Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control

Cheng Jie[♯], Prashanth L A[†], Michael Fu[$], Steve Marcus[‡] and Csaba Szepesvári[⋆]

***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.**

***Index Terms*—Prospect theory, reinforcement learning, stochastic optimization, simultaneous perturbation stochastic approximation.**

I. 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 [1], 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 rewards or costs [2], [3], 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 [4], [5], [6]. An approach that gained strong support amongst psychologists, behavioral scientists and economists (cf. [7], [8]) is based on [6]'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 [9], which superseded prospect theory (e.g.,[10]). 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.

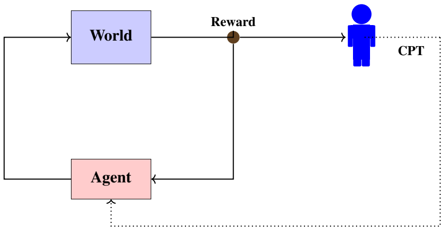


Fig. 1. Operational flow of a human-based decision making system

Our contributions[2]

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

[†] Institute for Systems Research, University of Maryland, College Park, Maryland, E-Mail: prashanth@isr.umd.edu,

[♯] Department of Computer Science and Automation, Indian Institute of Science, Bangalore, E-Mail: shalabh@csa.iisc.ernet.in,

[$] Robert H. Smith School of Business & Institute for Systems Research, University of Maryland, College Park, Maryland, E-Mail: mfu@isr.umd.edu,

[‡] Department of Electrical and Computer Engineering & Institute for Systems Research, University of Maryland, College Park, Maryland, E-Mail: marcus@umd.edu.

[⋆] Department of Computing Science, University of Alberta, E-Mail: szepesva@cs.ualberta.ca.

[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.

[2] A preliminary version of this paper, without the proofs, was published in ICML 2016 [11].

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 nonlinear transformation of the cumulative distribution functions associated with the random variable (cf. Section II 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* [12]. 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. In [13], the authors showed for a discounted MDP 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, in [14] the authors show that finding a globally mean-variance optimal policy in a discounted MDP is NP-hard. For average reward MDPs, in [15], the authors 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. [16], [17], [18]). 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.

In [19], the author proposes a CPT-measure for an abstract MDP setting. We differ from the aforementioned 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.

The rest of the paper is organized as follows: In Section II, we introduce the notion of CPT-value of a general random variable X. In Section III, we describe the empirical distribution based scheme for estimating the CPT-value of any random variable. In Section IV, we present the gradient-based algorithms for optimizing the CPT-value. We provide the proofs of convergence for all the proposed algorithms in Section V. We present the results from numerical experiments for the CPT-value estimation scheme in Section VI and finally, provide the concluding remarks in Section VII.

II. 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 $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ and $w^+, w^- : [0,1] \to [0,1]$ are continuous, with $u^+(x) = 0$ when $x \leq 0$ and $u^-(x) = 0$ when $x \geq 0$ (see assumptions (A1)-(A2) in Section III for precise requirements on u and w), is defined as

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

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 $(a)^+ = \max(a,0)$, $(a)^- = \max(-a,0)$, $\mathbb{C}(X) = \int_0^\infty \mathbb{P}(X > z)\, dz - \int_0^\infty \mathbb{P}(-X > z)\, dz = \mathbb{E}\left[(X)^+\right] - \mathbb{E}\left[(X)^-\right]$, showing the connection to expectations.

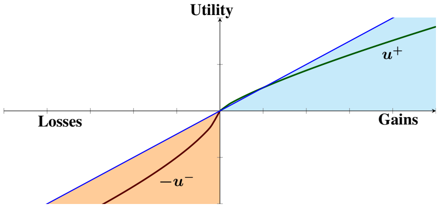


Fig. 2. An example of a utility function. A reference point on the x axis serves as the point of separating gains and losses. For losses, the disutility $-u^-$ is typically convex, for gains, the utility u^+ is typically concave; they are always non-decreasing and both of them take on the value of zero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

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.

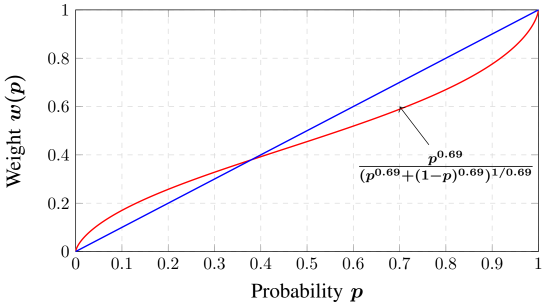


Fig. 3. An example of a weight function. A typical CPT weight function inflates small, and deflates large probabilities, capturing the tendency of humans doing the same when faces with decisions of uncertain outcomes.

The functions w^+, w^-, called the weight functions, capture the idea that humans deflate high-probabilities and inflate low-probabilities. For example, humans usually choose a stock that gives a large reward, e.g., one million dollars w.p. $1/10^6$ over one that gives \$1 w.p. 1 and the reverse when signs are flipped. Thus the value seen by a human subject is nonlinear in the underlying probabilities – an observation backed by strong empirical evidence [9], [10]. 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 $w^+(1) = w^-(1) = 1$. In [9], the authors recommend $w(p) = \frac{p^\eta}{(p^\eta+(1-p)^\eta)^{1/\eta}}$, while in [20], the author 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.*

III. Prediction of CPT-value

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., $\int_0^{+\infty} w^+ \left(\mathbb{P}\left(u^+(X) > z\right)\right) dz$ may diverge even if the first moment of random variable $u^+(X)$ is finite. For example, suppose U has the tail distribution function $\mathbb{P}(U > z) = \frac{1}{z^2}, z \in [1, +\infty)$, 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.

A. CPT-value prediction using quantiles

Let ξ_k^+ and ξ_k^- denote the kth quantiles of the r.v.s $u^+(X)$ and $u^-(X)$, respectively. Then, it can be seen that (see Proposition 6 in Section V-A1)

$$\lim_{n\to\infty} \sum_{i=1}^{n} \xi^+_{\frac{i}{n}} \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)$$
$$= \int_0^{+\infty} w^+ \left(\mathbb{P}\left(u^+(X) > z\right)\right) dz. \tag{2}$$

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

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

Algorithm 1 CPT-value prediction

```
Simulate $n$ i.i.d. samples from the distribution of $X$.
Order the samples and label them as follows: $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$. Note that $u^+(X_{[1]}), \ldots, u^+(X_{[n]})$ are also in ascending order.
Let
    $\overline{\mathbb{C}}_n^+ := \sum_{i=1}^{n} u^+(X_{[i]}) \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right).$
Apply $u^-$ on the sequence $\{X_{[1]}, X_{[2]}, \ldots, X_{[n]}\}$; notice that $u^-(X_{[i]})$ is in descending order since $u^-$ is a decreasing function.
Let
    $\overline{\mathbb{C}}_n^- := \sum_{i=1}^{n} u^-(X_{[i]}) \left(w^-\left(\frac{i}{n}\right) - w^-\left(\frac{i-1}{n}\right)\right).$
Return $\overline{\mathbb{C}}_n = \overline{\mathbb{C}}_n^+ - \overline{\mathbb{C}}_n^-$.
```

