Inherent Overestimation of DRL-Based Hybrid Beamforming for mmWave MIMO Systems: Behavioral Interpretation and Remedies

Dohyun Kim
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712, United States
Email: dohyun.kim@utexas.edu

Robert W. Heath Jr.
Department of Electrical and Computer Engineering
North Carolina State University
Raleigh, NC 27695, United States
Email: rwheathjr@ncsu.edu

***Abstract*—Recently, many machine learning-based hybrid beamforming algorithms have been studied to implement the practical mmWave dense MIMO systems with high spectral efficiency. Hybrid beamforming algorithm based on deep reinforcement learning (DRL), is claimed to be the state-of-the-art technique regarding the computation time to achieve high spectral efficiency. Nonetheless, DRL is known to suffer from *overestimation*, which reinforces the algorithm to converge to a suboptimal behavior. Herein, we investigate overestimation in DRL-based hybrid beamforming using the angle representation of analog precoder. We discuss possible directions, based on the behavioral interpretation, to handle the overestimation.**

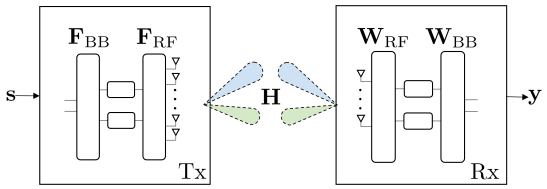
I. Introduction

Hybrid beamforming (HBF) enables the practical implementation of mmWave dense MIMO systems with high spectral efficiency [1]. Its effectiveness comes from the separation of the analog/digital domain, reducing the number of costly components such as converters between the analog/digital domain. Among a myriad of work applying the modern machine learning tools to HBF, in this paper, we focus on algorithms based on *deep reinforcement learning* (DRL). The benefit of DRL-based HBF is the short online computation time and robustness to channel estimation error [2].

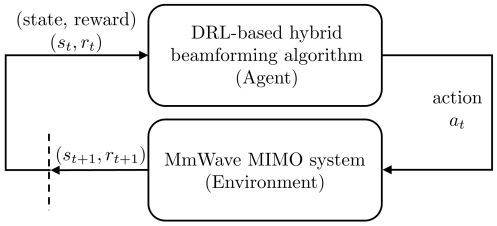
Overestimation in DRL is a well-known issue, that can make the algorithm converge to a suboptimal behavior. Moreover, overestimation is inherent and exists whenever the function estimator is imprecise [3]. Double Q-learning [4] is known as a ubiquitous solution for DRL with discrete states, but it is not suitable for DRL with continuous states which correspond to dense MIMO systems. Meanwhile, clipped double Q-learning [5] can handle DRL with continuous states.

To the best of the authors' knowledge, overestimation in DRL-based HBF has not been studied. Herein, we observe the overestimation behaviors of inherent states and discuss possible remedies.

The contribution of this paper are the following:



(a) Configuration of HBF for dense MIMO system



(b) Algorithm flow of DRL for HBF

Fig. 1: DRL-based HBF concepts: (a) Hardware, (b) Algorithm

- We properly observe and interpret the overestimation behavior of the inherent state in DRL-based HBF algorithms.
- We discuss remedies to the overestimation, providing experimental results on toy examples. We sketch to provide intuition towards an extension to practical problems.

II. Overestimation problem in DRL-based HBF

We implement an exemplary DRL-based HBF, similar to [2], with only one learning parameter $\mathbf{F}_{\mathrm{RF}}^{(t)}$. We consider a 128 by 16 MIMO system with 2 radio frequency chain and data stream in Figure 1a. We apply the learning model in Figure 1b with state $s_t = \{\mathbf{F}_{\mathrm{BB}}, \mathbf{F}_{\mathrm{RF}}^{(t-1)}, \mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}}\}$, action $a_t = \{\mathbf{F}_{\mathrm{BB}}, \mathbf{F}_{\mathrm{RF}}^{(t)}, \mathbf{W}_{\mathrm{RF}}, \mathbf{W}_{\mathrm{BB}}\}$, and reward r_t that corresponds to s_{t+1} and channel $\mathbf{H}$. We consider a narrowband channel model with $N_p = 2$ path clusters with angle of arrival vector $(0, \frac{\pi}{16})$ and angle of departure vec-

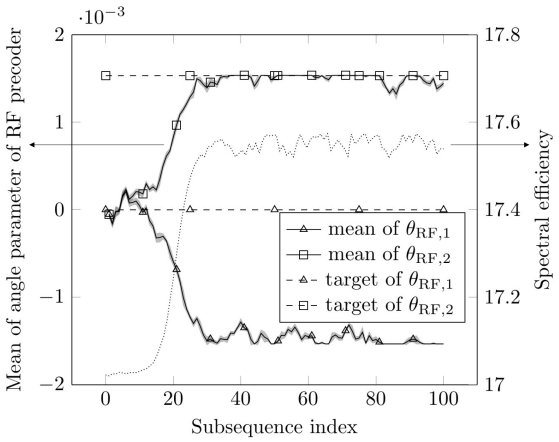


Fig. 2: Statistical interpretation of angle parameter $\boldsymbol{\theta}_{\mathrm{RF}}$ of RF precoder (per subsequence for selected SNR of 5dB and $\tau = 0.5 \cdot 10^{-3}$)

tor $(0, \frac{\pi}{128*16})$. We represent $\mathbf{F}_{\mathrm{RF}}^{(t)}$ with angle parameters $\theta_{\mathrm{RF},1}$ and $\theta_{\mathrm{RF},2}$ following the beam steering fashion [6].

