \delta) \leq e^{-\delta^2 / 2n}, \forall k.$$

And as a result,

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

□

Proof of Proposition 4

Proof. With u_k as defined in (32), we need to prove that

$$P\left(\left|\sum_{i=1}^K u_k \cdot (w(\hat{F}_k) - w(\hat{F}_{k+1})) - \sum_{i=1}^K u_k \cdot (w(F_k) - w(F_{k+1}))\right| \leq \epsilon\right) > 1 - \rho, \forall n > \frac{\ln(\frac{4K}{A})}{M}, \quad (33)$$

where w is Locally Lipschitz continuous with constants $L_1, \dots, L_K$ at the points $F_1, \dots, F_K$. From a parallel argument to that in the proof of Proposition 8, it is easy to infer that

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

Hence,

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

The claim in (33) now follows. □

D. Proofs for CPT-SPSA-G

To prove the main result in Theorem 1, 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 (Spall, 1992)) since we have a non-zero bias in the function evaluations, while the regular SPSA assumes the noise is zero-mean. Following this lemma, we complete the proof of Theorem 1 by invoking the well-known Kushner-Clark lemma (Kushner & Clark, 1978).

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

$$\left|\mathbb{E}\left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \middle| \mathcal{F}_n\right] - \nabla_i \mathbb{C}(X^{\theta_n})\right| \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (34)$$

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

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

We claim

$$\mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \mid \mathcal{F}_n \right] = \mathbb{E} \left[\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i} \mid \mathcal{F}_n \right] + \mathbb{E} [\eta_n \mid \mathcal{F}_n], \quad (35)$$

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

$$\mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \mid \mathcal{F}_n \right] \xrightarrow{n \rightarrow \infty} \mathbb{E} \left[\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i} \mid \mathcal{F}_n \right]. \quad (36)$$

We now analyse the RHS of (36). By using suitable Taylor's expansions,

$$\begin{aligned} \mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) &= \mathbb{C}(X^{\theta_n}) + \delta_n \Delta_n^\top \nabla \mathbb{C}(X^{\theta_n}) + \frac{\delta_n^2}{2} \Delta_n^\top \nabla^2 \mathbb{C}(X^{\theta_n}) \Delta_n + O(\delta_n^3), \\ \mathbb{C}(X^{\theta_n - \delta_n \Delta_n}) &= \mathbb{C}(X^{\theta_n}) - \delta_n \Delta_n^\top \nabla \mathbb{C}(X^{\theta_n}) + \frac{\delta_n^2}{2} \Delta_n^\top \nabla^2 \mathbb{C}(X^{\theta_n}) \Delta_n + O(\delta_n^3). \end{aligned}$$

From the above, it is easy to see that

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

Taking conditional expectation on both sides, we obtain

$$\begin{aligned} \mathbb{E} \left[\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i} \mid \mathcal{F}_n \right] &= \nabla_i \mathbb{C}(X^{\theta_n}) + \mathbb{E} \left[\sum_{j=1, j \neq i}^N \frac{\Delta_n^j}{\Delta_n^i} \nabla_j \mathbb{C}(X^{\theta_n}) \mid \mathcal{F}_n \right] + O(\delta_n^2) \\ &= \nabla_i \mathbb{C}(X^{\theta_n}) + O(\delta_n^2). \end{aligned} \quad (37)$$

The first equality above follows from the fact that Δ_n is distributed according to a d -dimensional vector of Rademacher random variables and is independent of $\mathcal{F}_n$. The second inequality follows by observing that Δ_n^i is independent of Δ_n^j , for any $i, j = 1, \dots, d, j \neq i$.