Notice the the following equivalence:

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

and also,

$$\sum_{i=1}^{n} u^{-}\left(X_{[i]}\right)\left(w^{-}\left(\frac{i}{n}\right)-w^{-}\left(\frac{i-1}{n}\right)\right)$$
$$=\int_{0}^{\infty} w^{-}\left(1-\hat{F}_{n}^{-}(x)\right) dx,$$

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

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

$$\overline{\mathbb{C}}_n=\int_{0}^{\infty} w^{+}\left(1-\hat{F}_{n}^{+}(x)\right) dx-\int_{0}^{\infty} w^{-}\left(1-\hat{F}_{n}^{-}(x)\right) dx. \tag{3}$$

Consider the special case when $w^{+}(p)=w^{-}(p)=p$ and u^{+} $(-u^{-})$, when restricted to the positive (respectively, negative) half line, are the identity functions. In this case, the CPT-value predictor $\overline{\mathbb{C}}_n$ coincides with the sample mean estimator for regular expectation.

B. Results for Hölder continuous weights

Recall the Hölder continuity property first:

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.*

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

Assumption (A1). The weight functions w^{+}, w^{-} are Hölder continuous with common order α. Further, there exists $\gamma \leq \alpha$ such that (s.t.) $\int_0^{+\infty}\mathbb{P}^{\gamma}(u^{+}(X)>z)dz<+\infty$ and $\int_0^{+\infty}\mathbb{P}^{\gamma}(u^{-}(X)>z)dz<+\infty$, where $\mathbb{P}^{\gamma}(\cdot)=(\mathbb{P}(\cdot))^{\gamma}$.

The above assumption ensures that the CPT-value as defined by (1) is finite - see Proposition 5 in Section V-A1 for a formal proof.

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

$$\overline{\mathbb{C}}_n \to \mathbb{C}(X) \text{ a.s. as } n\to\infty \tag{4}$$

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

Proof. See Section V-A1. □

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

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

Proposition 2. ***(Sample complexity.)*** *Assume (A1), (A2) and also that the utilities $u^{+}(X)$ and $u^{-}(X)$ are bounded above by $M<\infty$ w.p. 1. Then, $\forall \epsilon>0, \delta>0$, we have*

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

Corollary 1. *Under conditions of Proposition 2, we have*

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

where $\Gamma(\cdot)$ is the gamma function.

Proof. See Section V-A1. □

C. Results for Lipschitz continuous weights

In this section, we establish that the CPT-value predictor $\overline{\mathbb{C}}_n$ is asymptotically consistent when the weights are Lipschitz continuous, 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 obtain the asymptotic consistency claim in Proposition 1 for Lipschitz weight functions. However, this result is under a restrictive Lipschitz assumption on the distribution functions of $u^{+}(X)$ and $u^{-}(X)$. Using a different proof technique and (A1') in place of (A1), we can obtain a result similar to Proposition 1 without a Lipschitz assumption on the distribution functions. The following claim makes this precise.

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

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

Note that according to this proposition, our estimation scheme is sample-efficient (choosing the weights to be the identity function, the sample complexity cannot be improved).

Proof. See Section V-B. □

Remark 3. *For Hölder continuous weights, we incur a sample complexity of order $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ for accuracy $\epsilon>0$ and this is higher than the canonical Monte Carlo rate of $O\left(\frac{1}{\epsilon^{2}}\right)$. On the other hand, setting $\alpha=1$ in Proposition 2, we observe that one can achieve the canonical Monte Carlo rate for Lipschitz continuous weights.*

D. Locally Lipschitz weights and discrete-valued X

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

$$F_k=\sum_{i=1}^{k}p_k \text{ if } k\leq l \text{ and } \sum_{i=k}^{K}p_k \text{ if } k>l. \tag{5}$$

Then, the CPT-value is defined as

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

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

$$\hat{F}_k = \sum_{i=1}^{k} \hat{p}_k \text{ if } k \leq l \text{ and } \sum_{i=k}^{K} \hat{p}_k \text{ if } k > l. \tag{6}$$

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

$$\overline{\mathbb{C}}_n = u^-(x_1)w^-(\hat{p}_1) + \sum_{i=2}^{l} u^-(x_i)\Big(w^-(\hat{F}_i) - w^-(\hat{F}_{i-1})\Big) + \sum_{i=l+1}^{K-1} u^+(x_i)\Big(w^+(\hat{F}_i) - w^+(\hat{F}_{i+1})\Big) + u^+(x_K)w^+(\hat{p}_K).$$

Assumption (A1”). The weight functions $w^+(X)$ and $w^-(X)$ are locally Lipschitz continuous, i.e., for any x, there exist $L_x < \infty$ and $\rho > 0$, such that

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

Proposition 4. *Assume (A1”). Let $L = \max\{L_k, k = 2...K\}$, where L_k is the local Lipschitz constant of function $w^-(x)$ at points F_k, where $k = 1, \dots, l$, and of function $w^+(x)$ at points $k = l + 1, \dots, K$. Let $M = \max\{u^-(x_k), k = 1, \dots, l\} \bigcup \{u^+(x_k), k = l + 1, \dots, K\}$ 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*

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

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.

Proof. See Section V-C. □

A variant of Corollary 1 can be claimed by integrating the high-probability bound in Proposition 4 and we omit the details here.

IV. Control of CPT-value

A. 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

$$\text{Find } \theta^* = \arg\max_{\theta \in \Theta} \mathbb{C}(X^\theta), \tag{7}$$

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.

Algorithm 2 Structure of CPT-SPSA-G algorithm.

```
Input: initial parameter $\theta_0 \in \Theta$ where $\Theta$ is a compact and convex subset of $\mathbb{R}^d$, perturbation constants $\delta_n > 0$, sample sizes $\{m_n\}$, step-sizes $\{\gamma_n\}$, operator $\Pi : \mathbb{R}^d \to \Theta$.
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 $\overline{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n}$.
    CPT-value Estimation (Trajectory 2)
        Simulate $m_n$ samples using $(\theta_n - \delta_n \Delta_n)$.
        Obtain CPT-value estimate $\overline{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}$.
    Gradient Ascent
        Update $\theta_n$ using (9).
end for
Return $\theta_n$.
```

B. Gradient algorithm for CPT-value control (CPT-SPSA-G)

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 [21] 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 $i = 1, \dots, d$,

$$\widehat{\nabla}_i \mathbb{C}(X^\theta) = \frac{\overline{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n} - \overline{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}}{2\delta_n \Delta_n^i}, \tag{8}$$

where δ_n is a positive scalar that satisfies (A3) below, $\Delta_n = \left(\Delta_n^1, \dots, \Delta_n^d\right)^\mathsf{T}$, 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 $\overline{\mathbb{C}}_n^{\theta_n + \delta_n \Delta_n}$ (resp. $\overline{\mathbb{C}}_n^{\theta_n - \delta_n \Delta_n}$) denotes the CPT-value estimate that uses m_n samples of the r.v. $X^{\theta_n + \delta_n \Delta_n}$ (resp. $\overline{X}^{\theta_n - \delta_n \Delta_n}$). The (asymptotic) unbiasedness of the gradient estimate is proven in Lemma 3.

