A Prospect Theoretic Policy Gradient Algorithm for Behavioral Alignment in Reinforcement Learning

Anas Barakat

Joint work with Olivier Lepel

April 21th 2025 - Finance and RL Talks Series

ETH zürich

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Motivation: Beyond expected returns and risk-sensitive RL

- Classical RL: expected utility theory
- Limitation: Misalignment with human preferences due to complexities of human decision making and underlying psychological nuances of perception.
  - Asymmetric perception of gains and losses
  - Probability distortions inherent in human cognition,
    e.g. tendency to over-estimate rare events and underestimate frequent ones

Focus

Human-centered **sequential decision-making** models incorporating cognitive and psychological biases, essential for high-stakes, socially beneficial applications.

Historical Bit: Behavioral Economics and Prospect Theory

- Behavioral Economics: Infusing standard economics analysis with psychological understanding of how people make decisions.

- Daniel Kahneman awarded the Nobel Prize in Economic Sciences in 2002:

  'for having integrated insights from psychological research into economic science, especially concerning human judgment and decision-making under uncertainty.'

Cumulative Prospect Theoretic RL: Problem Formulation

- (a) Ref. point, (b) Utility $\mathcal{U} : \mathbb{R} \to \mathbb{R}_+$, (c) Prob. distortion $w : [0,1] \to [0,1]$

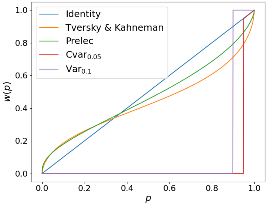
Cumulative Prospect Theory Value

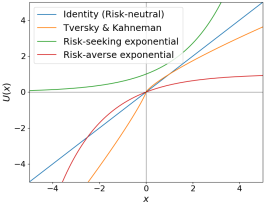
The CPT value of a real-valued random variable X is

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

CPT Policy Optimization

$$\max_{\pi \in \Pi} \mathbb{C}\left[\sum_{t=0}^{H-1} r_t\right] \qquad \text{(CPT-PO)}$$





Why CPT-RL? Personalized Treatment for Pain Management Example

Patients and clinicians make **sequential** decisions influenced by **psychological biases**.

- **Reference points**: patients pain level assessment and reporting (psych. bias).
- **Utility transformation:** *loss aversion* (patients might perceive pain increase or withdrawal symptoms as worse than equivalent gains in pain relief).
- **Probability distortion:** Low probability events such as severe side effects (e.g. dependency to medication) over or underweighted based on patient's psychology.

Prior Work

- **CPT in stateless static settings:** Wide adoption and widespread applications in:
  - Psychiatry [Sip et al., 2018, George et al., 2019, Mkrtchian et al., 2023]
  - Chronic diseases treatment [Zhao et al., 2023]
  - Emergency decision making [Sun et al., 2022]
  - Energy [Ebrahimigharehbaghi et al., 2022, Dorahaki et al., 2022]
  - **Finance** [Luxenberg et al., 2024]
- **CPT-RL:** Understanding and practical impact of CPT-RL remains limited despite:
  - A few works integrating CPT into RL [L.A. et al., 2016, Borkar and Chandak, 2021, Ramasubramanian et al., 2021, Danis et al., 2023].
  - Limited understanding of optimal policies in CPT-RL.
  - Computational challenges: CPT-SPSA-G algorithm, 0-th order algorithm (scaling issues, trajectory sampling, does not exploit sequential structure in rewards).

Our Contributions in a Nutshell

Central Question

How to align the agent's behavior with the given preferences by optimizing for CPT return values?

- Nature of optimal policies in CPT-RL
  - Existence of optimal deterministic Markovian policy like in standard MDPs?
  - What if we remove probability distortions in CPT?
  - Are there specific utility function classes for which there exist Markovian policies?
- **Policy Gradient Theorem for CPT-RL**
- **Policy Gradient Algorithm for CPT-RL**

Policy Gradient Theorem for CPT-RL

- Continuous utility functions u^-, u^+
- Lipschitz and differentiable weight functions w_-, w_+
- Differentiable policy parametrization $\theta \mapsto \pi_\theta(a|h)$

PG Theorem for CPT-RL

$$\forall \theta \in \mathbb{R}^d, \quad \nabla J(\theta) = \mathbb{E}\left[\varphi\left(R(\tau)\right) \sum_{t=0}^{H-1} \nabla_\theta \log \pi_\theta(a_t|h_t)\right],$$

$$\tau := (s_t, a_t, r_t)_{0 \le t \le H-1}, \quad R(\tau) := \sum_{t=0}^{H-1} r_t$$

$$\varphi(v) := \int_{z=0}^{\max(v,0)} w_+'(\mathbb{P}(u^+(R(\tau)) > z))dz - \int_{z=0}^{\max(-v,0)} w_-'(\mathbb{P}(u^-(R(\tau)) > z))dz$$

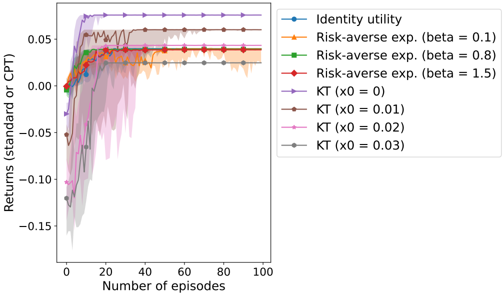
Policy Gradient Algorithm for CPT-RL

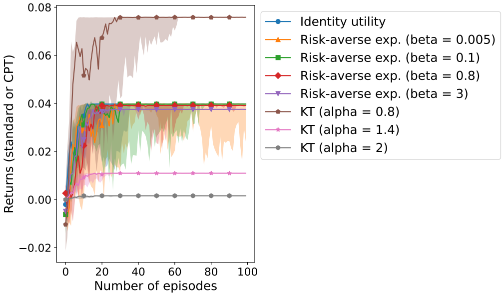
Algorithm 1 CPT-Policy Gradient Algorithm (CPT-PG)

