Soft-Robust Actor-Critic Policy-Gradient

YANG, ZONG-YING
Institute of Multimedia Engineering
National Yang Ming Chiao Tung University
Student ID: 309553038
wl02227976@gmail.com

1 Introduction

For some scenario, there may be the awful result from some choice. Therefore, the robust strategies will make the learning result too conservative. In this paper, they propose the idea to relax this conservativeness and construct a softer behavior that interpolates between being aggressive and robust.
The main contributions are:

- A soft-robust derivation of the objective function for policy-gradient.
- An Soft-Robust Actor-Critic(SR-AC) algorithm that uses stochastic approximation to learn a variant of distributionally robust policy in an online manner.
- Show the efficiency of soft-robust behaviors in a continuous action space as well.

The robust strategies may keep safe for making the choice. However, to get the better reward, a good trade-off between low risk and high returns is necessary.

2 Problem Formulation

Assumption 1: Under any policy π, the Markov chains resulting from any of the MDPs with transition laws $p \in P$ are irreducible and aperiodic.

d_p^π: the stationary distribution of the Markov chain that results from following policy π under transition model $p \in P$.

Soft-robust objective function: $\bar{J}(\pi) := E_{p\sim\omega}[J_p(\pi)]$
In the soft-robust objective function, there is the distribution ω, it introduces a softer form of robustness because it averages over the uncertainty set instead of considering the worst-case scenario.

The average transition model: $\bar{p} := E_{p\sim\omega}[p]$
The stationary distribution for average transition model: $\bar{d^\pi}(x) = E_{p\sim\omega}[d_p^\pi(x)]$

Soft-robust differential reward function: $\bar{Q}^\pi(x,a) := E_{p\sim\omega}[Q_p^\pi(x,a)]$
where $Q_p^\pi(x,a) := E^p[\sum_{t=0}^{+\infty} r_{t+1} - J_p(\pi)|x_0 = x, a_0 = a, \pi]$

soft-robust value function: $\bar{V}^\pi(x) := \sum_{a\in A} \pi(x,a)\bar{Q}^\pi(x,a) = E_{p\sim\omega}[V_p^\pi(x)]$
with $V_p^\pi(x) := \sum_{a\in A} \pi(x,a)Q_p^\pi(x,a)$

As in regular MDPs, the soft-robust average reward satisfies a Poisson equation. Reformulate it:
$\bar{J}(\pi) + \bar{V}^\pi(x) = \sum_{a\in A} \pi(x,a)(r(x,a) + \sum_{x'\in X} \bar{p}(x,a,x')\bar{V}^\pi(x'))$

Then, by $\bar{d}^{\pi}$, $\bar{J}(\pi) = \sum_{x \in X} \bar{d}^{\pi}(x) \sum_{a \in A} \pi(x,a) r(x,a)$

The goal is to learn a policy that maximizes the soft-robust average reward $\bar{J}$, use a policy-gradient method to reach it.
consider a class of parametrized stochastic policies $\pi_\theta : X \to M(A)$ with $\theta \in R^{d_1}$

The optimal set of parameters thus obtained is denoted by

$$\theta^* := arg \max_{\theta} \bar{J}(\pi_\theta) \tag{1}$$

Assumption 2: For any $(x,a) \in X \times A$, the mapping $\theta \longmapsto \pi_\theta(x,a)$ is continuously differentiable with respect to θ.

Soft-Robust Policy-Gradient:
For any MDP satisfying previous assumptions, there is

$$\bigtriangledown \bar{J}(\pi) = \sum_{x \in X} \bar{d}^{\pi}(x) \sum_{a \in A} \bigtriangledown_\theta \pi(x,a) \bar{Q}^{\pi}(x,a) \tag{2}$$

Soft-Robust Policy-Gradient with Function Approximation:
Let $f_\omega : X \times A \to R$ be a linear approximation of the soft-robust differential reward $\bar{Q}^{\pi}$.
If f_ω minimizes the mean squared error

$$\epsilon^{\omega} := \sum_{x \in X} \bar{d}^{\pi}(x) \sum_{a \in A} \pi(x,a) [\bar{Q}^{\pi}(x,a) - f_\omega(x,a)]^2 \tag{3}$$

and $\bigtriangledown_\omega f_\omega(x,a) = \bigtriangledown_\theta \log_\pi(x,a)$
Then,

$$\bigtriangledown_\theta \bar{J}(\pi) = \sum_{x \in X} \bar{d}^{\pi}(x) \sum_{a \in A} \bigtriangledown_\theta \pi(x,a) f_\omega(x,a) \tag{4}$$

In SR-AC algorithm, samples are generated using the nominal model and the current policy. Then samples are utilized to update the soft-robust average reward and the critic based on the TD-error. The soft-robust value function is critic according to which the actor parameters are updated. Then improve the policy by updating the policy parameters in the direction of a gradient estimate for the soft-robust objective. Repeat these until convergence.

3 Theoretical Analysis

Robust MDP:
An robust MDP is a tuple $\{X, Z, U, P, r, \gamma\}$, X is a finite set of states, Z is a set of absorbing terminal set, U is a finite set of actions, $r : X \times U \to R$ is a deterministic and bounded reward function, γ is a discount factor, and P, where $P(x,u) \subset M(X \cup U)$, denotes a known uncertainty in the state transitions.
In robust MDPs, it is typically interested in maximizing the worst case performance, so the robust value function for a policy π as its worst-case valuw function

$$V^{\pi}(x) = \inf_{p \in P} V^{\pi,p}(x) \tag{5}$$

Therefore, the learning result from robust MDP will be conservative to avoid the worst case performance. However, sometimes we have to accept the risk to get better performance.
The SR-AC algorithm is unlike robust MDP that maximize the worst-case performance. It added a distribution ω which can be thought as the adversary distribution over different transition models. The use of the distribution is to find a balance between aggressive and robust.

4 Conclusion

I thought the concept of robust strategies is similar to the constrained strategies. Therefore, combined the SR-AC with the constrained MDP may be the future research directions. The constrained MDP is to constrain the updated step. The robust MDP is to maximize the worst case. Combined one

may make the learning not avoid the worst case too much and find the better performance under the tolerable risk. Another future research directions mentioned in paper is to address the problem of learning the sequential game induced by an evolving adversarial distribution to derive an optimal soft-robust policy.

References

[1] Wolfram Wiesemann, Daniel Kuhn, Berç Rustem, (2012) Robust Markov Decision Processes. Mathematics of Operations Research 38(1):153-183.

[2] Tamar, A., Mannor, S. ; Xu, H.. (2014). Scaling Up Robust MDPs using Function Approximation.

[3] Ho, C.P., Petrik, M. ; Wiesemann, W.. (2018). Fast Bellman Updates for Robust MDPs.