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

$$\theta_{n+1}^i = \Pi_i\left(\theta_n^i + \gamma_n \widehat{\nabla}_i \mathbb{C}(X^{\theta_n})\right), \tag{9}$$

where γ_n is a step-size chosen to satisfy (A3) below and $\Pi = (\Pi_1, \dots, \Pi_d)$ is an operator that ensures that the update (9) stays bounded within the 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 (9) as follows:

$$\theta^i_{n+1} = \Pi_i \bigg(\theta^i_n + \gamma_n \bigg(\frac{\mathbb{C}(X^{\theta_n + \delta_n \Delta_n}) - \mathbb{C}(X^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta^i_n} \bigg) + \underbrace{\frac{(\epsilon_n^{\theta_n + \delta_n \Delta_n} - \epsilon_n^{\theta_n - \delta_n \Delta_n})}{2\delta_n \Delta^i_n}}_{\kappa_n} \bigg).$$

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 [22]):

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

While Theorems 1–2 show that the bias ϵ^θ is bounded above, to establish convergence of the policy gradient recursion (9), 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

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

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

Assumption (A4). CPT-value $\mathbb{C}(X^\theta)$ is a continuously differentiable function of θ, for any $\theta \in \Theta$.

In a typical RL setting, a sufficient condition for ensuring (A4) holds is to assume that the policy is continuously differentiable in θ.

Convergence result:

Theorem 1. *Assume (A1)-(A4). Consider the ordinary differential equation (ODE):*

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

where $\check{\Pi}_i(f(\theta)) := \lim\limits_{\vartheta \downarrow 0} \frac{\Pi_i(\theta + \vartheta f(\theta)) - \theta}{\vartheta}$, *for any continuous* $f(\cdot)$. *Let* $\mathcal{K} = \{\theta^* \mid \check{\Pi}_i \left(\nabla_i \mathbb{C}(X^{\theta^*}) \right) = 0, \forall i = 1, \ldots, d\}$. *Then, for* θ_n *governed by* (9), *we have*

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

Proof. See Section V-D. □

C. Newton algorithm for CPT-value control (CPT-SPSA-N)

Need for second-order methods: While stochastic gradient methods are useful in maximizing the CPT-value given biased estimates, they are sensitive to the choice of the step-size sequence $\{\gamma_n\}$. In particular, for a step-size choice $\gamma_n = \gamma_0/n$, if a_0 is not chosen to be greater than $1/\left(3\lambda_{min}(\nabla^2 \mathbb{C}(X^{\theta^*}))\right)$, 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 [23] and Ruppert [24]. The idea is to use larger step-sizes $\gamma_n = 1/n^\varsigma$, where $\varsigma \in (1/2, 1)$, for the update iteration (9) and average the iterates in the end, i.e., $\bar{\theta}_{n+1} = \frac{1}{n} \sum_{m=1}^n \theta_m$. However, it is well known that iterate averaging is optimal only in an asymptotic sense, while finite-time bounds show that the initial condition is not forgotten sub-exponentially fast (see Theorem 2.2 in [25]). Thus, it is optimal to average iterates only after a sufficient number of iterations have passed, which implies that the iterates are already close to the optimum and the updates can be stopped.

An alternative approach is to employ step-sizes of the form $\gamma_n = (a_0/n) M_n$, where M_n converges to $\left(\nabla^2 \mathbb{C}(X^{\theta^*})\right)^{-1}$, i.e., the inverse of the Hessian of the CPT-value at the optimum θ^*. Such a scheme gets rid of the step-size dependency (one can set $a_0 = 1$) and still obtains optimal convergence rates. This is the motivation behind having a second-order optimization scheme.

a) Gradient and Hessian estimation: We estimate the Hessian of the CPT-value function using the scheme suggested by [26]. 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 $\theta_n + \delta_n(\Delta_n + \widehat{\Delta}_n)$, $\theta_n - \delta_n(\Delta_n + \widehat{\Delta}_n)$ and θ_n, where $\{\Delta^i_n, \widehat{\Delta}^i_n, i = 1, \ldots, d\}$ are i.i.d. Rademacher and independent of $\theta_0, \ldots, \theta_{n-1}$. Using the CPT-value estimates for the aforementioned parameters, we estimate the Hessian and the gradient of the CPT-value function as follows: For $i, j = 1, \ldots, d$, set

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

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

Notice that the above estimates require three samples, while the second-order SPSA algorithm proposed first in [27] required four. Both the gradient estimate $\widehat{\nabla} \mathbb{C}(X_n^{\theta_n}) = [\widehat{\nabla}_i \mathbb{C}(X_n^{\theta_n})], i = 1, \ldots, d$, and the Hessian estimate $\widehat{H}_n = [\widehat{H}^{i,j}_n], i, j = 1, \ldots, d$, can be shown to be an $O(\delta_n^2)$ term away from the true gradient $\nabla \mathbb{C}(X_n^\theta)$ and Hessian $\nabla^2 \mathbb{C}(X_n^\theta)$, respectively (see Lemmas 4–5).

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

$$\theta^i_{n+1} = \Pi_i \left(\theta^i_n + \gamma_n \sum_{j=1}^d M^{i,j}_n \widehat{\nabla}_j \mathbb{C}(X^\theta_n) \right), \tag{10}$$

$$\overline{H}_n = (1 - \xi_n) \overline{H}_{n-1} + \xi_n \widehat{H}_n, \tag{11}$$

where ξ_n is a step-size sequence that satisfies $\sum_n \xi_n = \infty, \sum_n \xi_n^2 < \infty$ and $\frac{\gamma_n}{\xi_n} \to 0$ as $n \to \infty$. These conditions on ξ_n ensure that the updates to $\overline{H}_n$ proceed on a timescale that is faster than that of θ_n in (10) - see Chapter 6 of [22]. Further, Π is a projection operator as in CPT-SPSA-G and $M_n = [M^{i,j}_n] = \Upsilon(\overline{H}_n)^{-1}$. Notice that we invert $\overline{H}_n$ in each iteration, and to ensure that this inversion is feasible (so that the θ-recursion descends), we project $\overline{H}_n$ onto the set of positive definite matrices using the operator Υ. The operator has to be such that asymptotically $\Upsilon(\overline{H}_n)$ should be the same as $\overline{H}_n$ (since the latter would converge to the true Hessian), while ensuring inversion is feasible in the initial iterations. The assumption below makes these requirements precise.

Assumption (A5). For any $\{A_n\}$ and $\{B_n\}$, $\lim\limits_{n\to\infty} \|A_n - B_n\| = 0 \Rightarrow \lim\limits_{n\to\infty} \parallel \Upsilon(A_n) - \Upsilon(B_n) \parallel = 0$. Further, for any $\{C_n\}$ with $\sup\limits_n \parallel C_n \parallel < \infty$, $\sup\limits_n \left(\parallel \Upsilon(C_n) \parallel + \parallel \{\Upsilon(C_n)\}^{-1} \parallel \right) < \infty$.

