Annals of the Institute of Statistical Mathematics (2025) 77:535–562
https://doi.org/10.1007/s10463-025-00924-1

Offline minimax Q-function learning for undiscounted indefinite-horizon MDPs

Fengying Li[1] · Yuqiang Li[2] · Xianyi Wu[2] · Wei Bai[1]

Received: 5 September 2023 / Revised: 6 November 2024 / Accepted: 20 December 2024 /
Published online: 21 April 2025

Abstract
This work considers the offline evaluation problem for indefinite-horizon Markov Decision Processes. A minimax Q-function learning algorithm is proposed, which, instead of i.i.d. tuples (s, a, s', r), evaluates undiscounted expected return based by i.i.d. trajectories truncated at a given time step. The confidence error bounds are developed. Experiments using Open AI's Cart Pole environment are employed to demonstrate the algorithm.

Keywords Indefinite-horizon MDPs · Off-policy · Minimax Q-function learning · Policy evaluation · Occupancy measure

This research is supported by National Key R&D Program of China (Nos. 2021YFA1000100 and 2021YFA1000101), Natural Science Foundation of China (No. 72371103), and Ningxia Natural Science Foundation (No. 2023AAC03342).

✉ Wei Bai
nx_bw@nxnu.edu.cn

Fengying Li
lifengying01@126.com

Yuqiang Li
yqli@stat.ecnu.edu.cn

Xianyi Wu
xywu@stat.ecnu.edu.cn

1 School of Mathematics and Computer Science, Ningxia Normal University, Guyuan 756000, People's Republic of China

2 School of Statistics, KLATASDS-MOE, East China Normal University, Shanghai 200062, People's Republic of China