Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control

Prashanth L.A. PRASHLA@ISR.UMD.EDU
Institute for Systems Research, University of Maryland

Cheng Jie CJIE@MATH.UMD.EDU
Department of Mathematics, University of Maryland

Michael Fu MFU@ISR.UMD.EDU
Robert H. Smith School of Business & Institute for Systems Research, University of Maryland

Steve Marcus MARCUS@UMD.EDU
Department of Electrical and Computer Engineering & Institute for Systems Research, University of Maryland

Csaba Szepesvári SZEPESVA@CS.UALBERTA.CA
Department of Computing Science, University of Alberta

Abstract

Cumulative prospect theory (CPT) is known to model human decisions well, with substantial empirical evidence supporting this claim. CPT works by distorting probabilities and is more general than the classic expected utility and coherent risk measures. We bring this idea to a risk-sensitive reinforcement learning (RL) setting and design algorithms for both estimation and control. The RL setting presents two particular challenges when CPT is applied: estimating the CPT objective requires estimations of the *entire distribution* of the value function and finding a *randomized* optimal policy. The estimation scheme that we propose uses the empirical distribution to estimate the CPT-value of a random variable. We then use this scheme in the inner loop of 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.

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

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

[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.

transformation, such as the exponential function, of the rewards 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.

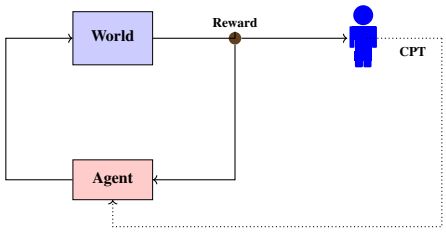


Figure 1. Operational flow of a human-based decision making system

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 nonlinear 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.

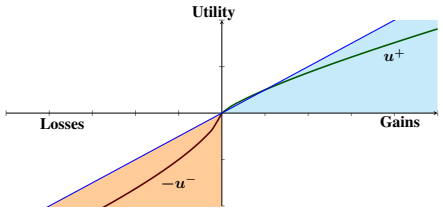


Figure 2. An example of a utility function. A reference point on the x axis serves as the point of separating gains and losses. For losses, the disutility $-u^-$ is typically convex, for gains, the utility u^+ is typically concave; they are always non-decreasing and both of them take on the value of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

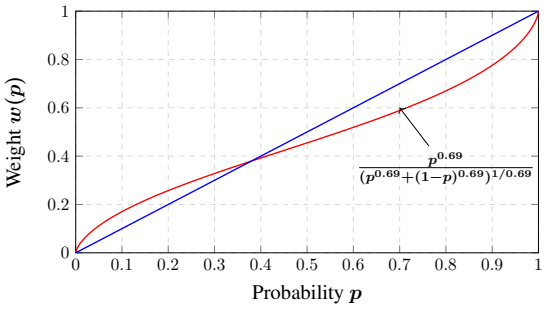


Figure 3. An example of a weight function. A typical CPT weight function inflates small, and deflates large probabilities, capturing the tendency of humans doing the same when faces with decisions of uncertain outcomes.

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 $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$ 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 3 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 $(a)^+ = \max(a, 0)$, $(a)^- = \max(-a, 0)$, $\mathbb{C}(X) = \int_0^{+\infty} \mathbb{P}(X > z)\, dz - \int_0^{+\infty} \mathbb{P}(-X > z)\, dz = \mathbb{E}[(X)^+] - \mathbb{E}[(X)^-]$, 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 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., $\int_0^{+\infty} w^+ \left(\mathbb{P}\left(u^+(X) > z\right)\right) dz$ 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.

3.1. Estimation scheme 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.*

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

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.) $\int_0^{+\infty} \mathbb{P}^\gamma(u^+(X) > z)dz < +\infty$ 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 (Prashanth et al., 2015) for a formal proof.

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

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

A similar claim holds with $u^-(X)$, ξ^-_α, w^- in place of $u^+(X)$, ξ^+_α, w^+, respectively. Here ξ^-_α denotes the αth 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: $X_{[1]}, X_{[2]}, \ldots, X_{[n]}$. Note that $u^+(X_{[1]}), \ldots, u^+(X_{[n]})$ are also in ascending order.
3: Let

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

4: 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.
5: Let

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

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

MAIN RESULTS

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*

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

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

Under additional assumptions on 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*

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

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 $O\left(\frac{1}{\epsilon^{2/\alpha}}\right)$ and this is higher than the canonical Monte Carlo rate of $O\left(\frac{1}{\epsilon^2}\right)$. 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 Dvoretzky-Kiefer-Wolfowitz (DKW) inequality, 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*

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

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

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

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).