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

Proof of Theorem 1

Proof. We first rewrite the update rule (8) as follows: For $i = 1, \dots, d$,

$$\theta_{n+1}^i = \theta_n^i + \gamma_n (\nabla_i \mathbb{C}(X^{\theta_n}) + \beta_n + \xi_n), \quad (38)$$

where

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

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

Convergence of (38) can be inferred from Theorem 5.3.1 on pp. 191-196 of (Kushner & Clark, 1978), provided we verify the necessary assumptions given as (B1)-(B5) below:

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

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

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

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

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

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

$$\dot{\theta}_t^i = \tilde{\Gamma}_i \left(-\nabla \mathbb{C}(X^{\theta_t^i}) \right), \text{ for } i = 1, \dots, d, \quad (39)$$

In the following, we verify the above assumptions for the recursion (8):

- (B1) holds by assumption in our setting.
- Lemma 5 above establishes that the bias β_n is $O(\delta_n^2)$ and since $\delta_n \rightarrow 0$ as $n \rightarrow \infty$, it is easy to see that (B2) is satisfied for β_n .
- (B3) holds by assumption (A3).
- We verify (B4) using arguments similar to those used in (Spall, 1992) for the classic SPSA algorithm:
We first recall Doob's martingale inequality (see (2.1.7) on pp. 27 of (Kushner & Clark, 1978)):

$$P \left(\sup_{m \geq 0} \|W_l\| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \lim_{l \rightarrow \infty} \mathbb{E} \|W_l\|^2. \quad (40)$$

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

$$P \left(\sup_{l \geq k} \left\| \sum_{n=k}^l \gamma_n \xi_n \right\| \geq \epsilon \right) \leq \frac{1}{\epsilon^2} \mathbb{E} \left\| \sum_{n=k}^{\infty} \gamma_n \xi_n \right\|^2 = \frac{1}{\epsilon^2} \sum_{n=k}^{\infty} \gamma_n^2 \mathbb{E} \|\eta_n\|^2. \quad (41)$$

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

$$\mathbb{E} \|\xi_n\|^2 \leq \mathbb{E} \left(\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \right)^2 \quad (42)$$

$$\leq \left(\left(\mathbb{E} \left(\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \right)^2 \right)^{1/2} + \left(\mathbb{E} \left(\frac{\bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i} \right)^2 \right)^{1/2} \right)^2 \quad (43)$$

$$\leq \frac{1}{4\delta_n^2} \left[\mathbb{E} \left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}} \right) \right]^{\frac{1}{1+\alpha_1}} \times \left(\left[\mathbb{E} \left[(\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n})^{2+2\alpha_2} \right] \right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E} \left[(\bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n})^{2+2\alpha_2} \right] \right]^{\frac{1}{1+\alpha_2}} \right) \quad (44)$$

$$\leq \frac{1}{4\delta_n^2} \left(\left[\mathbb{E} \left[(\bar{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n})^{2+2\alpha_2} \right] \right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E} \left[(\bar{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n})^{2+2\alpha_2} \right] \right]^{\frac{1}{1+\alpha_2}} \right) \quad (45)$$

$$\leq \frac{C}{\delta_n^2}, \text{ for some } C < \infty. \quad (46)$$

The inequality in (42) uses the fact that, for any random variable X , $\mathbb{E} \|X - E[X \mid \mathcal{F}_n]\|^2 \leq \mathbb{E} X^2$. The inequality in (43) follows by the fact that $\mathbb{E}(X + Y)^2 \leq ((\mathbb{E} X^2)^{1/2} + (\mathbb{E} Y^2)^{1/2})^2$. The inequality in (44) 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 (45) 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 (46) follows by using the fact that $\mathbb{C}(D^\theta)$ is bounded for any policy θ and the bias ϵ^θ is bounded by Proposition 2.

Thus, $\mathbb{E} \|\xi_n\|^2 \leq \frac{C}{\delta_n^2}$ for some $C < \infty$. Plugging this in (41), we obtain

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

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

- Observe that $\mathbb{C}(X^\theta)$ serves as a strict Lyapunov function for the ODE (39). This can be seen as follows:

$$\frac{d\mathbb{C}(X^\theta)}{dt} = \nabla \mathbb{C}(X^\theta) \dot{\theta} = \nabla \mathbb{C}(X^\theta) \tilde{\Gamma} (-\nabla \mathbb{C}(X^\theta)) < 0.$$

Hence, the set $\mathcal{K} = \{\theta \mid \tilde{\Gamma}_i (-\nabla \mathbb{C}(X^\theta)) = 0, \forall i = 1, \dots, d\}$ serves as the asymptotically stable attractor for the ODE (39).

