Bridging Offline Reinforcement Learning and Imitation Learning: A Tale of Pessimism

Paria Rashidinejad, Banghua Zhu, *Graduate Student Member, IEEE*, Cong Ma, Jiantao Jiao, *Member, IEEE*, and Stuart Russell

***Abstract*— Offline reinforcement learning (RL) algorithms seek to learn an optimal policy from a fixed dataset without active data collection. Based on the composition of the offline dataset, two main methods are used: imitation learning which is suitable for expert datasets, and vanilla offline RL which often requires uniform coverage datasets. From a practical standpoint, datasets often deviate from these two extremes and the exact data composition is usually unknown. To bridge this gap, we present a new offline RL framework, called *single-policy concentrability*, that smoothly interpolates between the two extremes of data composition, hence unifying imitation learning and vanilla offline RL. Under this new framework, we ask: can one develop an algorithm that achieves a minimax optimal rate adaptive to unknown data composition? To address this question, we consider a lower confidence bound (LCB) algorithm developed based on pessimism in the face of uncertainty in offline RL. We study finite-sample properties of LCB as well as information-theoretic limits in multi-armed bandits, contextual bandits, and Markov decision processes (MDPs). Our analysis reveals surprising facts about optimality rates. In particular, in both contextual bandits and RL, LCB achieves a fast convergence rate for nearly-expert datasets, analogous to the one achieved by imitation learning, contrary to the slow rate achieved in offline RL. In contextual bandits, we prove that LCB is adaptively optimal for the entire data composition range, achieving a smooth transition from imitation learning to offline RL. We further show that LCB is almost adaptively optimal in MDPs.**

***Index Terms*— Reinforcement learning theory, offline reinforcement learning, imitation learning, lower confidence bound (LCB), adaptive optimality.**

I. Introduction

REINFORCEMENT learning (RL) algorithms have recently achieved tremendous empirical success including beating Go champions [1], [2] and surpassing professionals in Atari games [3], [4], to name a few. Most success stories, however, are in the realm of online RL in which active data collection is necessary. This online paradigm falls short of leveraging previously-collected datasets and dealing with scenarios where online exploration is not possible [5]. To tackle these issues, offline (or batch) reinforcement learning [6], [7] arises in which the agent aims at achieving competence by exploiting a batch dataset without access to online exploration. This paradigm is useful in a diverse array of application domains such as healthcare [8]–[10], autonomous driving [11]–[13], and recommendation systems [14]–[16].

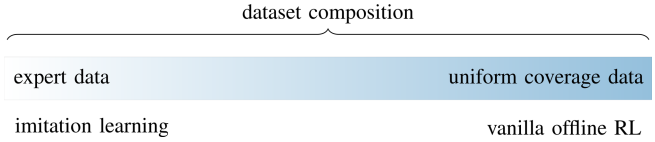

Fig. 1. Dataset composition range for offline RL problems. On one end, we have expert data for which imitation learning algorithms are well-suited. On the other end, we have uniform exploratory data for which vanilla offline RL algorithms can be used.

The key component of offline RL is a pre-collected dataset from an unknown stochastic environment. Broadly speaking, there exist two types of *data composition* for which offline RL algorithms have shown promising empirical and theoretical success; see Figure 1 for an illustration.

- **Expert data.** One end of the spectrum includes datasets collected by following an expert policy. For such datasets, imitation learning algorithms (e.g., behavior cloning [17]) are shown to be effective in achieving a small sub-optimality competing with the expert policy. In particular, it is recently shown in the work [18] that the behavior cloning algorithm achieves the minimal sub-optimality $1/N$ in episodic Markov decision processes, where N is the total number of samples in the expert dataset.
- **Uniform coverage data.** On the other end of the spectrum lies the datasets with uniform coverage. More specifically, such datasets are collected with an aim to cover *all* states and actions, even the states never visited or actions never taken by satisfactory policies. Most vanilla offline RL algorithms are only suited in this region and are shown to diverge for *narrower* datasets [5], [19], such as those collected via human demonstrations or hand-crafted policies, both empirically [20], [21] and theoretically [22], [23]. In this regime, a widely-adopted requirement is the *uniformly bounded*

