Closing the Gap: Achieving Global Convergence (Last Iterate) of Actor Critic under Markovian Sampling with Neural Network Parameterization

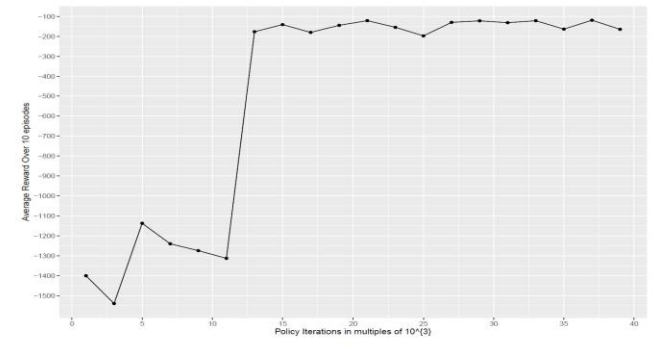
Mudit Gaur[1], Amrit Singh Bedi[2], Di Wang[3], Vaneet Aggarwal[1]
[1]Purdue University, [2]University of Central Florida, [3]KAUST

Problem Overview

- **M**ulti-Layer Neural Network representation for the critic
  - Some results use linear function to represent critic which makes them unsuitable for real world use
- Markovian Sampling Assumption
  - Some results assume samples from the MDP are independent. In practice samples are drawn from a Markov chain
- Continuous Action Spaces
  - All existing convergence results with neural network critic are restricted to finite action spaces
- Global Convergence
  - To establish global convergence, all existing results use natural policy gradient version of the algorithm, which requires calculating hessians. This is rarely done in practice.

Problem Overview (Cont.)

- Last Iterate Convergence
  - A typical convergence result of a practical implementation of actor critic looks like the following



  - To explain this result, last iterate of policy parameter should be shown close to optimal.

Algorithm Pseudocode

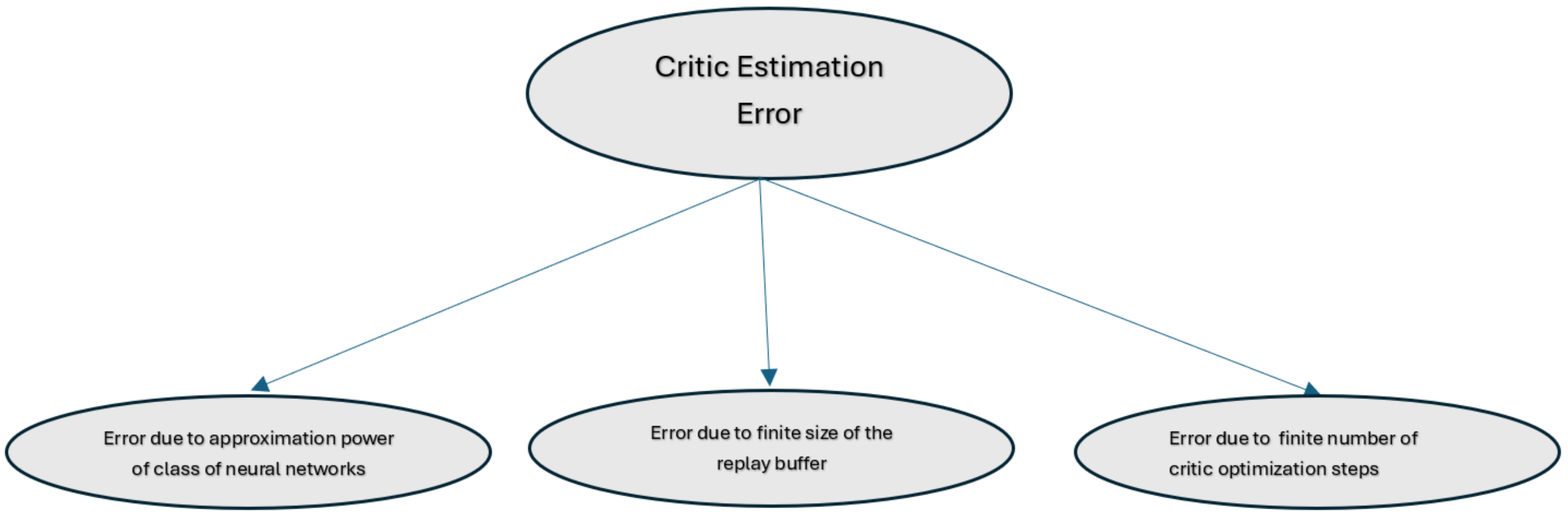
- Randomly initialize policy parameter
- Run the Actor loop for *K* iterations
  - Collect *n* samples and store in buffer using current policy iteration
  - Run critic loop for *J* iterations
    - Perform gradient ascent step by sampling from buffer
  - Perform actor update using policy gradient theorem using estimate of critic obtained at end of previous loop

