When Can Sequence Modelling Approaches Recover the Target Policy In Offline Reinforcement Learning? a Statistical Analysis

Abdelghani Ghanem[⋆†], Philippe Ciblat[†], and Mounir Ghogho[‡]
[⋆] TICLab, College of Engineering and Architecture,
International University of Rabat, Morocco
[†] LTCI, Department of Image, Data, Signal,
Telecom Paris, Institut Polytechnique de Paris, France
[‡] College of Computing,
University Mohammed VI Polytechnic, Morocco

***Abstract*—We present a theoretical analysis of sample complexity for learning the target policy in offline reinforcement learning (RL) using sequence modeling approaches. Our main theorem establishes bounds on the minimum required number of high-return samples. We identify distinct small-data and large-data regimes, characterized by a critical transition point, and reveal a potential trade-off between context coverage breadth and sampling depth. These findings offer insights into efficient data collection strategies and algorithm design for offline RL.**

***Index Terms*—Offline Reinforcement Learning, Sequence Modelling, Sample Complexity Analysis**

I. Introduction

Offline reinforcement learning (RL) addresses the challenge of learning effective policies from fixed datasets without online interaction with the environment [1], [2]. This paradigm is particularly relevant in domains such as robotics, logistics, and operations research, where exploration with untrained policies is impractical or unsafe. The offline RL setting has prompted the development of various approaches, initially focusing on adapting classical RL algorithms. These algorithms, such as off-policy methods [3], [4], were primarily designed for the online paradigm—a fundamentally different setting where the agent can interact with and learn from the environment in real-time. These adaptations typically incorporate mechanisms to mitigate action distribution shift while pursuing policy improvement [1], [5]–[8]. The goal is to learn an optimal policy that maximizes expected return, leveraging the information contained in the offline dataset, which often contains data from multiple policies or training stages. However, off-policy methods in offline RL settings are known for their sensitivity to hyperparameters and lack a theoretical basis for selecting among different distribution shift mitigation strategies [6], [7].

The limitations of classical RL methods in offline settings have motivated a shift towards framing offline RL as a supervised learning problem [9], [10]. Leveraging the inherently sequential nature of offline RL datasets, sequence modeling (SM) approaches have emerged as a promising direction [11], [12]. These methods offer several advantages over their off-policy counterparts, including algorithmic simplicity, reduced sensitivity to hyperparameters, and inherent resilience to action distribution shift [11]. Unlike off-policy techniques that aim to directly learn an optimal policy, SM approaches model the entire conditional distribution of policies present in the dataset, typically using transformer architectures [13], [14]. This comprehensive modeling approach, while powerful, introduces unique challenges. By capturing the full spectrum of policies, including suboptimal ones, these methods may be more susceptible to the influence of poor-quality data. During inference, the best policy is extracted by conditioning on the most high return contexts, but the success of this process heavily depends on the composition of the training dataset. The presence of suboptimal policies in the data could potentially hinder the extraction of truly optimal behavior, a challenge that is less pronounced in off-policy methods that explicitly target the best possible policy. This characteristic of SM approaches underscores the critical need for a thorough understanding of how dataset composition affects the quality of the extracted policy.

In this work, we address this challenge by providing a theoretical framework for analyzing the sample complexity of learning the target policy in offline RL using SM approaches. The analysis yields a novel bound on the required number of high-return samples, expressed in terms of the minimum number of samples and the expected minimum proportion of high-return data across contexts. This formulation allows for the characterization of the relationship between sample complexity and dataset composition, revealing distinct small-data and large-data regimes. A critical transition point between these regimes is identified and analyzed, providing insights into the diminishing returns of increasing dataset size beyond this point. The theoretical results suggest a fundamental trade-off between the breadth of context coverage and the depth of sampling within each context. This analysis may inform data collection strategies and algorithm design in offline RL, particularly in scenarios with imbalanced or limited data across different contexts.

II. Related Work

The study of sample complexity in reinforcement learning has a rich history, with seminal works establishing bounds for various settings [15], [16]. In the context of offline RL, recent research has focused on the challenges of distribution shift and policy constraint [5], [17]. SM approaches to RL, while relatively new, have shown promising empirical results [11], [18]. These methods draw inspiration from advances in natural language processing, particularly the use of transformer architectures [13]. Our work bridges the gap between these empirical advances and theoretical foundations, building on techniques from statistical learning theory [19].