Manuscript received 18 January 2022; accepted 5 June 2022. Date of publication 22 June 2022; date of current version 22 November 2022. The work of Paria Rashidinejad was supported in part by the Open Philanthropy Foundation and in part by the Leverhulme Trust. The work of Banghua Zhu and Jiantao Jiao was supported in part by NSF under Grant IIS-1901252, Grant CCF-1909499, and Grant DMS-2023505. *(Corresponding author: Cong Ma.)*

Paria Rashidinejad, Banghua Zhu, and Stuart Russell are with the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley (UC Berkeley), Berkeley, CA 94720 USA.

Cong Ma is with the Department of Statistics, The University of Chicago, Chicago, IL 60637 USA (e-mail: congma2015@gmail.com).

Jiantao Jiao is with the Department of Electrical Engineering and Computer Sciences and the Department of Statistics, University of California at Berkeley (UC Berkeley), Berkeley, CA 94720 USA.

Communicated by A. Krishnamurthy, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3185139.

Digital Object Identifier 10.1109/TIT.2022.3185139

concentrability coefficient which assumes that the ratio of the state-action occupancy density induced by *any policy* and the data distribution is bounded uniformly over all states and actions [24]–[27]. Another common assumption is uniformly lower bounded data distribution on all states and actions [28], [29], which ensures all states and actions are visited with sufficient probabilities. Algorithms developed for this regime are demonstrated to achieve a $1/\sqrt{N}$ sub-optimality competing with the optimal policy; see for example the papers [30]–[32].

A. Motivating Questions

Clearly, both of these two extremes impose strong assumptions on the dataset: at one extreme, we hope for a solely expert-driven dataset; at the other extreme, we require the dataset to cover every, even sub-optimal, actions. In practice, there are numerous scenarios where the dataset deviates from these two extremes, which has motivated the development of new offline RL benchmark datasets with different data compositions [5], [19]. With this need in mind, the first and foremost question is regarding offline RL formulations:

Question 1 (Formulation): Can we propose an offline RL framework that accommodates the entire data composition range?

We answer this question affirmatively by proposing a new formulation for offline RL that smoothly interpolates between two regimes: expert data and data with uniform coverage. More specifically, we characterize the data composition in terms of the ratio between the state-action occupancy density of an optimal policy[1] and that of the behavior distribution which we denote by $C^\star$; see Definition 1 for a precise formulation. In words, $C^\star$ can be viewed as a measure of the deviation between the behavior distribution and the distribution induced by the optimal policy. The case with $C^\star = 1$ recovers the setting with expert data since, by the definition of $C^\star$, the behavior policy is identical to the optimal policy. In contrast, when $C^\star > 1$, the dataset is no longer purely expert-driven: it could contain "spurious" samples—states and actions that are not visited by the optimal policy. As a further example, when the dataset has uniform coverage, say the behavior probability is lower bounded by $\mu_{\min}$ over all states and actions, it is straightforward to check that the new concentrability coefficient is also upper bounded by $\mu_{\min}^{-1}$.

Assuming a finite $C^\star$ is the weakest concentrability requirement [27], [33], [34] that is currently enjoyed only by some online algorithms such as CPI [35]. $C^\star$ imposes a much weaker assumption in contrast to other concentrability requirements which involve taking a maximum over all policies; see [33] for a hierarchy of different concentrability definitions. We would like to immediately point out that existing works on offline RL either do not specify the dependency of sub-optimality on data coverage [36], [37], or do not have a batch data coverage assumption that accommodates the entire data spectrum including the expert datasets [38], [39].

[1]In fact, our developments can accommodate arbitrary competing policies, however, we restrict ourselves to the optimal policy for ease of presentation.

