Risk-Sensitive Multi-Agent Reinforcement Learning in Network Aggregative Markov Games

Hafez Ghaemi, Hamed Kebriaei, Majid Nili, Alireza Ramezani
ghaemi.hafez@ut.ac.ir School of ECE, University of Tehran

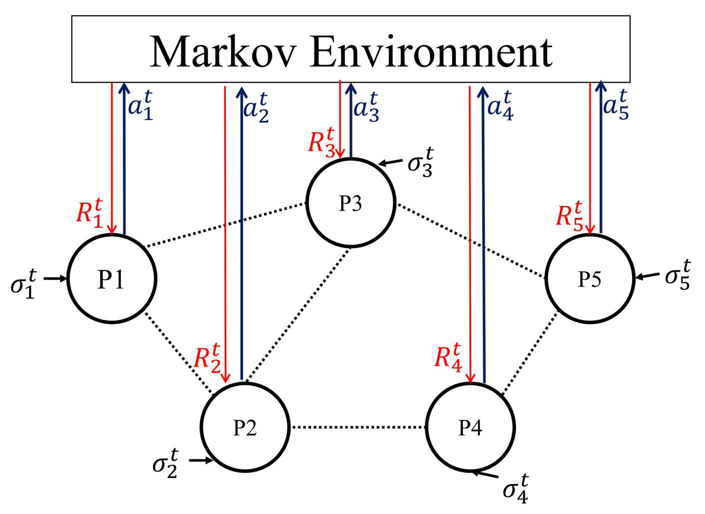
1. Overview

- Model **subjective** economic or social preferences in MARL
- Cumulative prospect theory (**CPT**) as risk. CPT generalizes coherent risk and explains **loss aversion** and **probability distortion** in humans.
- Net. Agg. Markov game (**NAMG**) to model distributed agent interactions
- **Nested CPT-AC**: distributed sampling-based actor-critic algorithm in NAMGs
- Possible convergence to a subjective Markov perfect **Nash** equilibrium
- Experiment shows agents with a higher CPT loss aversion are more inclined to **social isolation**.

arXiv **Code**

3. Net. Agg. Markov Games

Graph-based with $R^i(s, a^i, a^{-i}) = R^i(s, a^i, \sigma^i(a^{-i}))$, $\sigma^i(a^{-i}) = \sum_{j \in \mathcal{N} \setminus i} \omega_{ij} a^j$.



4. CPT Estimation via Sampling

Algorithm 1 CPT Value Estimation [1]

1: **Require:** Samples $X_1, \ldots, X_n$ from r.v. X, sorted in ascending order.
2: Let

$$\hat{\rho}^+_{cpt} := \sum_{i=1}^{n} u^+(X_i)\left(\omega^+\left(\frac{n+1-i}{n}\right) - \omega^+\left(\frac{n-i}{n}\right)\right)$$

$$\hat{\rho}^-_{cpt} := \sum_{i=1}^{n} u^-(X_i)\left(\omega^-\left(\frac{i}{n}\right) - \omega^-\left(\frac{i-1}{n}\right)\right)$$

3: Return $\hat{\rho}_{cpt} = \hat{\rho}^+_{cpt} - \hat{\rho}^-_{cpt}$.

8. Future Work and References

A plausible future direction is the use of function approximation and deep RL in CPT-sensitive MARL for large-scale human behavior simulations, albeit w/o theoretical gurantees.

[1] Cheng Jie, LA Prashanth, Michael Fu, Steve Marcus, and Csaba Szepesvári. Stochastic optimization in a cumulative prospect theory framework. *IEEE Transactions on Automatic Control*, 63(9):2867–2882, 2018.

2. Cumulative Prospect Theory

Given r.v. X, weighting function such as $\omega(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{(1/\gamma)}}$, and **convex-concave utility** function such as $u^+(x) = x^\alpha$ for $x \geq 0$ and $-u^-(x) = -\lambda(-x)^\beta$ for $x < 0$, CPT value is defined as:

$$\mathbb{CPT}_{\mathbb{P}}[X] := \sum_{i=0}^{n} \phi^+(\mathbb{P}(X = x_i)) u^+(x_i - x^0) - \sum_{-m}^{-1} \phi^-(\mathbb{P}(X = x_i)) u^-(x_i - x^0), \tag{1}$$

where $\phi^{\pm}$ is a cumulative probability weighting function for gains and losses. Humans are generally loss-averse (due to convex-concave utility), and they overestimate/underestimate small/large probabilities (due to probability weighting). **CPT MARL objective** in NAMG:

$$\max_{\pi^i} V^i_\pi(s_0) = \max_{\pi^i} {}_{\pi^i(a^i_0|s_0)\times(\sigma^{-i}_0|s_0)\times(s_1|s_0,a_0)} \left[R^i(s_0, a_0) + \gamma V^i_\pi(s_1)\right]. \tag{2}$$

5. Nested CPT Policy Gradient

Theorem 1. The gradient of the CPT return for agent i, $V^i_{\pi_\theta}(s_0)$, with respect to the policy parameter θ^i is

$$\nabla V^i_{\pi_\theta}(s_0) \propto_{\mu^i_{cpt}(s)} \left[\sum_{a,s'} \frac{\partial \phi}{\partial(\pi^i_\theta(a^i|s)(\sigma^{-i}|s)(s'|s,a))} (\sigma^{-i}|s)(s'|s,a)(\nabla \pi_{\theta^i}(a^i|s)) u(R^i(s,a^i,\sigma^{-i},s') + \gamma V^i_{\pi_\theta}(s'))\right], \tag{3}$$

where distribution μ^i_{cpt} is a *subjective* steady-state probability distribution of the MDP.