III. System Model

We cast offline RL as a sequence modeling problem within a Markov Decision Process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma)$, where $\mathcal{S}$, $\mathcal{A}$, $\mathcal{P}$, R, and γ denote the state space, action space, transition probability function, reward function, and discount factor, respectively. We assume discrete data, noting that continuous spaces can be addressed through discretization techniques [18]

The core of our analysis revolves around a fixed-size static training dataset $\mathcal{T}$, comprising trajectories generated by K distinct *unknown* policies $\{\pi_k\}_{k=1}^K$. We transform $\mathcal{T}$ into a sequence modeling dataset $\mathcal{D} = \{(x_l, y_l)\}_{l=1}^N$, where $x_l \in \mathcal{X}$ represents the context (e.g., previous states, actions, returns) and $y_l \in \mathcal{Y}$ represents the next token (typically an action). The vocabulary size $V = |\mathcal{Y}|$ corresponds to the number of possible actions, while $C = |\mathcal{X}|$ denotes the number of possible contexts. To enhance interpretability, our approach prioritizes actions as tokens, though the methodology generalizes to vocabularies that incorporate state and return tokens. Additionally, we adopt returns —defined as the cumulative sum of future rewards— as our primary reward metric. This choice is made without loss of generality, as alternative metrics, such as Monte Carlo value estimates [18], are also applicable.

To characterize the dataset composition, we define α_k as the *expected* proportion of samples in $\mathcal{D}$ generated by policy π_k, ensuring $\sum_{k=1}^K \alpha_k = 1$. We distinguish between high-return and low-return contexts, denoting $\mathcal{X}^h \subset \mathcal{X}$ as the set of high-return contexts and $\mathcal{X}^l = \mathcal{X} \setminus \mathcal{X}^h$ as the set of low-return contexts, with $C^h = |\mathcal{X}^h|$. For each policy π_k, we decompose $\alpha_k = \alpha_k^h + \alpha_k^l$, where α_k^h and α_k^l represent the proportions of high-return and low-return data, respectively. The overall expected proportions of high-return and low-return data are denoted as $\alpha^h := \sum_{k=1}^K \alpha_k^h$ and $\alpha^l := \sum_{k=1}^K \alpha_k^l$.

For a context $c \in \mathcal{X}^h$, let $N_c = N_c^h + N_c^l$ denote the number of samples in $\mathcal{D}$ containing c, where N_c^h and N_c^l represent the numbers of high-return and low-return samples, respectively. For $c \in \mathcal{X}^l$, we have that $N_c = N_c^l$. This decomposition is crucial, as even if a context c is in $\mathcal{X}^h$, not all of its occurrences in the dataset are necessarily optimal. This is particularly evident in episodic environments where trajectories near the end timesteps may have the same return, but the chosen actions might be sub-optimal depending on the policy. The illustrative example below provides further clarification on this point.

Illustrative Example: Consider a 5x5 grid world where an agent must navigate from a start position to a goal. We define high-return trajectories as those reaching the goal in 10 steps or fewer. Let context A represent the agent's position one step away from the goal, and context B represent the starting position. Context A is classified as a high-return context ($A \in \mathcal{X}^h$) due to its proximity to the goal. Consider a dataset $\mathcal{T}$ which contains the following two trajectories:

- τ_1: $B \to \cdots \to A \to$ Goal (10 steps, high-return)
- τ_2: $B \to \cdots \to A \to$ [suboptimal actions] $\to$ Goal (20 steps, low-return)

In this example, $N_A = 2$ (total occurrences of context A), while $N_A^h = 1$ (occurrences of A in high-return trajectories). Thus, $N_A \neq N_A^h$ despite $A \in \mathcal{X}^h$. This discrepancy arises because context A's classification as high-return is based on its potential for high returns, but the actual returns depend on subsequent actions in the trajectory. □

Finally, we use $\beta_c^h = \frac{N_c^h}{N}$ to denote the estimated proportion of high-return samples in which context c appeared. By design, we have that $\mathbb{E}\left[\sum_{c \in \mathcal{X}^h} \beta_c^h\right] = \alpha^h$.