With this formulation in mind, a natural next step is designing offline RL algorithms that handle various data compositions, i.e., for all $C^\star \geq 1$. Recently, efforts have been made toward reducing the offline dataset requirements based on a shared intuition: the agent should act conservatively and avoid states and actions less covered in the offline dataset. Based on this intuition, a variety of offline RL algorithms are proposed that achieve promising empirical results. Examples include model-based methods that learn pessimistic MDPs [37], [39], [40], model-free methods that reduce the Q-functions on unseen state-action pairs [41]–[43], and policy-based methods that minimize the divergence between the learned policy and the behavior policy [20], [21], [44]–[49].

However, it is observed empirically that existing policy-based methods perform better when the dataset is nearly expert-driven (toward the left of data spectrum in Figure 1) whereas existing model-based methods perform better when the dataset is randomly-collected (toward the right of data spectrum in Figure 1) [37], [50]. It remains unclear whether a single algorithm exists that performs well regardless of data composition—an important challenge from a practical perspective [5], [19], [51]. More importantly, the knowledge of the dataset composition may not be available a priori to assist in selecting the right algorithm. This motivates the second question on the algorithm design:

Question 2 (Adaptive Algorithm Design): Can we design algorithms that can achieve minimal sub-optimality when facing different dataset compositions (i.e., different $C^\star$)? Furthermore, can this be achieved in an adaptive manner, i.e., without knowing $C^\star$ beforehand?

To answer the second question, we analyze a *pessimistic* variant of a value-based method in which we first form a lower confidence bound (LCB) for the value function of a policy using the batch data and then seek to find a policy that maximizes the LCB. A similar algorithm design has appeared in the recent work [36]. It turns out that such a simple algorithm—fully agnostic to the data composition—is able to achieve *almost* optimal performance in multi-armed bandits and Markov decision processes, and optimally solve the offline learning problem in contextual bandits. See the section below for a summary of our theoretical results.

B. Main Results

In this subsection, we give a preview of our theoretical results; see Table I for a summary. Under the new framework defined via $C^\star$, we instantiate the LCB approach to three different decision-making problems with increasing complexity: (1) multi-armed bandits, (2) contextual bandits, and (3) infinite-horizon discounted Markov decision processes. We will divide our discussions on the main results accordingly. Throughout the discussion, N denotes the number of samples in the batch data, S denotes the number of states, and we ignore the log factors.

1) Multi-Armed Bandits: To address the offline learning problem in multi-armed bandits, LCB starts by forming a lower confidence bound—using the batch data—on the mean reward associated with each action and proceeds to select

TABLE I
A SUMMARY OF OUR THEORETICAL RESULTS WITH ALL THE LOG FACTORS IGNORED

Multi-armed bandits	$C^\star \in [1,2)$	$C^\star \in [2,\infty)$
Algorithm 1 (MAB-LCB) sub-optimality (Theorem 1)	$\sqrt{\frac{C^\star}{N}}$	$\sqrt{\frac{C^\star}{N}}$
Information-theoretic lower bound (Theorem 2)	$\exp\left(-(2-C^\star)\cdot\log\left(\frac{2}{C^\star-1}\right)\cdot N\right)$	$\sqrt{\frac{C^\star}{N}}$
Most played arm (Proposition 2)	$\exp\left(-N\cdot\mathsf{KL}\left(\mathrm{Bern}\left(\frac{1}{2}\right)\|\mathrm{Bern}\left(\frac{1}{C^\star}\right)\right)\right)$	N/A
Contextual bandits	$C^\star \in [1,\infty)$	
Algorithm 2 (CB-LCB) sub-optimality (Theorem 4)	$\sqrt{\frac{S(C^\star-1)}{N}}+\frac{S}{N}$	
Information-theoretic lower bound (Theorem 5)	$\sqrt{\frac{S(C^\star-1)}{N}}+\frac{S}{N}$	
Markov decision processes	$C^\star \in [1,1+1/N)$	$C^\star \in [1+1/N,\infty)$
Algorithm 3 (VI-LCB) sub-optimality (Theorem 6)	$\frac{S}{(1-\gamma)^4 N}$	$\sqrt{\frac{SC^\star}{(1-\gamma)^5 N}}$
Information-theoretic lower bound (Theorem 7)	$\sqrt{\frac{S(C^\star-1)}{(1-\gamma)^3 N}}+\frac{S}{(1-\gamma)^2 N}$	$\sqrt{\frac{S(C^\star-1)}{(1-\gamma)^3 N}}+\frac{S}{(1-\gamma)^2 N}$