The claim follows from the Kushner-Clark lemma. $\square$

E. Newton algorithm for CPT-value optimization (CPT-SPSA-N)

E.1. Need for second-order methods

While stochastic gradient descent methods are useful in minimizing the CPT-value given biased estimates, they are sensitive to the choice of the step-size sequence $\{\gamma_n\}$. In particular, for a step-size choice $\gamma_n = \gamma_0/n$, if a_0 is not chosen to be greater than $1/3\lambda_{\min}(\nabla^2 \mathbb{C}(X^{\theta^*}))$, then the optimum rate of convergence is not achieved, where $\lambda_{\min}$ denotes the minimum eigenvalue, while $\theta^* \in \mathcal{K}$ (see Theorem 1). A standard approach to overcome this step-size dependency is to use iterate averaging, suggested independently by Polyak (Polyak & Juditsky, 1992) and Ruppert (Ruppert, 1991). The idea is to use larger step-sizes $\gamma_n = 1/n^\varsigma$, where $\varsigma \in (1/2, 1)$, and then combine it with averaging of the iterates. However, it is well known that iterate averaging is optimal only in an asymptotic sense, while finite-time bounds show that the initial condition is not forgotten sub-exponentially fast (see Theorem 2.2 in (Fathi & Frikha, 2013)). 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 $(\nabla^2 \mathbb{C}(X^{\theta^*}))^{-1}$, i.e., the inverse of the Hessian of the CPT-value at the optimum θ^* . Such a scheme gets rid of the step-size dependency (one can set $a_0 = 1$) and still obtains optimal convergence rates. This is the motivation behind having a second-order optimization scheme.

E.2. Gradient and Hessian estimation

We estimate the Hessian of the CPT-value function using the scheme suggested by (Bhatnagar & Prashanth, 2015). As in the first-order method, we use Rademacher random variables to simultaneously perturb all the coordinates. However, in this case, we require three system trajectories with corresponding parameters $\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)$, $\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)$ and θ_n , where $\{\Delta_n^i, \hat{\Delta}_n^i, i = 1, \dots, d\}$ are i.i.d. Rademacher and independent of $\theta_0, \dots, \theta_n$. Using the CPT-value estimates for the aforementioned parameters, we estimate the Hessian and the gradient of the CPT-value function as follows: For $i, j = 1, \dots, d$, set

$$\begin{aligned} \hat{\nabla}_i \mathbb{C}(X_n^{\theta_n}) &= \frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}}{2\delta_n \Delta_n^i}, \\ \hat{H}_n^{i,j} &= \frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} + \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)} - 2\bar{\mathbb{C}}_n^{\theta_n}}{\delta_n^2 \Delta_n^i \hat{\Delta}_n^j}. \end{aligned}$$

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

Algorithm 3 Structure of CPT-SPSA-N algorithm.

Input: initial parameter $\theta_0 \in \Theta$ where Θ is a compact and convex subset of $\mathbb{R}^d$, perturbation constants $\delta_n > 0$, sample sizes $\{m_n\}$, step-sizes $\{\gamma_n, \xi_n\}$, operator $\Gamma : \mathbb{R}^d \rightarrow \Theta$.

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

Generate $\{\Delta_n^i, \hat{\Delta}_n^i, i = 1, \dots, d\}$ using Rademacher distribution, independent of $\{\Delta_m, \hat{\Delta}_m, m = 0, 1, \dots, n-1\}$.

CPT-value Estimation (Trajectory 1)

Simulate m_n samples using parameter $(\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n))$.

Obtain CPT-value estimate $\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)}$.

CPT-value Estimation (Trajectory 2)

Simulate m_n samples using parameter $(\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n))$.

Obtain CPT-value estimate $\bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}$.

CPT-value Estimation (Trajectory 3)

Simulate m_n samples using parameter θ_n .

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

Newton step

Update the parameter and Hessian according to (47)–(48).

end for

Return θ_n .