A simple way to 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 [28], [27] for a similar operator.

Convergence result:

Theorem 2. *Assume (A1)-(A5). Consider the ODE:*

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

where $\bar{\Pi}_i$ *is as defined in Theorem 1. Let* $\mathcal{K} = \{\theta \in \Theta \mid \nabla \mathbb{C}(X^{\theta^i}) \check{\Pi}_i \left(-\Upsilon(\nabla^2 \mathbb{C}(X^{\theta}))^{-1} \nabla \mathbb{C}(X^{\theta^i}) \right) = 0, \forall i = 1, \ldots, d\}$. *Then, for* θ_n *governed by* (10), *we have*

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

Proof. See Section V-E. □

V. Convergence Proofs

A. Proofs for CPT-value predictor

1) Hölder continuous weights: For proving Propositions 1 and 4, we require Hoeffding's inequality, which is given below.

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

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

Proposition 5. *Under (A1), the CPT-value* $\mathbb{C}(X)$ *as defined by* (1) *is finite.*

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

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

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

Proposition 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*

Cs: Fix the ugly parenthesis! Also, why

$$\lim_{n\to\infty} \sum_{i=1}^n \xi^+_{\frac{i}{n}} \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right) \right)$$
$$= \int_0^\infty w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right) dz < \infty, \tag{12}$$
$$\lim_{n\to\infty} \sum_{i=1}^n \xi^-_{\frac{i}{n}} \left(w^-\left(\frac{i}{n}\right) - w^-\left(\frac{i-1}{n}\right) \right)$$
$$= \int_0^\infty w^-\left(\mathbb{P}\left(u^-(X) > z\right)\right) dz < \infty. \tag{13}$$

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

Proof. We shall focus on proving the first part of equation (12). Notice that the following linear combination of simple functions:

$$\sum_{i=1}^n w^+\left(\frac{i}{n}\right) I_{\left[\xi^+_{\frac{n-i}{n}}, \xi^+_{\frac{n+1-i}{n}}\right]}(z) \tag{14}$$

will converge almost everywhere to the function $w^+\left(\mathbb{P}\left(u^+(X) > z\right)\right)$ in the interval $[0, \infty)$. Further, for all $z \in [0, \infty)$, we have

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

The integral of (14) can be simplified as follows:

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

The Hölder continuity property assures the fact that $\lim_{n\to\infty} |w^+(\frac{n-i}{n}) - w^+(\frac{n-i-1}{n})| = 0$, and the limit in (12) holds through an application of the dominated convergence theorem.

The second part of (12) can be justified in a similar fashion. □

Proof of Proposition 1

Proof. Without loss of generality, assume the Hölder constants H of the weight functions w^+ and w^- are both 1. We prove the claim for the first integral in the CPT-value predictor $\overline{\mathbb{C}}_n$ in Algorithm 1, i.e., we show that

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

$$\lim_{n\to\infty} \sum_{i=1}^n u^+\left(X_{[i]}\right) \left(w^+\left(\frac{n-i+1}{n}\right) - w^+\left(\frac{n-i}{n}\right) \right)$$
$$\xrightarrow{n\to\infty} \int_0^\infty w^+\left(P\left(u^+(X) > z\right)\right) dz, \text{ a.s.} \tag{15}$$

For any given $\epsilon > 0$, we have

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

$$
\begin{aligned}
&\quad -\sum_{i=1}^{n} \xi^+_{\frac{i}{n}} \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\Bigg| > \epsilon \Bigg) \\
&\le \mathbb{P}\Bigg(\bigcup_{i=1}^{n} \Bigg\{ \Bigg| u^+\left(X_{[i]}\right) \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right) \\
&\quad -\xi^+_{\frac{i}{n}} \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\Bigg| > \frac{\epsilon}{n}\Bigg\}\Bigg) \\
&\le \sum_{i=1}^{n} \mathbb{P}\Bigg(\Bigg| u^+\left(X_{[i]}\right) \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right) \\
&\quad -\xi^+_{\frac{i}{n}} \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\Bigg| > \frac{\epsilon}{n}\Bigg) \\
&= \sum_{i=1}^{n} \mathbb{P}\Bigg(\Bigg| \left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}} \right) \\
&\quad \times \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\Bigg| > \frac{\epsilon}{n}\Bigg) \\
&\le \sum_{i=1}^{n} \mathbb{P}\left(\left| \left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}} \right) \left(\frac{1}{n}\right)^{\alpha} \right| > \frac{\epsilon}{n}\right) \qquad (16) \\
&= \sum_{i=1}^{n} \mathbb{P}\left(\left| \left(u^+\left(X_{[i]}\right) - \xi^+_{\frac{i}{n}} \right) \right| > \frac{\epsilon}{\cdot n^{1-\alpha}}\right). \qquad (17)
\end{aligned}
$$

In the above, (16) follows from the fact that w^+ is Hölder with constant 1.

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

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

We focus on the first term above.

Let $W_t = I_{\left(u^+(X_t) > \xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)}, t = 1, \ldots, n.$

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

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

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

$$
\begin{aligned}
&\mathbb{P}\Bigg(\sum_{i=1}^{n} W_t - n \cdot \left[1 - F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right)\right] \\
&\qquad > n \cdot \left[F^+\left(\xi^+_{\frac{i}{n}} + \frac{\epsilon}{n^{(1-\alpha)}}\right) - \frac{i}{n}\right]\Bigg) \le e^{-2n\cdot\delta_t^{'}},
\end{aligned}
$$

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 $\delta_i^{'} \le L^+ \cdot \left(\frac{\epsilon}{n^{(1-\alpha)}}\right)$. Hence, we obtain

$$
\begin{aligned}
\mathbb{P}\left(u^+(X_{[i]}) - \xi^+_{\frac{i}{n}} > \frac{\epsilon}{n^{(1-\alpha)}}\right) &< e^{-2n\cdot L^+ \frac{\epsilon}{n^{(1-\alpha)}}} \\
&= e^{-2n^{\alpha}\cdot L^+ \epsilon} \qquad (18)
\end{aligned}
$$

In a similar fashion, one can show that

$$
\mathbb{P}\left(u^+(X_{[i]}) - \xi^+_{\frac{i}{n}} < -\frac{\epsilon}{n^{(1-\alpha)}}\right) \le e^{-2n^{\alpha}\cdot L^+ \epsilon}. \qquad (19)
$$

Combining (18) and (19), we obtain

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

Plugging the above in (17), we obtain

$$
\begin{aligned}
&\mathbb{P}\Bigg(\Bigg| \sum_{i=1}^{n} u^+\left(X_{[i]}\right) \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right) \\
&\quad -\sum_{i=1}^{n} \xi^+_{\frac{i}{n}} \cdot \left(w^+\left(\frac{n+1-i}{n}\right) - w^+\left(\frac{n-i}{n}\right)\right)\Bigg| > \epsilon \Bigg) \\
&\le 2n \cdot e^{-2n^{\alpha}\cdot L^+ \epsilon}. \qquad (20)
\end{aligned}
$$

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

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

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

which implies (15).

