2021-Zheng

Stackelberg Actor-Critic: A Game-Theoretic Perspective

- **Date**: 2021
- **Link:** paper
- **Authors:**
  - Zheng, Liyuan
  - Fiez, Tanner
  - Alumbaugh, Zane
  - Chasnov, Benjamin
  - Ratliff, Lillian J
- **Cites:**
- **Cited by:**
- **Keywords:** #stackelberg-games #actor-critic #stackelberg-actor-critic #bilevel-optimization
- **Collections:**
- **Status:** #completed

---

0 - Abstract

- *Actor-critic RL method*:
  - *Critic* - updates its approximate expected return of the actor.
  - *Actor* - updates its policy in a direction based on the critic's estimation.

- The actor-critic method has an intrinsic hierarchical structure between the *actor* and the *critic*.

> *Key Idea*: Exploit this hierarchy to formulate the actor-critic method as a two-player general-sum Stackelberg game.

- The algorithm:
  - Leverage the Stackelberg gradient update following the total derivative.
  - The *actor* optimizes utilizing the knowledge that the *critic* responds near-optimally to the update by the *actor*.
    - i.e. the *actor* takes an action, the *critic* observes the action and makes a best response action, just like in Stackelberg games.

- The algorithm out performs normal actor-critic in experiments.

- Propose that the algorithm could be generalized to general actor-critic based methods.

1 - Introduction

- Intrinsic hierarchical structure of actor-critic method:
  - *critic* seeks to be at an optimum given the parameters of the *actor*.
  - *actor* seeks to be at an optimum knowing that the critic responds near-optimally to the parameters selected by the actor.
  - This can be viewed as a Stackelberg game!

- *Stackelberg games* - characterize the interaction between a leader and a follower:
  - *Leader* - acts before the follower, therefore must account for how the follower will respond.
  - *Follower* - selects a best response to the leader's action.

- Actor-critic method as a two-player general-sum Stackelberg game formulation:
  - The *actor* is the *leader*.
  - The *Actor* must solve a bilevel optimization problem in which the actor objective is a function of the *critic*'s parameters.
  - The *critic* is the *follower*.
  - The *critic* responds optimally with respect to its own parameters.

- *Stackelberg actor-critic* - novel algorithm that explicitly takes into account the interaction structure between the players.

- Shown experimentally to produce more robust solutions.

1.1 - Related Work

- Many papers studying game-theoretic frameworks for multi-agent RL.

- There are fewer papers studying game-theoretic frameworks for single-agent RL, like this paper does.
  - The existing paper only considers gradient descent-ascent to approx. the Stackelberg dynamics.

- Previous actor-critic works have used local second-order information to construct a constrained optimization problem.
  - In this work, a bilevel optimization is used that allows for actor-critic interactions to be characterized.

2 - Preliminaries

2.1 - Actor-Critic

- Problem description:
  - Discrete-time MDP.
  - Continuous state $\mathcal{S}$ and action $\mathcal{A}$ spaces.
  - Initial state s_0 is determined by the initial state density $s_0 \sim \rho(s)$.

- Expected return of a policy π after executing a_t in state s_t is expressed by the Q function:

$$Q^\pi(s_t,\, a_t) = E_{\tau\sim\pi}\left[\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'},\, a_{t'}) \mid s_t, a_t\right]$$

- Expected return of a policy π in state s_t can be expressed by the V function:

$$V^\pi(s_t) = E_{\tau\sim\pi}\left[\sum_{t'=t}^{T} \gamma^{t'-t} r(s_{t'},\, a_{t'}) \mid s_t\right]$$

- Goal of RL is to find an optimal policy that maximizes the expected return.

$$J(\pi) = E_{\tau\sim\pi}\left[\sum_{t=0}^{T} \gamma^{t} r(s_t,\, a_t)\right]$$

$$J(\pi) = \int_\tau p(\tau \mid \pi)\, R(\tau)\, d\tau$$

$$J(\pi) = E_{s\sim\rho,\, a\sim\pi(\cdot|s)}[Q^\pi(s, a)]$$

  Note: We sample a trajectory following π, then we weight the cumulative reward of this trajectory by the probability of it occurring, given our policy. We then sum this over all trajectories, giving the expected cumulative reward for the π.