Notice that the above estimates require three samples, while the second-order SPSA algorithm proposed first in (Spall, 2000) required four. Both the gradient estimate $\hat{\nabla} \mathbb{C}(X_n^{\theta_n}) = [\hat{\nabla}_i \mathbb{C}(X_n^{\theta_n})], i = 1, \dots, d$, and the Hessian estimate $\hat{H}_n = [\hat{H}_n^{i,j}], i, j = 1, \dots, d$, can be shown to be an $O(\delta_n^2)$ term away from the true gradient $\nabla \mathbb{C}(X_n^\theta)$ and Hessian $\nabla^2 \mathbb{C}(X_n^\theta)$, respectively (see Lemmas 7–8).

E.3. Update rule

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

$$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i + \gamma_n \sum_{j=1}^d M_n^{i,j} \hat{\nabla}_j \mathbb{C}(X_n^\theta) \right), \quad (47)$$

$$\bar{H}_n = (1 - \xi_n) \bar{H}_{n-1} + \xi_n \hat{H}_n, \quad (48)$$

where ξ_n is a step-size sequence that satisfies $\sum_n \xi_n = \infty$, $\sum_n \xi_n^2 < \infty$ and $\frac{\gamma_n}{\xi_n} \rightarrow 0$ as $n \rightarrow \infty$. These conditions on ξ_n ensure that the updates to $\bar{H}_n$ proceed on a timescale that is faster than that of θ_n in (47) - see Chapter 6 of (Borkar, 2008). Further, Γ is a projection operator as in CPT-SPSA-G and $M_n = [M_n^{i,j}] = \Upsilon(\bar{H}_n)^{-1}$. Notice that we invert $\bar{H}_n$ in each iteration, and to ensure that this inversion is feasible (so that the θ -recursion descends), we project $\bar{H}_n$ onto the set of positive definite matrices using the operator Υ . The operator has to be such that asymptotically $\Upsilon(\bar{H}_n)$ should be the same as $\bar{H}_n$ (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 $\{B_n\}$, $\lim_{n \rightarrow \infty} \|A_n - B_n\| = 0 \Rightarrow \lim_{n \rightarrow \infty} \|\Upsilon(A_n) - \Upsilon(B_n)\| = 0$. Further, for any $\{C_n\}$ with $\sup_n \|C_n\| < \infty$, $\sup_n (\|\Upsilon(C_n)\| + \|\{\Upsilon(C_n)\}^{-1}\|) < \infty$.

A simple way to ensure the above is to have $\Upsilon(\cdot)$ as a diagonal matrix and then add a positive scalar δ_n to the diagonal elements so as to ensure invertibility - see (Gill et al., 1981), (Spall, 2000) for a similar operator.

Algorithm 3 presents the pseudocode.

E.4. Convergence result

Theorem 6. Assume (A1)-(A4). Consider the ODE:

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

where $\tilde{\Gamma}_i$ is as defined in Theorem 1. Let $\mathcal{K} = \{\theta \in \Theta \mid \nabla \mathbb{C}(X^{\theta^i}) \tilde{\Gamma}_i \left(-\Upsilon(\nabla^2 \mathbb{C}(X^{\theta}))^{-1} \nabla \mathbb{C}(X^{\theta^i}) \right) = 0, \forall i = 1, \dots, d\}$. Then, for θ_n governed by (47), we have

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

Proof. Before proving Theorem 6, we bound the bias in the SPSA based estimate of the Hessian in the following lemma.