IV. Statistical Trajectory Model

We define each policy π_k as a conditional probability distribution over actions given contexts. Specifically, $\pi_k(v|c)$ represents the probability of taking action v given context c under policy k. For each context $c \in \mathcal{X}$, $\pi_k(\cdot|c)$ forms a probability distribution over the action vocabulary $[V]$, satisfying $\sum_{v=1}^V \pi_k(v|c) = 1$. Subsequently, we define the behavior policy π as a mixture of π_k and the target policy π^* for offline RL as follows:

$$\pi = \sum_{k=1}^K \alpha_k \pi_k; \qquad \pi^* = \frac{1}{\alpha^h} \sum_{k=1}^K \alpha_k^h \pi_k$$

Note that π^* uses α_k^h as coefficients, distinguishing it from the behavior policy π. An effective offline RL algorithm should aim to approximate π^*. Off-policy methods attempt this by directly modeling the policy with the maximum Q-value, effectively targeting π^*. In contrast, SM approaches model π, but attempt to recover π^* during inference by conditioning on contexts $c \in \mathcal{X}^h$ at each timestep. This approach implicitly assumes that high-return contexts are predominantly generated by π^*, theoretically allowing its recovery.

We define our learned model p as an estimate of the behavior policy, where $p(v|c) = \frac{1}{N_c} \sum_{l=1}^{N_c} X_l^{c,v}$, and $X_l^{c,v}$ are indicators of the occurrence of the pair (c, v) given c for the l-th sample in $\mathcal{D}$. We assume that p is an unbiased estimator of the true underlying distribution π, i.e. $\mathbb{E}[p] = \pi$. Consequently, $X_l^{c,v}$ can be interpreted as a Bernoulli random variable, with the probability of $X_l^{c,v} = 1$ given by the conditional probability $\pi(v|c)$. Additionally, to simplify our theoretical analysis, we assume that the $X_l^{c,v}$'s are independent across different samples. It can be shown that the empirical conditional distribution p is obtained by minimizing the known

categorical cross-entropy loss [20]. For theoretical analysis, we assume a simplified model where the *canonical* vectors of context-action pairs are directly used. In practice, these would be derived from a transformer model [13].

V. Problem Statement

Given a dataset with a minimum number of samples generated by the behavior policy for any high-return context, our goal is to find a lower bound on the minimum number of high-return samples needed so that our learned model approximates the target policy in offline RL.

Formally, let $\nu^h_{min} := \beta^h_{min} N_{min}$ be the minimum number of samples generated by π for any $c \in \mathcal{X}^h$ in $\mathcal{D}$, where $\beta^h_{min} = \min_{c\in\mathcal{X}^h} \mathbb{E}[\beta^h_c]$, and $N_{min} = \min_{c\in\mathcal{X}^h} N_c$. Our goal is to find a lower bound on ν^h_{min} such that our learned conditional model p approximates π^* on all $c \in \mathcal{X}^h$. To achieve this, we consider the matrix representations of p and π^*:

$$\mathbf{p} = (p(v|c))_{c\in\mathcal{X},v\in[V]} \in [0,1]^{C\times V} \tag{1}$$

$$\boldsymbol{\pi}^* = (\pi^*(v|c))_{c\in\mathcal{X},v\in[V]} \in [0,1]^{C\times V} \tag{2}$$

In order to measure the approximation error, we use the 1-norm:

$$\|\boldsymbol{p} - \boldsymbol{\pi}^*\|_1 = \sum_{c\in\mathcal{X}^h}\sum_{v\in\mathcal{Y}} |p(v|c) - \pi^*(v|c)| \tag{3}$$

Throughout the rest of the paper, we use $\sum_{c,v}$ to denote this double summation for brevity. It is worth noting that the 1-norm is twice the total variation distance $\|\boldsymbol{p}-\boldsymbol{\pi}^*\|_{\mathrm{TV}}$, another commonly used metric for probability distributions [20], [21]. Additionally, we compare π^* and p only on contexts from $\mathcal{X}^h$, which delineates a key flexibility of offline RL compared to scenarios of *offline imitation learning* (i.e., BC) [22] where we would require the model to approximate π^* on all $c \in \mathcal{X}$.