- Policy-based approach - parameterizes π by θ and finds optimal θ^* by maximizing the expected return:

$$\max_\theta J(\theta) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}[Q^\pi(s, a)]$$

- Applying the policy gradient theorem:

$$\nabla_\theta J(\theta) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) Q^\theta(s, a)\right]$$

- This optimization problem can be solved by gradient ascent.

- Actor-critic method adds another parameterization w for the Q function, $Q_w(s, a)$:

$$\max_\theta J(\theta) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[Q_w(s, a)\right]$$

- Optimization is solved by gradient ascent:

$$\nabla_\theta J(\theta) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[\nabla_\theta \log \pi_\theta(a|s) Q_w(s, a)\right]$$

- The critic is optimized by minimizing the error between true value functions:

$$\min_w L(w) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[(Q_w(s, a) - Q^\pi(s, a))^2\right]$$

  where the true value is estimated by MC or bootstrapping.

- Actor-critic methods typically perform direct descent-ascent:

$$\theta \leftarrow \theta + \alpha_\theta \nabla_\theta J(\theta)$$
$$w \leftarrow w - \alpha_w \nabla_w L(w)$$

2.2 - Stackelberg Game

- Definitions:
  - $f_1(x_1,\, x_2)$ and $f_2(x_1, x_2)$ - objective functions the leader and follower want to minimize.
  - x_1 and x_2 - player's decision variables.

- Leader's objective function:

$$\min_{x_1}\{f_1(x_1,\, x_2) \mid x_2 = \arg\min_y f_2(x_1, y)\}$$

- *Best response* - $x_2^* = \arg\min_y f_2(x_1, y)$. Leader assumes the follower always selects it's best action given the leader's action.

- Total derivative of the leader's cost function:

$$\frac{df_1(x_1,\, x_2^*(x_1))}{dx_1} = \frac{\partial f_1(x_1, x_2)}{\partial x_1} + \frac{dx_2^*(x_1)}{dx_1}\,\frac{\partial f_1(x_1,\, x_2)}{\partial x_2}$$

- Implicit Jacobian term, obtained using the implicit function theorem:

$$\frac{dx_2^*(x_1)}{dx_1} = -\left(\frac{\partial^2 f_2(x_1,\, x_2)}{\partial x_1 \partial x_2}\right)\left(\frac{\partial^2 f_2(x_1,\, x_2)}{\partial x_2^2}\right)^{-1}$$

3 - Stackelberg Actor-Critic

- Formulate actor-critic as a two-player general sum Stackelberg game.
- Actor and critic can only pick their own parameters while their objectives depend on both player's parameters.
- Critic objective:

$$L(\theta,\, w) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[\left(Q_w(s,a) - Q^\pi(s,a)\right)^2\right]$$

> *Key Idea* - to yield an accurate approximation, the critic should be selecting a best response:
>
> $$w^*(\theta) = \arg\min_\phi L(\theta,\, \phi)$$
>
> Thus, the *critic* assumes the role of the *follower*.

- Actor aims to solve the bilevel optimization problem given by

$$\begin{aligned} &\max\nolimits_\theta \; J(\theta,\, w^*(\theta)) \\ \text{s.t.} \;\; & w^*(\theta) = \arg\min\nolimits_\phi L(\theta, \phi) \end{aligned}$$

- The actor objective is also a function of both parameters:

$$J(\theta, w) = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[Q_w(s,a)\right]$$

