Cumulative Prospect Theory Meets Reinforcement Learning: Prediction and Control

Prashanth L.A.

Joint work with Cheng Jie, Michael Fu, Steve Marcus and Csaba Szepesvári

University of Maryland, College Park
AI that benefits humans

Reinforcement learning (RL) setting with rewards evaluated by humans

Cumulative prospect theory (CPT) captures human preferences
For a given r.v. $X$, CPT-value $C(X)$ is

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

- **Utility functions** $u^+, u^- : \mathbb{R} \to \mathbb{R}_+$, $u^+ (x) = 0$ when $x \leq 0$, $u^- (x) = 0$ when $x \geq 0$

- **Weight functions** $w^+, w^- : [0, 1] \to [0, 1]$ with $w(0) = 0$, $w(1) = 1$
For a given r.v. $X$, CPT-value $C(X)$ is

$$C(X) := \int_{0}^{+\infty} w^+ (\mathbb{P}(u^+(X) > z)) \, dz - \int_{0}^{+\infty} w^- (\mathbb{P}(u^-(X) > z)) \, dz$$

Gains

Losses

Utility functions $u^+, u^- : \mathbb{R} \rightarrow \mathbb{R}_+$, $u^+(x) = 0$ when $x \leq 0$, $u^-(x) = 0$ when $x \geq 0$

Weight functions $w^+, w^- : [0, 1] \rightarrow [0, 1]$ with $w(0) = 0$, $w(1) = 1$

Connection to expected value:

$$C(X) = \int_{0}^{+\infty} \mathbb{P}(X > z) \, dz - \int_{0}^{+\infty} \mathbb{P}(-X > z) \, dz$$

$$= \mathbb{E} [(X)^+] - \mathbb{E} [(X)^-]$$

$(a)^+ = \max(a, 0)$, $(a)^- = \max(-a, 0)$
Utility and weight functions

**Utility functions**

For losses, the disutility $-u^-$ is convex, for gains, the utility $u^+$ is concave.

**Weight function**

Overweight low probabilities, underweight high probabilities.
Kahneman & Tversky (1979) “Prospect Theory: An analysis of decision under risk” is the second most cited paper in economics during the period, 1975-2000
Our Contributions

\[ \mathbb{C}(X^\theta) := \int_0^{+\infty} w^+ \left( \mathbb{P} \left( u^+ (X^\theta) > z \right) \right) \, dz - \int_0^{+\infty} w^- \left( \mathbb{P} \left( u^- (X^\theta) > z \right) \right) \, dz \]

Find \( \theta^* = \arg \max_{\theta \in \Theta} \mathbb{C}(X^\theta) \)

- CPT-value estimation using empirical distribution functions
- SPSA-based policy gradient algorithm
- sample complexity bounds for estimation + asymptotic convergence of policy gradient
- traffic signal control application
Problem: Given samples $X_1, \ldots, X_n$ of $X$, estimate

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

Nice to have: Sample complexity $O \left( \frac{1}{\epsilon^2} \right)$ for accuracy $\epsilon$
Empirical distribution function (EDF): Given samples $X_1, \ldots, X_n$ of $X$,

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

Using EDFs, the CPT-value $\mathbb{C}(X)$ is estimated by

$$\overline{C}_n = \int_{0}^{\infty} w^+(1 - \hat{F}_n^+(x))dx - \int_{0}^{\infty} w^-(1 - \hat{F}_n^-(x))dx$$

Part (I) \hspace{2cm} Part (II)
Empirical distribution function (EDF): Given samples $X_1, \ldots, X_n$ of $X$,

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

Using EDFs, the CPT-value $\overline{C}(X)$ is estimated by

$$\overline{C}_n = \int_0^{+\infty} w^+ (1 - \hat{F}_n^+(x)) \, dx - \int_0^{+\infty} w^- (1 - \hat{F}_n^-(x)) \, dx$$

Part (I) \hspace{1cm} Part (II)

**Computing Part (I):** Let $X_1, X_2, \ldots, X_n$ denote the order-statistics

$$\text{Part (I)} = \sum_{i=1}^{n} u^+(X[i]) \left( w^+ \left( \frac{n + 1 - i}{n} \right) - w^+ \left( \frac{n - i}{n} \right) \right),$$
(A1). Weights $w^+, w^-$ are Hölder continuous, i.e.,
\[ |w^+(x) - w^+(y)| \leq H|x - y|^\alpha, \forall x, y \in [0, 1] \]