VI. Main Results

This section presents our main theoretical results, establishing sample complexity bounds for learning near-optimal policies in offline RL using SM approaches, and analyzes the implications of these bounds across different data regimes.

Theorem (Sample Complexity Bound)**.** *For any $\epsilon > 0$, if ν^h_{min} satisfies:*

$$\nu^h_{min} \geq \max\left\{\beta^h_{min}\left(\frac{C^h V}{\epsilon}\right)^2, N_{min}\left(1 - \frac{\epsilon}{4C^h}\right)\right\}, \tag{4}$$

then, we have that:

$$\mathbb{E}[\|\boldsymbol{p} - \boldsymbol{\pi}^*\|_1] < \epsilon \tag{5}$$

Proof. We begin by decomposing the error into variance and bias terms using the triangle inequality:

$$\mathbb{E}[\|\boldsymbol{p} - \boldsymbol{\pi}^*\|_1] \leq \mathbb{E}[\|\boldsymbol{p} - \mathbb{E}[\boldsymbol{p}]\|_1] + \|\mathbb{E}[\boldsymbol{p}] - \boldsymbol{\pi}^*\|_1 \tag{6}$$

Step 1: Bounding the variance term $\mathbb{E}[||\boldsymbol{p} - \mathbb{E}[\boldsymbol{p}]||_1]$.
We begin by expressing the 1-norm and applying linearity of expectation:

$$\mathbb{E}[||\boldsymbol{p} - \mathbb{E}[\boldsymbol{p}]||_1] = \sum_{c,v} \mathbb{E}[|p(v|c) - \mathbb{E}[p(v|c)]|] \tag{7}$$

Now, we apply Jensen's inequality to each term:

$$\begin{aligned}\sum_{c,v} \mathbb{E}[|p(v|c) - \mathbb{E}[p(v|c)]|] &\leq \sum_{c,v} \sqrt{\mathbb{E}[(p(v|c) - \mathbb{E}[p(v|c)])^2]} \\ &= \sum_{c,v} \sqrt{\mathrm{Var}(p(v|c))} \end{aligned} \tag{8}$$

For each $p(v|c)$, we have that:

$$\mathrm{Var}(p(v|c)) = \frac{\pi(v|c)(1-\pi(v|c))}{N_c} \leq \frac{1}{4N_c} \leq \frac{1}{4\min_{c\in\mathcal{X}^h} N_c} \tag{9}$$

Therefore,

$$\mathbb{E}[||\boldsymbol{p} - \mathbb{E}[\boldsymbol{p}]||_1] \leq \frac{C^h V}{2\sqrt{N_{min}}} \tag{10}$$

Now, to ensure that $\mathbb{E}[||\boldsymbol{p} - \mathbb{E}[\boldsymbol{p}]||_1] \leq \epsilon/2$, we need:

$$\nu^h_{min} \geq \beta^h_{min}\left(\frac{C^h V}{\epsilon}\right)^2 \tag{11}$$

Step 2: Bounding the bias term $\|\mathbb{E}[\boldsymbol{p}] - \boldsymbol{\pi}^*\|_1$.
Since $\mathbb{E}[\boldsymbol{p}] = \boldsymbol{\pi} = \sum_{i=1}^K \alpha_i \boldsymbol{\pi}_i$, where $\boldsymbol{\pi}$ is the matrix representation of π, we have that:

$$\|\mathbb{E}[\boldsymbol{p}] - \boldsymbol{\pi}^*\|_1 = \|\sum_{i=1}^K \alpha_i \boldsymbol{\pi}_i - \frac{1}{\alpha^h}\sum_{i=1}^K \alpha_i^h \boldsymbol{\pi}_i\|_1 \tag{12}$$

$$= \|\sum_{i=1}^K (\alpha_i^l + \alpha_i^h - \frac{\alpha_i^h}{\alpha^h})\boldsymbol{\pi}_i\|_1 \tag{13}$$