6. Distributed Nested CPT Actor-Critic

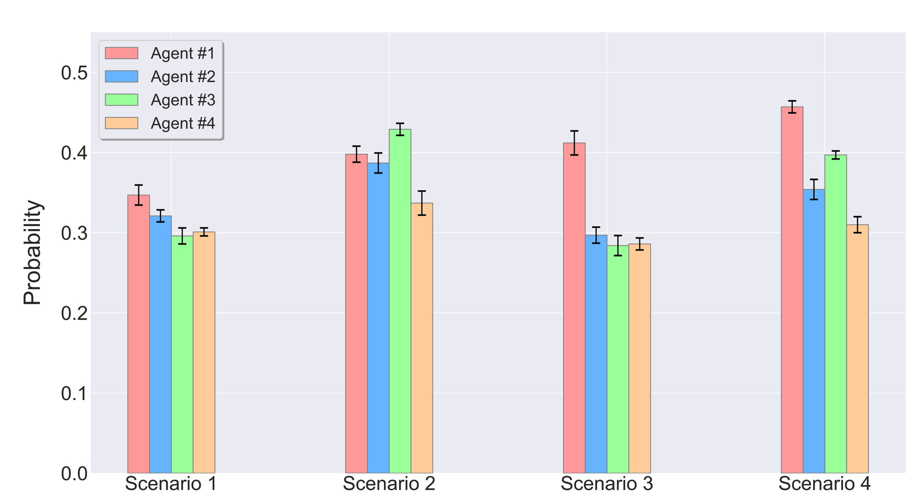
Algorithm 2 Distributed Nested CPT Actor-Critic

1: **For each agent n, repeat until convergence:**
2: Sample a^n_t from $\pi_{\theta^n_t}(.|s_t)$. Execute a^n_t and observe r^n_t, s_{t+1}, and σ^{-n}_t. Push $(r_t, s_{t+1}, \sigma^{-n}_t)$ to $ExpDict^n(s_t, a^n_t, \sigma^{-n}_t)$.
3: **Critic value estimation:**
4: **for** $i = 1, 2, \ldots, n_{max}$, **do**
5: Sample $\hat{a}^n_t$ from $\pi_{\theta^n_t}(.|s_t)$ and construct $\hat{\sigma}^{-n}_t$ by observing neighbors. Sample $(\hat{r}^n_t, \hat{s}_{t+1})$ from $ExpDict(s_t, \hat{a}^n_t, \hat{\sigma}^{-n}_t)$ or a simulator of the environment.
6: Let $X_i := \hat{r}^n_t + \gamma V^n_{\pi_\theta}(\hat{s}_{t+1})$. If the sample came from a simulator, push $(\hat{r}^n_t, \hat{s}_{t+1})$ to $ExpDict(s_t, \hat{a}^n_t, \hat{\sigma}^{-n}_t)$.
7: **end for**
8: Estimate $\hat{V}^n_{\pi_{\theta_t}}(s_t)$ using array of X and Algorithm 1.
9: **Critic step:**
10: $\delta_t := \hat{V}^n_{\pi_{\theta_t}}(s_t) - V^n_{\pi_{\theta_t}}(s_t)$, $V^n_{\pi_{\theta_t}}(s_t) \leftarrow V^n_{\pi_{\theta_t}}(s_t) + \alpha_{cr,t}\delta_t$.
11: **Actor step:** Compute $\nabla V^n_{\pi_{\theta_t}}(s_0)$ using the gradient estimation scheme based on Theorem 1, and then $\theta^n_{t+1} := \theta^n_t + \alpha_{ac,t} \nabla V^n_{\pi_{\theta_t}}(s_0)$.

Under a set of assumptions, Algorithm 2 asymptotically converges to the unique CPT-sensitive Markov perfect Nash equilibrium of the NAMG. If the assumptions do not hold, we can only ensure convergence to locally optimal policies.

7. Numerical Experiment

Experiment: $R^i(s, a^i, \sigma^i(a^{-i})) = R^i_{self}(s) + \sigma^i(a^{-i}) R^i_{com}(s) a^i$, with $R_{self}(s, a^i) \sim \mathcal{N}(0.5, 0.1)$ and $R^i_{com}(s) \sim 5 \cdot Unif(-0.5, 0.5)$, and $\sigma^i(a^{-i}) = \frac{1}{N-1}(\sum_{j \in \mathcal{N} \setminus i} a_j)$. The setup implies a high risk for agent if it decides to take $a^i > 0$, to become socially involved with its neighboring community and tie its received reward to their actions. Probability of $a = 0$ is a quantitative indicator of **social conservatism** and we observed it is also **proportional to CPT loss-aversion** coeff. of agents.



Converged probability of $a = 0$ in s_0 for different loss aversion scenarios. Scenario 1: all agents risk-neutral, scenario 2: all agents risk-sensitive ($\lambda = 2.6$), scenario 3: only Agent 1 is risk-sensitive ($\lambda = 2.6$), scenario 4: Agent 1 has a higher CPT loss aversion coefficient ($\lambda = 3.2$) than others ($\lambda = 2.6$).