Note: The version of the algorithm we are using is like the Twin Delayed DDPG (Dankwa et al. 2019) which has been shown to be effective for continuous action setups.

Challenges and their Solutions

- **Multi Layer Neural Network Critic Representation and Markovian Sampling**
- **Associated Challenges:**
  - For a non-linear critic, the TD update devised in Sutton 1992 for linear critic and the corresponding convergence results can no longer be used
  - Existing analyses for non-linear critic require i.i.d sampling assumptions
- **Solution**
  - We decompose the error incurred in estimating the critic into three components as follows

Challenges and their Solutions (cont.)



- The left most term accounts for the multi layer neural network structure
- The middle term accounts for the Markovian sampling and finite buffer size
- The right most term accounts for the error incurred due to the finite number of steps in the critic loop

Challenges and their Solutions (cont.)

- **Last Iterate and Global Convergence , Continuous Action Space**
- **Associated Challenges:**
  - Since multi layer neural networks are non-convex, convergence on the term $J(\lambda^*) - J(\lambda^K)$ is not possible without further structural assumptions.
  - Instead works with neural network critic prove an upper bound on the regret given by $$\frac{1}{K}\sum_{i=1}^{K} (J(\lambda^*) - J(\lambda^i)).$$
  - This is done by using the smoothness property of the average expected return.
  - Techniques leveraging smoothness are restricted to finite action spaces as the upper bounds involve cardinality of the action space.
  - Results leveraging smoothness also rely on natural actor critic involving calculation of Hessians.

Challenges and their Solutions (cont.)

- **Solution**
  - We exploit the Polyak-Lojasiewicz (PL) property of the MDP (proved in Ding et al. 2022) to obtain a global convergence of the last iterate for vanilla actor critic. The PL property is given as

$$\sqrt{\mu}\,(J(\lambda^*) - J(\lambda)) \leq ||\nabla J(\lambda)|| + \epsilon'$$

  - This property allows us to obtain the following result

$$J(\lambda^*) - J(\lambda^K) \leq \tilde{O}\left(\frac{1}{K}\right) + \frac{1}{K}\sum_{i=1}^{K}\left(|Q^i - Q^{\lambda_i}|\right) + \varepsilon_{bias}$$

- The second term on the right-hand side is the cumulative error in critic estimation till iteration *K* of the algorithm
- ϵ_{bias} is a measure of how compatible are the actor and critic networks

Final Results

$$J(\lambda^*) - J(\lambda^K) \leq \tilde{O}\left(\frac{1}{K}\right) + \tilde{O}\left(\frac{1}{\sqrt{n}}\right) + \tilde{O}(\gamma^J) + \varepsilon_{error} + \varepsilon_{bias}$$

- The resulting sample complexity of ϵ^{-3} for global convergence matches the best sample complexity obtained even for the linear critic case obtained in Xu et al. 2020.
- Our result holds for the vanilla actor critic and not natural policy gradient, as is done in all other global convergence results.

Final Results (cont.)

Work	MMCGL	Sample Complexity
This work	✓	ϵ^{-3}
Fu et al. 2021	M, C, L	ϵ^{-6}
Cayci et al. 2022	M, **M**, C, L	ϵ^{-4}
Xu et al. 2020	**M**, C, L	ϵ^{-3}
Tian et al. 2023	**M**, C, L , **G**	ϵ^{-2}