$$\leq C^h\left(\sum_{i=1}^K \alpha_i^l + |(1-\frac{1}{\alpha^h})|\sum_{i=1}^K \alpha_i^h\right) \tag{14}$$

$$= 2C^h(1-\alpha^h) = 2C^h\left(1 - \mathbb{E}[\sum_{c\in\mathcal{X}^h} \beta^h_c]\right) \tag{15}$$

$$\leq 2C^h(1-\beta^h_{min}) \tag{16}$$

The key step is recognizing that $\|\boldsymbol{\pi}_i\|_1 = C^h$ for all i. To ensure that this is at most equal to $\epsilon/2$, the following inequality must be satisfied:

$$\nu^h_{min} \geq N_{min}\left(1 - \frac{\epsilon}{4C^h}\right) \tag{17}$$

Combining the two bounds completes the proof. □

Proposition 1 (Sample Complexity Regimes)**.** *The sample complexity ν^h_{min} exhibits distinct regimes as a function of the minimum number of samples N_{min}:*

1) *For $N_{min} \ll N^*_{min}$: $\nu^h_{min} \approx \beta^h_{min}\left(\frac{C^h V}{\epsilon}\right)^2$ (small-data regime)*

2) *For $N_{min} \gg N^*_{min}$: $\nu^h_{min} \approx N_{min}\left(1 - \frac{\epsilon}{4C^h}\right)$ (large-data regime)*

where the transition point N^*_{min} is given by:

$$N^*_{min} = \frac{4\beta^h_{min} {C^h}^3 V^2}{\epsilon^2(4C^h - \epsilon)} \tag{18}$$

The analysis highlights key factors influencing sample complexity in offline RL, particularly emphasizing the role of worst-case context coverage. In data-scarce scenarios, reducing action space complexity and carefully selecting high-return contexts become critical. As data increases, the focus shifts to improving the minimum sample count for any high-return context, with diminishing returns from increasing overall dataset size. This suggests a fundamental trade-off between ensuring a good proportion of high-return samples across contexts (breadth) and sufficient samples per context (depth). These findings have significant implications for data collection strategies in offline RL, suggesting adaptive sampling methods that balance exploration of diverse contexts with exploitation of known high-return areas, particularly focusing on underrepresented high-return contexts in larger datasets.

Proposition 2 (Approximated Critical Minimum Sample Size)**.** *Let N^*_{min} be the critical minimum sample size at which the two terms in the bound of the main theorem are equal. For $\epsilon \ll C^h$, N^*_{min} can be approximated as:*

$$N^*_{min} \approx \frac{\beta^h_{min} {C^h}^2 V^2}{\epsilon^2} + \frac{\beta^h_{min} C^h V^2}{4\epsilon} \tag{19}$$

Proof. To express N^*_{min} in a more analytically tractable form, we consider the typical case where $\epsilon \ll C^h$. Under this assumption, we use a Taylor series expansion:

$$\frac{1}{4C^h - \epsilon} = \frac{1}{4C^h} \cdot \frac{1}{1 - \frac{\epsilon}{4C^h}} \tag{20}$$

$$\approx \frac{1}{4C^h}\left(1 + \frac{\epsilon}{4C^h} + \left(\frac{\epsilon}{4C^h}\right)^2 + \cdots\right) \tag{21}$$

$$\approx \frac{1}{4C^h}\left(1 + \frac{\epsilon}{4C^h}\right) \quad \text{(keeping only first-order terms)} \tag{22}$$

Substituting this back into our expression for N^*_{min} from (18) yields the result in (19). □

The formula for the critical minimum sample size N^*_{min} highlights key factors driving the transition between small-data and large-data regimes in offline RL. The dominant term, $\frac{\beta^h_{min} {C^h}^2 V^2}{\epsilon^2}$, shows that N^*_{min} scales quadratically with the number of high-return contexts (C^h) and the action space size (V), and inversely with the accuracy (ϵ). This indicates that larger datasets are required for problems with complex action spaces or many high-return contexts to transition into the large-data regime.

The secondary term, $\frac{\beta^h_{min} C^h V^2}{4\epsilon}$, becomes more significant as ϵ increases, suggesting that lower accuracy thresholds require more data to balance growth. The linear dependence on β^h_{min} across the formula emphasizes that datasets with more evenly distributed high-return samples across contexts will require proportionally larger sizes to reach the transition point.