(A2). Utilities $u^+(X)$ and $u^-(X)$ are bounded above by $M < \infty$

**Sample Complexity:**

Under (A1) and (A2), for any $\epsilon, \delta > 0$, we have

\[
\mathbb{P} \left( |\bar{C}_n - C(X)| \leq \epsilon \right) > 1 - \delta, \forall n \geq \ln \left( \frac{1}{\delta} \right) \cdot \frac{4H^2M^2}{\epsilon^{2/\alpha}}
\]
(A1). Weights \( w^+, w^- \) are Hölder continuous, i.e.,
\[
|w^+(x) - w^+(y)| \leq H|x - y|^\alpha, \forall x, y \in [0, 1]
\]

(A2). Utilities \( u^+(X) \) and \( u^-(X) \) are bounded above by \( M < \infty \)

Sample Complexity:

Under (A1) and (A2), for any \( \epsilon, \delta > 0 \), we have

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

Special Case: Lipschitz weights \( (\alpha = 1) \)

Sample complexity \( O \left( \frac{1}{\epsilon^2} \right) \) for accuracy \( \epsilon \)
CPT-value optimization

Find \( \theta^* = \arg \max_{\theta \in \Theta} C(X^\theta) \)

RL application: \( \theta = \) policy parameter, \( X^\theta = \) return

Two-Stage Solution:

inner stage Obtain samples of \( X^\theta \) and estimate \( C(X^\theta) \);

outer stage Update \( \theta \) using gradient ascent

\( \nabla_i C(X^\theta) \) is not given
Update rule: \( \theta_{n+1}^i = \Gamma_i \left( \theta_n^i + \gamma_n \hat{\nabla}_i C(X^\theta_n) \right) \), \( i = 1, \ldots, d \).

Challenge: estimating \( \nabla_i C(X^{\theta}) \) given only biased estimates of \( C(X^{\theta}) \)

Solution: use SPSA [Spall’92]

\[
\hat{\nabla}_i C(X^{\theta}) = \frac{\overline{C}_{n+\delta_n \Delta_n}^{\theta_n} - \overline{C}_{n-\delta_n \Delta_n}^{\theta_n}}{2 \delta_n \Delta_n^i}
\]

\( \Delta_n \) is a vector of independent Rademacher r.v.s and \( \delta_n > 0 \) vanishes asymptotically.
Measurement Oracle $\rightarrow f(x) + \xi$  

Zero mean

Simulation optimization

$X, \epsilon \rightarrow$ CPT Estimator $\rightarrow C(X) + \epsilon$

Controlled bias

$\delta_n \Delta_n$

$m_n$ samples

Prediction

$\frac{\theta_n + \delta_n \Delta_n}{C_n}$

Update $\theta_n$

(Gradient ascent)

$\theta_{n+1}$

$\delta_n \Delta_n$

$m_n$ samples

Prediction

$\frac{\theta_n - \delta_n \Delta_n}{C_n}$

Control

Figure 1: Overall flow of CPT-SPSA

How to choose $m_n$ to ignore estimation bias? Ensure $\frac{1}{m_n \alpha/2 \delta_n} \rightarrow 0$
Application: Traffic signal control

- For any path $i = 1, \ldots, M$, let $X_i$ be the delay gain
  - calculated with a pre-timed traffic light controller as reference
- CPT captures the road users’ evaluation of the delay gain $X_i$
- Goal: Maximize

$$
CPT(X_1, \ldots, X_M) = \sum_{i=1}^{M} \mu^i C(X_i)
$$

$\mu^i$: proportion of traffic on path $i$
(a) AVG-SPSA

(b) EUT-SPSA

(c) CPT-SPSA

Figure 2: Histogram of CPT-value of the delay gain: AVG uses plain sample means (no utility/weights), EUT uses utilities but no weights and CPT uses both.
Conclusions

• Want AI to be beneficial to humans

• CPT - a very popular paradigm for modeling human decisions
Conclusions

- Want AI to be beneficial to humans
- **CPT** - a very popular paradigm for modeling human decisions
- We lay the foundations for using **CPT** in an **RL** setting
  - Prediction: Sample means (TD) won’t work, but empirical distributions do!
  - Control: No Bellman, but SPSA can be employed

**Future directions:**

- **Crowdsourcing** experiment to validate CPT online
- **Robustness** to unknown utility and weight function parameters
Thanks! Questions?