- Computing the total derivative of $J(\theta,\, w^*(\theta))$ using the eqs. from 2.2:

$$\begin{aligned} \frac{dJ(\theta, w^*(\theta))}{d\theta} &= \frac{\partial J(\theta, w)}{\partial\theta} + \frac{dw^*(\theta)}{d\theta}\frac{\partial J(\theta, w)}{\partial w} \\ &= \frac{\partial J(\theta, w)}{\partial\theta} - \left(\frac{\partial^2 L(\theta, w)}{\partial\theta\, \partial w}\right)\left(\frac{\partial^2 L(\theta, w)}{\partial w^2}\right)^{-1}\frac{\partial J(\theta, w)}{\partial w} \end{aligned}$$

(Eq. 19)

- Computing the $\frac{\partial J(\theta,w)}{\partial\theta}$ term - computed by policy gradient theorem.
- Computing the $\frac{\partial^2 L(\theta,w)}{\partial w^2}$ term - computed by taking the direct derivative:

$$\frac{\partial J(\theta, w)}{\partial w} = E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[\frac{dQ_w(s,a)}{dw}\right]$$

- Computing the $\frac{\partial^2 L(\theta,w)}{\partial w^2}$ -

$$\begin{aligned} \frac{\partial^2 L(\theta, w)}{\partial w^2} &= E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[\frac{\partial}{\partial w^2}\left(Q_w(s,a) - Q^\pi(s,a)\right)^2\right] \\ &= E_{s\sim\rho,\, a\sim\pi_\theta(\cdot|s)}\left[2\frac{dQ_w(s,a)}{dw}\frac{dQ_w(s,a)}{dw}^T + 2\left(Q_w(s,a) - Q^\pi(s,a)\right)\frac{d^2 Q_w(s,a)}{dw^2}\right] \end{aligned}$$

- See Theorem 1 and Proposition 1 in the paper for how to compute the $\frac{\partial^2 L(\theta,w)}{\partial w\, \partial\theta}$ term.
- Once the terms have been computed, we can apply the Stackelberg gradient update:

$$\begin{aligned} \theta &\leftarrow \theta + \alpha_\theta \tfrac{dJ(\theta, w^*(\theta))}{d\theta} \\ w &\leftarrow w - \alpha_w \tfrac{\partial L(\theta, w)}{\partial w} \end{aligned}$$

- *Note:* In practice, in order to maintain best response in the inner level with an iterative optimization algorithm, a number of unrolling gradient steps of critic update are performed.
  (I'm not sure I understand this part)

3.1 - Hessian Regularization

- In Eq. 19 we compute the inverse of the critic hessian, $\frac{\partial^2 L(\theta,w)}{\partial w^2}$.
- **Problem:**
  - If the critic parameter w is not in a neighborhood of critical points, the Hessian matrix might be ill-conditioned, depending on L and Q_w.

- **Solution:**
  - To avoid this, STAC computes the inverse of a regularized Hessian, $\frac{\partial^2 L(\theta,w)}{\partial w^2} + \lambda I$.

- λ - parameter that controls the mix of Stackelberg and normal gradient update:
  - $\lambda \to \infty$ - eigenvalues become zero and the update reduces to a normal gradient update.
  - $\lambda \to 0$ - pure Stackelberg update.

4 - Experiments

- Performance is evaluated on the OpenAI gym platform.

- The only difference between STAC and AC are the update rules, as derived in Section 3.

- Metric - average episode return versus the time steps.

- Different actor-critic learning rates and k-steps are tested.

> *Key Idea* - the best performance overall is achieved by STAC with multiple critic unrolling steps.

- $\lambda = 0$ for cartpole.

- $\lambda = 500$ for Reacher, Hopper, and Walker2D.
  - This is likely because these are more complex games.
  - This results in STAC performance closer to that of AC.

5 - Discussion and Future Work

- STAC could be extended to a general Stackelberg learning meta-framework for any actor-critic based method.

- Experiment with switching the actor and the critic order.
  - When the *actor* is the *leader*, we get generalized policy iteration.
  - When the *critic* is the *leader*, we get more value-based methods.

- Experiment with decaying λ.
  - Less regularization should be needed as the learning approaches the neighborhood of the equilibrium.