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 $\xi^-_{\frac{i}{n}}$. □

Proof of Proposition 2

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

Lemma 2. ***(Dvoretzky-Kiefer-Wolfowitz (DKW) inequality)*** *Let $\hat{F}_n(u) = \frac{1}{n}\sum_{i=1}^{n} I_{[(u(X_i)) \le u]}$ denote the empirical distribution of a r.v. $u(X)$, with $u(X_1), \ldots, u(X_n)$ being sampled from the r.v $u(X)$. The, for any n and $\epsilon > 0$, we have*

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

The reader is referred to Chapter 2 of [29] for a detailed description of empirical distributions in general and the 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

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

Now, plugging in the DKW inequality, we obtain

$$\begin{aligned}
&\mathbb{P}\left(\left|\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X)>t\right)\right)dt\right.\right.\\
&\qquad\left.\left.-\int_0^\infty w^+\left(1-\hat{F}_n^+(t)\right)dt\right|>\epsilon\right)\\
&\le\mathbb{P}\left(HM\sup_{t\in\mathbb{R}}\left|(\mathbb{P}\left(u^+(X)<t\right)-\hat{F}_n^+(t)\right|^\alpha>\epsilon\right)\\
&\le e^{-n\frac{\epsilon^{(2/\alpha)}}{2H^2M^2}}. \qquad (21)
\end{aligned}$$

The claim follows. □

Proof of Corollary 1

Proof. Integrating the high-probability bound in Proposition 2, we obtain

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

□

B. Lipschitz continuous weights

Ch: Can't we have a unified proof of Hölder and Lipschitz cases?

Setting $\alpha=\gamma=1$ in the proof of Proposition 3, it is easy to see that the CPT-value (1) 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 3. ***(Generalized Dominated Convergence theorem)*** *Let $\{f_n\}_{n=1}^\infty$ be a sequence of measurable functions on a measurable space E that converge pointwise a.e. on a 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 $|f_n|\le g_n$ for all $n\in\mathbb{N}$. If $\lim\limits_{n\to\infty}\int_E g_n=\int_E g$, then $\lim\limits_{n\to\infty}\int_E f_n=\int_E f$.*

Proof. This is a standard result that can be found in any textbook on measure theory. For instance, see Theorem 2.3.11 in [30]. □

Proof of Proposition 3: Asymptotic convergence

Proof. We first prove the asymptotic convergence claim for the first integral (3) in the CPT-value predictor in Algorithm 1, i.e., we show

$$\int_0^\infty w^+\left(1-\hat{F}_n^+(x)\right)dx\to\int_0^\infty w^+\left(\mathbb{P}\left(u^+(X)>x\right)\right)dx. \qquad (22)$$

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

We have

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

and

$$\int_0^\infty\left(1-\hat{F}_n^+(x)\right)dx=\int_0^\infty\int_x^\infty d\hat{F}_n(t)\,dx. \qquad (23)$$

Since $\hat{F}_n^+(x)$ has bounded support on $\mathbb{R}$ $\forall n$, the integral in (23) is finite. Applying Fubini's theorem to the RHS of (23), we obtain

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

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

Notice that

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

From the foregoing,

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

The claim in (22) 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 $L\cdot(1-\hat{F}_n(x))\xrightarrow{a.s.}L(\mathbb{P}\left(u^+(X)>x\right))$ uniformly $\forall x$. The latter fact is implied by the Glivenko-Cantelli theorem (cf. Chapter 2 of [29]).

Following similar arguments, it is easy to show that

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

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

C. Proofs for discrete valued X

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

Proposition 7. *Let F_k and $\hat{F}_k$ be as defined in* (5), (6)*, Then, for every $\epsilon > 0$,*

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

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

$$\begin{aligned} &P(\left|\hat{F}_k - F_k\right| > \epsilon) \\ &= P(\left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \geq x_k\}} - \frac{1}{n}\sum_{i=1}^{n} E(I_{\{X_i \geq x_k\}})\right| > \epsilon) \\ &= P(\left|\sum_{i=1}^{n} I_{\{X_i \geq x_k\}} - \sum_{i=1}^{n} E(I_{\{X_i \geq x_k\}})\right| > n\epsilon) \qquad (24) \\ &\leq 2e^{-2n\epsilon^2}, \qquad (25) \end{aligned}$$

where the last inequality above follows by an application of Hoeffding inequality after observing that X_i are independent of each other and for each i, the corresponding r.v. in (24) is an indicator that is trivially bounded above by 1. □

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, we have*

$$\begin{aligned} &\mathbb{P}\left(\left|\sum_{i=1}^{K} u_k w(\hat{F}_k) - \sum_{i=1}^{K} u_k w(F_k)\right| > \epsilon\right) \\ &\leq K\left(e^{-2n\rho^2} + e^{-2n\epsilon^2/(KLM)^2}\right), \text{ where} \\ &u_k = u^-(x_k) \text{ if } k \leq l \text{ and } u^+(x_k) \text{ if } k > l. \qquad (26) \end{aligned}$$

Proof. Observe that

$$\begin{aligned} &\mathbb{P}\left(\left|\sum_{k=1}^{K} u_k w(\hat{F}_k) - \sum_{k=1}^{K} u_k w(F_k)\right| > \epsilon\right) \\ &= \mathbb{P}\left(\bigcup_{k=1}^{K} \left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right) \\ &\leq \sum_{k=1}^{K} \mathbb{P}\left(\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right) \qquad (27) \end{aligned}$$

For each $k = 1, ...K$, the function w is locally Lipschitz on $[p_k - \rho, p_k + \rho)$ with common constant L. Therefore, for each k, we can decompose the corresponding probability in (27) as follows:

$$\begin{aligned} &\mathbb{P}\left(\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right) \\ &= \mathbb{P}\left(\left\{\left|F_k - \hat{F}_k\right| > \rho\right\} \bigcap \left\{\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right\}\right) \\ &+ \mathbb{P}\left(\left\{\left|F_k - \hat{F}_k\right| \leq \rho\right\} \bigcap \left\{\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right\}\right) \\ &\leq \mathbb{P}\left(\left|F_k - \hat{F}_k\right| > \rho\right) \\ &+ \mathbb{P}\left(\left\{\left|F_k - \hat{F}_k\right| \leq \rho\right\} \bigcap \left\{\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right\}\right). \\ &\qquad (28) \end{aligned}$$

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

$$\begin{aligned} &\mathbb{P}\left(\left\{\left|F_k - \hat{F}_k\right| \leq \rho\right\} \bigcap \left\{\left|u_k w(\hat{F}_k) - u_k w(F_k)\right| > \frac{\epsilon}{K}\right\}\right) \\ &\leq \mathbb{P}\left(u_k L \left|F_k - \hat{F}_k\right| > \frac{\epsilon}{K}\right) \\ &\leq e^{-2n\epsilon/(KLu_k)^2} \leq e^{-2n\epsilon/(KLM)^2}, \forall k. \qquad (29) \end{aligned}$$