Input: $\theta_0 \in \mathbb{R}^d$, utility functions u^+, u^-, weight functions w_+, w_-, step size $\alpha > 0$.
for $k = 0, \cdots, K$, **do**
    /Policy gradient estimation
    Sample a trajectory $\tau := (s_t, a_t, r_t)_{0 \leq t \leq H-1}$, with $s_0 \sim \rho$ following π_{θ_k}
    // Quantile estimation
    Sample n trajectories $\tau_j := (s_t^j, a_t^j, r_t^j)_{0 \leq t \leq H-1}$, $1 \leq j \leq n$ with $s_0^j \sim \rho$ following π_{θ_k}
    Compute and order $R(\tau_j)$, label them as: $R(\tau_{[1]}) < R(\tau_{[2]}) < \cdots < R(\tau_{[n]})$
    $\hat{\xi}^+_{\frac{i}{n}} = u^+(R(\tau_{[i]}))$; $\hat{\xi}^-_{\frac{i}{n}} = u^-(R(\tau_{[i]}))$
    //Approximation of $\varphi(R(\tau))$
    $\hat{\phi}_n^\pm = \sum_{i=0}^{j_n-1} w'_\pm\left(\frac{i}{n}\right)\left(\hat{\xi}^\pm_{\frac{n-i}{n}} - \hat{\xi}^\pm_{\frac{n-i-1}{n}}\right) + w'_\pm\left(\frac{j_n}{n}\right)\left(R(\tau) - \hat{\xi}^\pm_{\frac{n-j_n-1}{n}}\right)$
    $\hat{g}_k = (\hat{\phi}_n^+ - \hat{\phi}_n^-)\sum_{t=0}^{H-1} \nabla_\theta \log \pi_{\theta_k}(a_t|h_t)$
    /Policy gradient update
    $\theta_{k+1} = \theta_k + \alpha\, \hat{g}_k$
end for

Application to Trading in Financial Markets

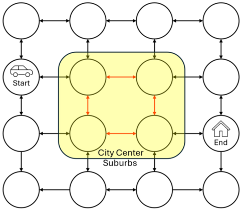
- **Gym Trading Environment**: Bitcoin USD (BTC-USD) market data on 4 years.
- **States**: few extracted features ('open', 'high', 'low', 'close') prices and volume.
- **Rewards**: log values of the ratio of the portfolio valuations at times t and $t-1$.

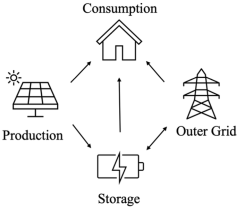




Conclusion

- Main goal: human-centered sequential decision-making models incorporating cognitive and psychological biases.
- Several other applications we explore: electricity energy management, traffic control on a grid, control on MuJoCo ...
- **Potential for integration and further implementation impact in practice**





Check out our paper for more details!

References I

Borkar, V. S. and Chandak, S. (2021).
Prospect-theoretic q-learning.
Systems & Control Letters, 156:105009.

Danis, D., Parmacek, P., Dunajsky, D., and Ramasubramanian, B. (2023).
Multi-agent reinforcement learning with prospect theory.
2023 Proceedings of the Conference on Control and its Applications (CT), pages 9–16.

Dorahaki, S., Rashidinejad, M., Ardestani, S. F. F., Abdollahi, A., and Salehizadeh, M. R. (2022).
A home energy management model considering energy storage and smart flexible appliances: A modified time-driven prospect theory approach.
Journal of Energy Storage, 48:104049.

References II

- Ebrahimigharehbaghi, S., Qian, Q. K., de Vries, G., and Visscher, H. J. (2022).
  Application of cumulative prospect theory in understanding energy retrofit decision: A study of homeowners in the netherlands.
  Energy and Buildings, 261:111958.
- George, S. A., Sheynin, J., Gonzalez, R., Liberzon, I., and Abelson, J. L. (2019).
  Diminished value discrimination in obsessive-compulsive disorder: A prospect theory model of decision-making under risk.
  Frontiers in Psychiatry, 10:469.
- L.A., P., Jie, C., Fu, M., Marcus, S., and Szepesvari, C. (2016).
  Cumulative prospect theory meets reinforcement learning: Prediction and control.
  In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1406–1415, New York, New York, USA. PMLR.

References III

Luxenberg, E., Schiele, P., and Boyd, S. (2024).
Portfolio optimization with cumulative prospect theory utility via convex optimization.
Computational Economics, pages 1–21.

Mkrtchian, A., Valton, V., and Roiser, J. P. (2023).
Reliability of decision-making and reinforcement learning computational parameters.
Computational Psychiatry, 7(1):30.

Ramasubramanian, B., Niu, L., Clark, A., and Poovendran, R. (2021).
Reinforcement learning beyond expectation.
In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 1528–1535.

References IV

Sip, K. E., Gonzalez, R., Taylor, S. F., and Stern, E. R. (2018).
Increased loss aversion in unmedicated patients with obsessive–compulsive disorder.
Frontiers in Psychiatry, 8:309.

Sun, J., Zhou, X., Zhang, J., Xiang, K., Zhang, X., and Li, L. (2022).
A cumulative prospect theory-based method for group medical emergency decision-making with interval uncertainty.
BMC Medical Informatics and Decision Making, 22(1):124.

Zhao, M., Wang, Y., Meng, X., and Liao, H. (2023).
A three-way decision method based on cumulative prospect theory for the hierarchical diagnosis and treatment system of chronic diseases.
Applied Soft Computing, 149:110960.