the one with the largest LCB. We show in Theorem 1 that LCB achieves a $\sqrt{C^\star/N}$ sub-optimality competing with the optimal action for all $C^\star \geq 1$. It turns out that LCB is adaptively optimal in the regime $C^\star \in [2,\infty)$ in the sense that it achieves the minimal sub-optimality $\sqrt{C^\star/N}$ without the knowledge of the $C^\star$; see Theorem 2. We then turn to the case with $C^\star \in [1,2)$, in which the optimal action is pulled with more than probability $1/2$. In this regime, it is discovered that the optimal rate has an exponentially decays with N, i.e., $\exp(-N)$, and is achieved by the naive algorithm of selecting the most played arm (cf. Theorem 2). To complete the picture, we also prove in Theorem 3 that LCB cannot be adaptively optimal for all ranges of $C^\star \geq 1$ if the knowledge of $C^\star$ range is not available.

At first glance, it may seem that LCB for offline RL mirrors upper confidence bound (UCB) for online RL by simply flipping the sign of the bonus. However, our results reveal that the story in the offline setting is much more subtle than that in the online case. Contrary to UCB that achieves optimal regret in multi-armed bandits [52], LCB is provably *not* adaptively optimal for solving offline bandit problems under the $C^\star$ framework.

2) Contextual Bandits: The LCB algorithm for contextual bandits shares a similar design to that for multi-armed bandits. However, the performance upper and lower bounds are more intricate and interesting when we consider contextual bandits with at least two states. With regards to the upper bound, we show in Theorem 4 that LCB exhibits two different behaviors depending on the data composition $C^\star$. When $C^\star \geq 1+S/N$, LCB enjoys a $\sqrt{S(C^\star-1)/N}$ sub-optimality, whereas when $C^\star \in [1,1+S/N)$, LCB achieves a sub-optimality with the rate S/N; see Figure 2(b) for an illustration. The latter regime ($C^\star \approx 1$) is akin to the imitation learning case where the batch data is close to the expert data. LCB matches the performance of behavior cloning for the extreme case $C^\star = 1$. In addition, in the former regime ($C^\star \geq 1+S/N$), the performance upper bound depends on the data composition through $C^\star - 1$, instead of $C^\star$. This allows the rate of sub-optimality to smoothly transition from $1/N$ to $1/\sqrt{N}$ as $C^\star$ increases. More importantly, both rates are shown to be minimax optimal in Theorem 3, hence confirming the adaptive optimality of LCB for solving offline contextual bandits—in stark contrast to the bandit case. On the other hand, this showcases the advantage of the $C^\star$ framework as it provably interpolates the imitation learning regime and the (non-expert) offline RL regime.

On a technical front, to achieve a tight dependency on $C^\star - 1$, a careful decomposition of the sub-optimality is necessary. In Section IV-C, we present the four levels of decomposition of the sub-optimality of LCB that allow us to accomplish the goal. The key message is this: the sub-optimality is incurred by both the value difference and the probability of choosing a sub-optimal action. A purely value-based analysis falls short of capturing the probability of selecting the wrong arm and yields a $1/\sqrt{N}$ rate regardless of $C^\star$. In contrast, the decomposition laid out in Section IV-C delineates the cases in which the value difference (or the probability of choosing wrong actions) plays a bigger role.