Using Proposition 7, we obtain

$$\mathbb{P}\left(\left|F_k - \hat{F}_k\right| > \rho\right) \leq e^{-2n\rho^2}, \forall k. \qquad (30)$$

Using (29) and (30) in (28), we obtain

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

The claim follows. □

Proof of Proposition 4

Proof. With u_k as defined in (26), we need to prove that, $\forall n \geq \frac{1}{\kappa}\ln\left(\frac{1}{\delta}\right)\ln\left(\frac{4K}{M}\right)$, the following high-probability bound holds

$$\begin{aligned} &\mathbb{P}\left(\left|\sum_{i=1}^{K} u_k \left(w\left(\hat{F}_k\right) - w\left(\hat{F}_{k+1}\right)\right)\right.\right. \\ &\left.\left. - \sum_{i=1}^{K} u_k \left(w\left(F_k\right) - w\left(F_{k+1}\right)\right)\right| \leq \epsilon\right) > 1 - \delta. \qquad (31) \end{aligned}$$

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

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

Hence,

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

The claim in (31) 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 [31]) 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 [32].

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

Lemma 3. *Let $\mathcal{F}_n = \sigma(\theta_m, m \le n)$, $n \ge 1$. Then, for any $i = 1, \ldots, d$, we have almost surely,*

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

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.

Notice that

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] \tag{32}$$

$$= \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] + \mathbb{E}\left[\eta_n \mid \mathcal{F}_n\right], \tag{33}$$

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

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

Thus,

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] \xrightarrow{n\to\infty} \mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right]. \tag{34}$$

We now analyse the RHS of (34). By using suitable Taylor's expansions,

$$\mathbb{C}(X^{\theta_n\pm\delta_n\Delta_n}) = \mathbb{C}(X^{\theta_n}) \pm \delta_n \Delta_n^{\mathsf{T}} \nabla \mathbb{C}(X^{\theta_n}) + \frac{\delta^2}{2} \Delta_n^{\mathsf{T}} \nabla^2 \mathbb{C}(X^{\theta_n}) \Delta_n + O(\delta_n^3).$$

From the above, it is easy to see that

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

Taking conditional expectation on both sides, we obtain

$$\mathbb{E}\left[\frac{\mathbb{C}(X^{\theta_n+\delta_n\Delta_n}) - \mathbb{C}(X^{\theta_n-\delta_n\Delta_n})}{2\delta_n\Delta_n^i} \mid \mathcal{F}_n \right] = \nabla_i \mathbb{C}(X^{\theta_n}) + \mathbb{E}\left[\sum_{j=1, j\ne i}^{N} \frac{\Delta_n^j}{\Delta_n^i} \right] \nabla_j \mathbb{C}(X^{\theta_n}) + O(\delta_n^2) = \nabla_i \mathbb{C}(X^{\theta_n}) + O(\delta_n^2). \tag{35}$$

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

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

Proof of Theorem 1

Proof. We first rewrite the update rule (9) as follows: For $i = 1, \ldots, d$,

$$\theta_{n+1}^i = \Pi_i\left(\theta_n^i + \gamma_n (\nabla_i \mathbb{C}(X^{\theta_n}) + \beta_n + \xi_n) \right), \tag{36}$$

where

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

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

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

Convergence of (36) can be inferred from Theorem 5.3.1 on pp. 191-196 of [32], 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 $\beta_n, n \ge 0$ is a bounded random sequence with $\beta_n \to 0$ almost surely as $n \to \infty$.

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

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

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

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

$$\dot{\theta}_t^i = \check{\Pi}_i\left(-\nabla \mathbb{C}(X^{\theta_t^i}) \right), \text{ for } i = 1, \ldots, d, \tag{37}$$

In the following, we verify the above assumptions for the recursion (9):

- (B1) holds by assumption in our setting.
- Lemma 3 above establishes that the bias β_n is $O(\delta_n^2)$ and since $\delta_n \to 0$ as $n \to \infty$, it is easy to see that (B2) is satisfied for β_n.

- (B3) holds by assumption (A3).
- We verify (B4) using arguments similar to those used in [31] for the classic SPSA algorithm:
  We first recall Doob's martingale inequality (see (2.1.7) on pp. 27 of [32]):

$$\mathbb{P}\left(\sup_{l\geq 0}\left\|W_l\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\lim_{l\to\infty}\mathbb{E}\left\|W_l\right\|^2. \tag{38}$$

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

$$\mathbb{P}\left(\sup_{l\geq k}\left\|\sum_{n=k}^{l}\gamma_n\xi_n\right\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2}\mathbb{E}\left\|\sum_{n=k}^{\infty}\gamma_n\xi_n\right\|^2$$
$$= \frac{1}{\epsilon^2}\sum_{n=k}^{\infty}\gamma_n^2\mathbb{E}\left\|\eta_n\right\|^2. \tag{39}$$

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

$$\mathbb{E}\left\|\xi_n\right\|^2 \leq \mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n} - \overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2 \tag{40}$$

$$\leq\left(\left(\mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2} + \left(\mathbb{E}\left(\frac{\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n}}{2\delta_n\Delta_n^i}\right)^2\right)^{1/2}\right)^2 \tag{41}$$

$$\leq\frac{1}{4\delta_n^2}\left[\mathbb{E}\left(\frac{1}{(\Delta_n^i)^{2+2\alpha_1}}\right)\right]^{\frac{1}{1+\alpha 1}} \times\left(\left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) \tag{42}$$

$$\leq\frac{1}{4\delta_n^2}\left(\left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n+\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}} + \left[\mathbb{E}\left[(\overline{\mathbb{C}}_n^{\theta_n-\delta_n\Delta_n})\right]^{2+2\alpha_2}\right]^{\frac{1}{1+\alpha_2}}\right) \tag{43}$$

$$\leq\frac{C}{\delta_n^2},\ \text{for some } C < \infty. \tag{44}$$

The inequality in (40) uses the fact that, for any random variable X, $\mathbb{E}\left\|X - E[X \mid \mathcal{F}_n]\right\|^2 \leq \mathbb{E}X^2$. The inequality in (41) follows by the fact that $\mathbb{E}(X+Y)^2 \leq \left((\mathbb{E}X^2)^{1/2} + (\mathbb{E}Y^2)^{1/2}\right)^2$. The inequality in (42) uses Hölder inequality, with $\alpha_1, \alpha_2 > 0$ satisfying $\frac{1}{1+\alpha_1} + \frac{1}{1+\alpha_2} = 1$. The equality in (43) 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 (44) follows by using the fact that $\mathbb{C}(X^\theta)$ is bounded for any parameter θ and the bias ϵ^θ is bounded by Corollary 1.