3.2. Estimation scheme for locally Lipschitz weights and discrete X

Here we assume that the r.v. X is discrete valued. Let $p_i, i = 1, \ldots, K$, 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

$$F_k = \sum_{i=1}^{k} p_k \text{ if } k \leq l \text{ and } \sum_{i=k}^{K} p_k \text{ if } k > l. \tag{4}$$

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

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

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

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

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

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 $i = 1, \ldots, d$,

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

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

4.2. Update rule

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

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

where γ_n is a step-size chosen to satisfy (A3) below and $\Gamma = (\Gamma_1, \ldots, \Gamma_d)$ is an operator that ensures that the update (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:

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

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 $\Gamma : \mathbb{R}^d \to \Theta$.
for $n = 0, 1, 2, \ldots$ **do**
  Generate $\{\Delta_n^i, i = 1, \ldots, d\}$ using Rademacher distribution, independent of $\{\Delta_m, m = 0, 1, \ldots, n-1\}$.
  CPT-value Estimation (Trajectory 1)
    Simulate m_n samples using $(\theta_n + \delta_n\Delta_n)$.
    Obtain CPT-value 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 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)):

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

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

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

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

4.3. Convergence result[2]

Theorem 1. *Assume (A1)-(A3) and also that $\mathbb{C}(X^\theta)$ is a continuously differentiable function of θ, for any $\theta \in \Theta$[3]. Consider the ordinary differential equation (ODE):*

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

[2]The detailed proof is available in (Prashanth et al., 2015).

[3]In 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* $\mathcal{K} = \{\theta \mid \check{\Gamma}_i\left(\nabla_i \mathbb{C}(X^\theta)\right) = 0, \forall i = 1, \ldots, d\}$*. Then, for* θ_n *governed by* (8)*, we have*

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

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

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

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

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

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 (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 $i = 1, \ldots, \mathcal{M}$, with $d_n^{i,j}, j = 1, \ldots, n_i$ 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 (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.

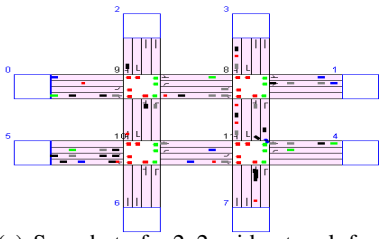
Figures 4(b)–4(d) present the histogram of the CPT-values from the test phase for AVG-SPSA, EUT-SPSA and CPT-SPSA, respectively. A similar exercise for pre-timed TLC resulted in a CPT-value of -46.14. It is evident that each algorithm converges to a different policy. However, the CPT-value of the resulting 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

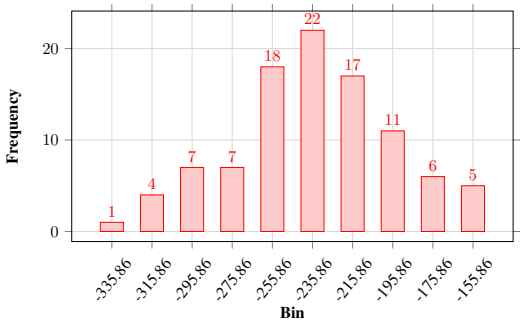
CPT has been a very popular paradigm for modeling human decisions among psychologists/economists, but has escaped the radar of the AI community. This work is the first step in incorporating CPT-based criteria into an RL framework. However, both prediction and control of CPT-based value is challenging.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

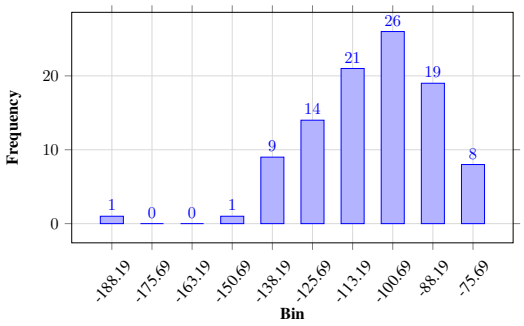
This work was supported by the Alberta Innovates Technology Futures through the Alberta Ingenuity Centre for Machine Learning, NSERC, the National Science Foundation (NSF) under Grants CMMI-1434419, CNS-1446665, and CMMI-1362303, and by the Air Force Office of Scientific Research (AFOSR) under Grant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



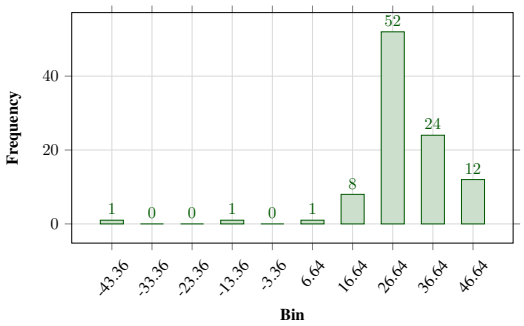
(a) Snapshot of a 2x2-grid network from GLD simulator. The figure shows eight edge nodes that generate traffic, four traffic lights and four-laned roads carrying cars.



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

References

Allais, M. Le comportement de l'homme rationel devant le risque: Critique des postulats et axioms de l'ecole americaine. *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.