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

$$\left| \mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} + \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)} - 2\bar{\mathbb{C}}_n^{\theta_n}}{\delta_n^2 \Delta_n^i \hat{\Delta}_n^j} \middle| \mathcal{F}_n \right] - \nabla_{i,j}^2 \mathbb{C}(X^{\theta_n}) \right| \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (49)$$

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

$$\begin{aligned} & \mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} + \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)} - 2\bar{\mathbb{C}}_n^{\theta_n}}{\delta_n^2 \Delta_n^i \hat{\Delta}_n^j} \middle| \mathcal{F}_n \right] \\ & \xrightarrow{n \rightarrow \infty} \mathbb{E} \left[\frac{\mathbb{C}(X^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)}) + \mathbb{C}(X^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}) - 2\mathbb{C}(X^{\theta_n})}{\delta_n^2 \Delta_n^i \hat{\Delta}_n^j} \middle| \mathcal{F}_n \right]. \end{aligned} \quad (50)$$

Now, the RHS of (50) approximates the true gradient with only an $O(\delta_n^2)$ error; this can be inferred using arguments similar to those used in the proof of Proposition 4.2 of (Bhatnagar & Prashanth, 2015). We provide the proof here for the sake of completeness. Using Taylor's expansion as in Lemma 5, we obtain

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

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

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

$$\left| \mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}}{2\delta_n \Delta_n^i} \middle| \mathcal{F}_n \right] - \nabla_i \mathbb{C}(X^{\theta_n}) \right| \rightarrow 0 \text{ as } n \rightarrow \infty. \quad (51)$$

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

$$\mathbb{E} \left[\frac{\bar{\mathbb{C}}_n^{\theta_n + \delta_n(\Delta_n + \hat{\Delta}_n)} - \bar{\mathbb{C}}_n^{\theta_n - \delta_n(\Delta_n + \hat{\Delta}_n)}}{2\delta_n \Delta_n^i} \middle| \mathcal{F}_n \right] \xrightarrow{n \rightarrow \infty} \mathbb{E} \left[\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta_n^i} \middle| \mathcal{F}_n \right].$$

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

Proof of Theorem 6

Before we prove Theorem 6, we show that the Hessian recursion (48) converges to the true Hessian, for any policy θ .

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

$$\|H_{n+1}^{i,j} - \nabla_{i,j}^2 \mathbb{C}(X^{\theta_n})\| \rightarrow 0, \text{ and } \|\Upsilon(\bar{H}_n)^{-1} - \Upsilon(\nabla_{i,j}^2 \mathbb{C}(X^{\theta_n}))^{-1}\| \rightarrow 0.$$

Proof. Follows in a similar manner as in the proofs of Lemmas 7.10 and 7.11 of (Bhatnagar et al., 2013). $\square$

Proof. (Theorem 6) The proof follows in a similar manner as the proof of Theorem 7.1 in (Bhatnagar et al., 2013); we provide a sketch below for the sake of completeness.

We first rewrite the recursion (47) as follows: For $i = 1, \dots, d$

$$\theta_{n+1}^i = \Gamma_i \left(\theta_n^i + \gamma_n \sum_{j=1}^d \bar{M}^{i,j}(\theta_n) \nabla_j \mathbb{C}(X_n^\theta) + \gamma_n \zeta_n + \chi_{n+1} - \chi_n \right), \quad (52)$$

where

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

In lieu of Lemmas 7–9, it is easy to conclude that $\zeta_n \rightarrow 0$ as $n \rightarrow \infty$, χ_n is a martingale difference sequence and that $\chi_{n+1} - \chi_n \rightarrow 0$ as $n \rightarrow \infty$. Thus, it is easy to see that (52) is a discretization of the ODE:

$$\dot{\theta}_t^i = \check{\Gamma}_i \left(-\nabla \mathbb{C}(X^{\theta_t^i}) \Upsilon(\nabla^2 \mathbb{C}(X^{\theta_t}))^{-1} \nabla \mathbb{C}(X^{\theta_t^i}) \right). \quad (53)$$

Since $\mathbb{C}(X^\theta)$ serves as a Lyapunov function for the ODE (53), it is easy to see that the set

$\mathcal{K} = \{\theta \mid \nabla \mathbb{C}(X^{\theta^i}) \check{\Gamma}_i \left(-\Upsilon(\nabla^2 \mathbb{C}(X^\theta))^{-1} \nabla \mathbb{C}(X^{\theta^i}) \right) = 0, \forall i = 1, \dots, d\}$ is an asymptotically stable attractor set for the ODE (53). The claim now follows from Kushner-Clark lemma. $\square$

$\square$