Figure 2 depicts the overestimation issue in DRL-based HBF. The spectral efficiency staggers up to Subsequence 10, reaching almost 17. After Subsequence 15, we observe an increase in spectral efficiency, linearly up to Subsequence 30. Again, after Subsequence 40, we observe negligible oscillation. The means of angle parameter $\theta_{\mathrm{RF},1}$ and $\theta_{\mathrm{RF},2}$ are both zero at Subsequence 1. The mean of angle parameter $\theta_{\mathrm{RF},2}$ tends to increase up to Subsequence 30, where it lies near its target. However, the mean of angle parameter $\theta_{\mathrm{RF},1}$ does not converge to its target. It tends to decrease up to Subsequence 30, where it lies near $-1.5 \cdot 10^{-3}$ not converging it to its target of zero. We interpret that the overestimation of the state $\theta_{\mathrm{RF},1} = -1.5 \cdot 10^{-3}$ is the main source leading to suboptimal behavior. To be specific, the baseline starting with $\theta_{\mathrm{RF},1} = 0$, explores $\theta_{\mathrm{RF},1}$ around 0. Due to the imprecision of function estimator, the value of a negative $\theta_{\mathrm{RF},1}$ becomes higher than that of $\theta_{\mathrm{RF},1} = 0$. The overestimation of value induces a poor policy to select negative $\theta_{\mathrm{RF},1}$. The poor policy then results in a bad estimation of value. Overall, the overestimation accumulates throughout the recursive update of value [3]. We observe the accumulated error as a "drift" in $\theta_{\mathrm{RF},1}$, causing its tendency of decreasing.

III. Control of overestimation in DRL

Overestimation in DRL with discrete states can be improved by the use of separate networks, respectively for selecting and evaluating an action in the max operator of value updates [4]. Similarly, for continuous states, deep deterministic policy gradient (DDPG) separately trains target networks and online networks in an actor-critic learning fashion [7]. The target networks are delayed copies of online networks, where a parameter τ controls the delay. For stable learning, DDPG requires a small τ. The small τ, however, slows the change of target actor network, eventually making the target networks and online networks similar. Therefore, the practical use of DDPG needs further solution of overestimation.

On one hand, using the *minimum value estimate* of two separate networks is a quick remedy of overestimation in DDPG, at the cost of additional computation from the extra networks [5]. On the other hand, *multi-step bootstrapping* method [8] in the value estimate without additional networks introduces underestimation that needs further investigation in DRL-based HBF. Overall, the behavioral interpretation of DRL-based HBF using angle representation is interesting to observe the effect of remedies introduced by [5], [8].

IV. Conclusive Remark

Interpretation in DRL-based HBF is important, in the sense that it allows us to observe the behavioral details more than just its explicit performance. We illustrated the overestimation behavior in DRL-based HBF, with an exemplary implementation using angles, which is not explicit based on observed spectral efficiency. As a specific result, the poor "beam steering angle" behaviors accumulate overestimation errors, eventually lead to a suboptimal value of spectral efficiency throughout the learning process.

The behavioral interpretation, and remedies of the overestimation, of DRL-based HBF using angle representation make more margin of the tradeoff between computation time and spectral efficiency. Using the minimum value estimate and multi-step bootstrapping method can be further combined to control the overestimation.

Acknowledgment

This work was partially supported by the U.S. Army Research Labs under grant W911NF-19-1-0221.

References

[1] R. W. Heath *et al.*, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE journal of selected topics in signal processing*, vol. 10, no. 3, pp. 436–453, 2016.

[2] Q. Wang *et al.*, "Precodernet: Hybrid beamforming for millimeter wave systems with deep reinforcement learning," *IEEE Wireless Communi. Letters*, vol. 9, no. 10, pp. 1677–1681, 2020.

[3] S. Thrun and A. Schwartz, "Issues in using function approximation for reinforcement learning," in *Proc. Connectionist Models Summer School Hillsdale, NJ. Lawrence Erlbaum*, 1993.

[4] H. Van Hasselt *et al.*, "Deep reinforcement learning with double Q-learning," *arXiv preprint arXiv:1509.06461*, 2015.

[5] S. Fujimoto *et al.*, "Addressing function approximation error in actor-critic methods," *arXiv preprint arXiv:1802.09477*, 2018.

[6] O. El Ayach *et al.*, "The capacity optimality of beam steering in large millimeter wave mimo systems," in *IEEE Int. Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE, 2012, pp. 100–104.

[7] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," *arXiv:1509.02971*, 2015.

[8] L. Meng *et al.*, "The effect of multi-step methods on overestimation in deep reinforcement learning," *arXiv:2006.12692*, 2020.