P: Need to update the above arguments

Thus, $\mathbb{E}\left\|\xi_n\right\|^2 \leq \frac{C}{\delta_n^2}$ for some $C < \infty$. Plugging this in (39), we obtain

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

The equality above follows from (A3) in the main paper.
- Observe that $\mathbb{C}(X^\theta)$ serves as a strict Lyapunov function for the ODE (37). This can be seen as follows:

$$\frac{d\mathbb{C}(X^\theta)}{dt} = \nabla\mathbb{C}(X^\theta)\dot{\theta} = \nabla\mathbb{C}(X^\theta)\check{\Pi}\left(-\nabla\mathbb{C}(X^\theta)\right) < 0.$$

Hence, the set $\mathcal{K} = \{\theta \mid \check{\Pi}_i\left(-\nabla\mathbb{C}(X^\theta)\right) = 0, \forall i = 1,\ldots,d\}$ serves as the asymptotically stable attractor for the ODE (37).

The claim follows from the Kushner-Clark lemma. □

E. Proofs for CPT-SPSA-N

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

Before proving Theorem 2, we bound the bias in the SPSA based estimate of the Hessian in the following lemma.

Lemma 4. *For any $i, j = 1,\ldots,d$, we have almost surely,*

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

Proof. As in the proof of Lemma 3, we can ignore the bias from the CPT-value estimation scheme and conclude that

$$\mathbb{E}\left[\frac{\overline{\mathbb{C}}_n^{\theta_n+\delta_n(\Delta_n+\widehat{\Delta}_n)} + \overline{\mathbb{C}}_n^{\theta_n-\delta_n(\Delta_n+\widehat{\Delta}_n)} - 2\overline{\mathbb{C}}_n^{\theta_n}}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right]$$
$$\xrightarrow{n\to\infty}\mathbb{E}\left[\frac{\mathbb{C}(X^+) + \mathbb{C}(X^-) - 2\mathbb{C}(X^{\theta_n})}{\delta_n^2\Delta_n^i\widehat{\Delta}_n^j} \mid \mathcal{F}_n\right]. \tag{45}$$

Now, the RHS of (45) 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 [26]. We provide the proof here for the sake of completeness. Using Taylor's expansion as in Lemma 3, we obtain

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

Taking conditional expectation, we observe that the first and last term above become zero, while the second term becomes

$\nabla^2_{ij}\mathbb{C}(X^{\theta_n})$. The claim follows by using the fact that $\delta_n \to 0$ as $n \to \infty$. □

Lemma 5. *For any $i = 1, \ldots, d$, we have almost surely,*

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

Proof. Follows by using completely parallel arguments to that in Lemma 3. □

The following lemma establishes that the Hessian recursion (11) converges to the true Hessian, for any policy θ.

Lemma 6. *For any $i, j = 1, \ldots, d$, we have almost surely,*

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

Proof. Follows in a similar manner as in the proofs of Lemmas 7.10 and 7.11 of [21]. □

*Proof. **(Theorem 2)*** The proof follows in a similar manner as the proof of Theorem 7.1 in [21]; we provide a sketch below for the sake of completeness.

We first rewrite the recursion (10) as follows: For $i = 1, \ldots, d$

$$\theta_{n+1}^i = \Pi_i\Bigg(\theta_n^i + \gamma_n \sum_{j=1}^{d} \bar{M}^{i,j}(\theta_n)\nabla_j \mathbb{C}(X_n^{\theta}) + \gamma_n \zeta_n + \chi_{n+1} - \chi_n\Bigg), \tag{46}$$

where

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

In lieu of Lemmas 4–6, it is easy to conclude that $\zeta_n \to 0$ as $n \to \infty$, χ_n is a martingale difference sequence and that $\chi_{n+1} - \chi_n \to 0$ as $n \to \infty$. Thus, it is easy to see that (46) is a discretization of the ODE:

$$\dot{\theta}_t^i = \check{\Pi}_i\left(-\nabla\mathbb{C}(X^{\theta_t^i})\Upsilon(\nabla^2\mathbb{C}(X^{\theta_t}))^{-1}\nabla\mathbb{C}(X^{\theta_t^i})\right). \tag{47}$$

Since $\mathbb{C}(X^{\theta})$ serves as a Lyapunov function for the ODE (47), it is easy to see that the set $\mathcal{K} = \{\theta \mid \nabla\mathbb{C}(X^{\theta^i})\check{\Pi}_i\left(-\Upsilon(\nabla^2\mathbb{C}(X^{\theta}))^{-1}\nabla\mathbb{C}(X^{\theta^i})\right) = 0, \forall i = 1, \ldots, d\}$ is an asymptotically stable attractor set for the ODE (47). The claim now follows from Kushner-Clark lemma. □

VI. 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 [33] 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 $i = 1, \ldots, \mathcal{N}$. Let $d_n^{i,j}$ denote the delay experienced by jth road user on ith path, for any $i = 1, \ldots, \mathcal{M}$ and $j = 1, \ldots, n_i$, where n_i denotes the number of road users on path i. We specify the various components of the traffic control MDP below. The state $s_n = (q_n^1, \ldots, q_n^{\mathcal{N}}, t_n^1, \ldots, t_n^{\mathcal{N}}, d_n^{1,1}, \ldots, d_n^{\mathcal{M},n_{\mathcal{M}}})^{\mathsf{T}}$ 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 $i = 1, \ldots, \mathcal{M}$. 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

$$\text{CPT}(X_1, \ldots, X_{\mathcal{M}}) = \sum_{i=1}^{\mathcal{M}} \mu^i \mathbb{C}(X_i). \tag{48}$$

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

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

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

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

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

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 [34]) 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 (48) is to be understood as a *differential delay*.

Using a Boltzmann policy that has the form

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

with features $\phi_{s,a}$ as described in Section V-B of [35], 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:

