View Meta-Reviews

Paper ID
361

Paper Title
Offline Reinforcement Learning: Role of State Aggregation and Trajectory Data

Track Name
COLT2024

META-REVIEWER #1

META-REVIEW QUESTIONS

1. Verdict on this paper
Accept

2. Brief Justification.
Reviewers have read the rebuttal and discussed. All three of reviewers are in favor of acceptance.

View Reviews

Paper ID

361

Paper Title

Offline Reinforcement Learning: Role of State Aggregation and Trajectory Data

Track Name

COLT2024

Reviewer #1

Questions

1. Please summarize the main claims and contributions of the paper (1-2 paragraphs).

This work settles a problem that was left open in the prior works of Xie and Jiang (2021) and Foster et al. (2022):
Can Offline Policy Evluation (OPE) be done using a number of samples that is poly(\log |F|, H, C) (F being the function class and C the concentrability coefficient) under value function realizability, and **natural offline data distribution such as admissible / trajectory data**?
The lower bound in Foster et al. (2022) relied on offline data distribution that is not generated by a behavior policy. The present paper proves a similar lower bound but with trajectory data, and establishes \Omega(2^H) samples are necessary.

This work further provides a characterization of the sample complexity in terms of "aggregated conecntrability coefficient" C'. In short, the way to the lower bound is by showing (1, Thm 2) The sample complexity of OPE is \Omega(C'); (2, Prop 3) The lower bound MDPs may have C=O(H^3) but C'=Omega(2^H) even with admissible data; (3; Thm 4) There is a generic reduction that converts hard instances with admissible data to trajectory data.

2. Describe the strengths and weaknesses of the work. Describe the significance and novelty of the contributions, relation to the prior work, clarity, and relevance to COLT.

The paper is well written, settles the issue regarding data admissibility that was left open in Foster et al. (2022), and further provides a charcterization in terms of the aggregated concentrability coefficient.
I don't see any significant weaknesses. The work is also highly relevant to COLT, of

course, as it studies fundamental limits of statistical efficiency in offline RL with function approximation.

4. Please provide questions for authors to address during the author feedback period.

(*) Would your results carry to the infinite horizon discounted setting?

Minor comments

(*) Page 2 “aggregated concealability coefficient” ---> "aggregated concentrability coefficient"

(*) Page 4 "The learner is given ... $X \times A \to [-1, 1]$" ---> $X \times A \to [-H, H]$

5. Please provide an overall score for the submission.

Accept

7. [Post rebuttal only] Check the box if you have read and discussed author’s feedback post rebuttal.

Yes

Reviewer #2

Questions

1. Please summarize the main claims and contributions of the paper (1-2 paragraphs).

This paper solves an open problem that if offline RL with polynomial sample complexity is possible with value function realizability and finite concentrability coefficient but without Bellman completeness. The authors provide a negative answer to this question for the task of offline policy evaluation. They reveal that the sample complexity of offline policy evaluation depends on a concept named aggregated concentrability coefficient, which can be exponentially large without stronger assumption. The lower bounds and upper bounds in this paper provide a complete picture for offline policy evaluation with only value function realizability.

2. Describe the strengths and weaknesses of the work. Describe the significance and novelty of the contributions, relation to the prior work, clarity, and relevance to COLT.

Strengths: this is a very interesting paper, and is well written. The idea in this paper significantly contributes to the field of offline reinforcement learning with function approximation.

Weakness: I didn’t find any significant weakness, just a few questions, see below.

4. Please provide questions for authors to address during the author feedback period.

There seems to be a typo on page 2, where you wrote “aggregate concealability coefficient” instead of “aggregate concentrability coefficient”.

I’m curious how do the insights and findings of this work (offline policy evaluation) correlate with or contribute to offline policy optimization?

This paper considers the setting where the state space is layered across time. How do the results in this work suggest/extend to the more general and commonly studied setting where stages share the same state space? I understand that the lower bound is still valid. How about the upper bound? Is the fact that the state space is layered across time essential in showing the upper bound?

On page 4, why does the function set $\mathcal{F}$ consist of functions of range $[-1,1]$? The realizability assumption assumes that the true value function, whose range is $[-H,H]$, is in the function class. So is the function class large enough to satisfy the realizability assumption? Or should the function set $\mathcal{F}$ consist of functions of range $[-H,H]$?

For the aggregated transition dynamics defined in (3), why does it depend on π and offline data distribution μ? Can you provide some intuition or a perspective on seeing the reasoning behind this definition?

5. Please provide an overall score for the submission.

Accept

7. [Post rebuttal only] Check the box if you have read and discussed author’s feedback post rebuttal.

Yes

Reviewer #3

Questions

1. Please summarize the main claims and contributions of the paper (1-2 paragraphs).

The paper addresses the issue of off-line RL, where we have accumulated data from one behavioral policy, and would like to evaluate a different policy.

The concentrability coefficient is a well know way to match the two. It is known that in the tabular setting the concentrability coefficient is the right measure. The paper looks at the function approximation case, where the class of function includes the Q-function of the evaluated policy.

The paper generalizes the concentrability coefficient to an aggregated concentrability coefficient, where, conceptually, one can cluster states. It shows that the new measure can be exponentially different than the previous one, and it captures the correct sample complexity.

Compared to previous works, the main additions is that it uses a data generation which is due to one policy (and not just an arbitrary distribution) and addresses the issue of trajectory feedback.

2. Describe the strengths and weaknesses of the work. Describe the significance and novelty of the contributions, relation to the prior work, clarity, and relevance to COLT.

PRO

* The requirement that the distribution can be generated by some policy is natural and important
* The understanding of the trajectory data is important and relevant

CON

* The algorithm is similar to previous work (as the autor say). This is not a serious issue since the main contributions are the lower bounds.

5. Please provide an overall score for the submission.

Weak accept

7. [Post rebuttal only] Check the box if you have read and discussed author's feedback post rebuttal.

Yes