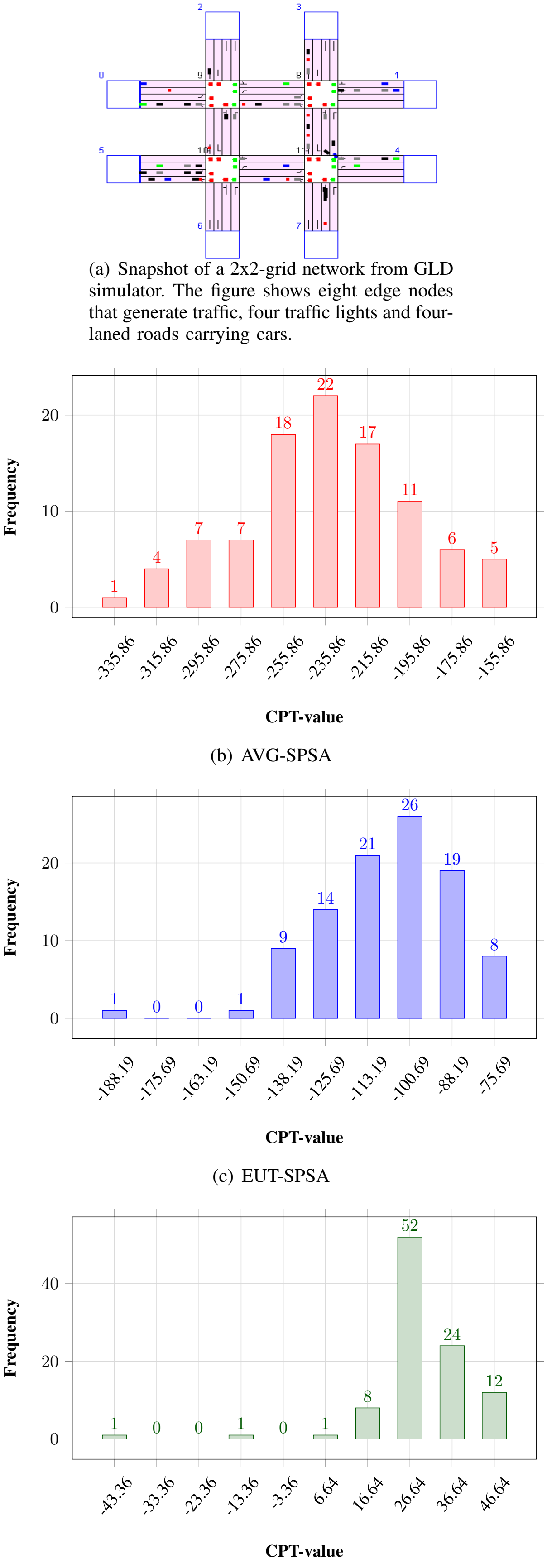
$$w^+(p) = \frac{p^{\eta_1}}{\left(p^{\eta_1} + (1-p)^{\eta_1}\right)^{\frac{1}{\eta_1}}}, \text{ and}$$

$$w^-(p) = \frac{p^{\eta_2}}{\left(p^{\eta_2} + (1-p)^{\eta_2}\right)^{\frac{1}{\eta_2}}},$$

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 [9] and have been used earlier in a traffic application (see [36]). For all the algorithms, motivated by standard guidelines (see [37]), 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 (48) 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.



(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

Fig. 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 utilites and weights. Note: larger values are better.

VII. Conclusions

CPT has been a very popular paradigm for modeling human decisions among psychologists/economists, but has escaped the radar of the reinforcement learning 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.

References

[1] H. A. Simon, "Theories of decision-making in economics and behavioral science," *The American Economic Review*, vol. 49, pp. 253–283, 1959.

[2] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton: Princeton University Press, 1944.

[3] P. Fishburn, *Utility Theory for Decision Making*. Wiley, New York, 1970.

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

[5] D. Ellsberg, "Risk, ambiguity and the Savage's axioms," *The Quarterly Journal of Economics*, vol. 75, no. 4, pp. 643–669, 1961.

[6] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica: Journal of the Econometric Society*, pp. 263–291, 1979.

[7] C. Starmer, "Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk," *Journal of economic literature*, pp. 332–382, 2000.

[8] J. Quiggin, *Generalized Expected Utility Theory: The Rank-dependent Model*. Springer Science & Business Media, 2012.

[9] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[10] N. C. Barberis, "Thirty years of prospect theory in economics: A review and assessment," *Journal of Economic Perspectives*, vol. 27, no. 1, pp. 173–196, 2013.

[11] L. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvári, "Cumulative prospect theory meets reinforcement learning: Prediction and control," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 1406–1415.

[12] M. C. Fu, Ed., *Handbook of Simulation Optimization*. Springer, 2015.

[13] M. Sobel, "The variance of discounted Markov decision processes," *Journal of Applied Probability*, pp. 794–802, 1982.

[14] S. Mannor and J. N. Tsitsiklis, "Algorithmic aspects of mean–variance optimization in Markov decision processes," *European Journal of Operational Research*, vol. 231, no. 3, pp. 645–653, 2013.

[15] J. Filar, L. Kallenberg, and H. Lee, "Variance-penalized Markov decision processes," *Mathematics of Operations Research*, vol. 14, no. 1, pp. 147–161, 1989.

[16] V. Borkar and R. Jain, "Risk-constrained Markov decision processes," in *IEEE Conference on Decision and Control (CDC)*, 2010, pp. 2664–2669.

[17] L. A. Prashanth, "Policy Gradients for CVaR-Constrained MDPs," in *Algorithmic Learning Theory*. Springer International Publishing, 2014, pp. 155–169.

[18] A. Tamar, Y. Glassner, and S. Mannor, "Optimizing the CVaR via sampling," *arXiv preprint arXiv:1404.3862*, 2014.

[19] K. Lin, "Stochastic Systems with Cumulative Prospect Theory," Ph.D. Thesis, University of Maryland, College Park, 2013.

[20] D. Prelec, "The probability weighting function," *Econometrica*, pp. 497–527, 1998.

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

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

[23] B. T. Polyak and A. B. Juditsky, "Acceleration of stochastic approximation by averaging," *SIAM Journal on Control and Optimization*, vol. 30, no. 4, pp. 838–855, 1992.

[24] D. Ruppert, "Stochastic approximation," *Handbook of Sequential Analysis*, pp. 503–529, 1991.

[25] M. Fathi and N. Frikha, "Transport-entropy inequalities and deviation estimates for stochastic approximation schemes," *Electronic Journal of Probability*, vol. 18, no. 67, pp. 1–36, 2013.

[26] S. Bhatnagar and L. A. Prashanth, "Simultaneous perturbation Newton algorithms for simulation optimization," *Journal of Optimization Theory and Applications*, vol. 164, no. 2, pp. 621–643, 2015.

[27] J. C. Spall, "Adaptive stochastic approximation by the simultaneous perturbation method," *IEEE Trans. Autom. Contr.*, vol. 45, pp. 1839–1853, 2000.

[28] P. Gill, W. Murray, and M. Wright, *Practical Optimization*. Academic Press, 1981.

[29] L. A. Wasserman, *All of Nonparametric Statistics*. Springer, 2015.

[30] K. B. Athreya and S. N. Lahiri, *Measure Theory and Probability Theory*. Springer Science & Business Media, 2006.

[31] J. C. Spall, "Multivariate stochastic approximation using a simultaneous perturbation gradient approximation," *IEEE Trans. Auto. Cont.*, vol. 37, no. 3, pp. 332–341, 1992.

[32] H. Kushner and D. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.

[33] M. Wiering, J. Vreeken, J. van Veenen, and A. Koopman, "Simulation and optimization of traffic in a city," in *IEEE Intelligent Vehicles Symposium*, June 2004, pp. 453–458.

[34] L. Prashanth and S. Bhatnagar, "Reinforcement Learning With Function Approximation for Traffic Signal Control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 412 –421, 2011.

[35] ——, "Threshold Tuning Using Stochastic Optimization for Graded Signal Control," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 9, pp. 3865 –3880, 2012.

[36] S. Gao, E. Frejinger, and M. Ben-Akiva, "Adaptive route choices in risky traffic networks: A prospect theory approach," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 5, pp. 727–740, 2010.

[37] J. C. Spall, *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley & Sons, 2005, vol. 65.