VII. Numerical Results

We empirically validate our theoretical findings through controlled experiments.

Experimental Setup: We designed a synthetic, controlled environment with $|\mathcal{S}| = 10$ states and $|\mathcal{A}| = 5$ actions, using an optimal policy (favoring first action with 0.7 probability), a uniform random policy, and a suboptimal policy. The behavior policy weights follow $\alpha^h_k \leq \alpha_k$ with controlled β^h_{min} and N_{min} to ensure consistent coverage across high-return contexts.

Results and Analysis: For values of $\epsilon \in \{0.2, 0.3, 0.4, 0.5\}$, we calculated the theoretical minimum sample size ν^h_{min} and tested whether the resulting empirical error remained below the theoretical threshold. As shown in

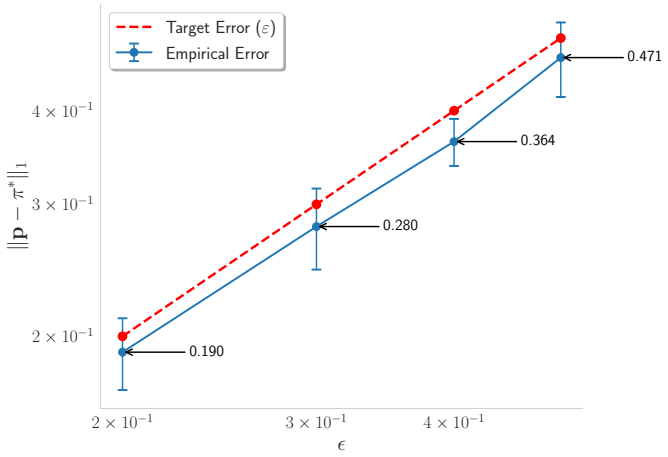

Fig. 1. Empirical approximation error versus theoretical thresholds for varying ϵ values. Error bars represent standard deviation across 10 random seeds.

Figure 1, the empirical error $\|\mathbf{p} - \pi^*\|_1$ consistently remains below the theoretical threshold ϵ across all tested values.

We then varied high-return samples as a fraction $\{0.1, 0.75, 1.0, 1.5, 2.0\}$ of the theoretical minimum to test bound tightness. For $\epsilon = 0.2$ (Figure 2), using 0.75× the theoretical sample size yields errors above threshold, while the full theoretical size maintains error below threshold, confirming our bound's tightness. With the more stringent $\epsilon = 0.1$ (Figure 3), the required samples increase significantly, and using fewer samples than predicted consistently fails to achieve the target error.

Discussion: Our experiments confirm that using the minimum number of high-return samples prescribed by our theory ensures the empirical error remains below the target threshold, while using fewer samples results in higher errors. This simultaneously validates both the effectiveness and tightness of our bounds, providing theoretically grounded guidance for minimum data requirements in offline RL.

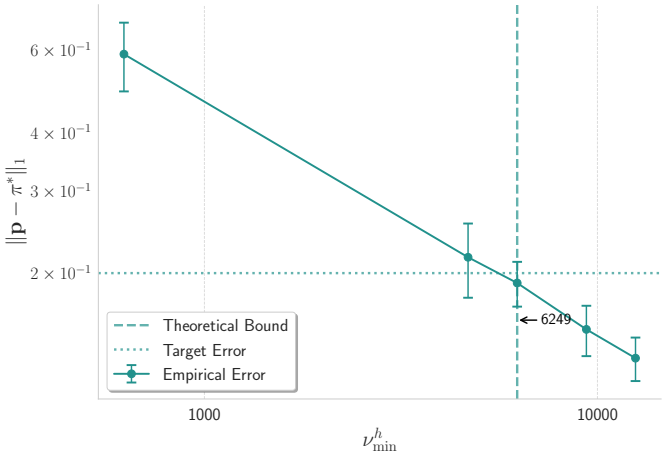

Fig. 2. Approximation error as a function of high-return samples for $\epsilon = 0.2$. Vertical line indicates theoretical minimum sample size.

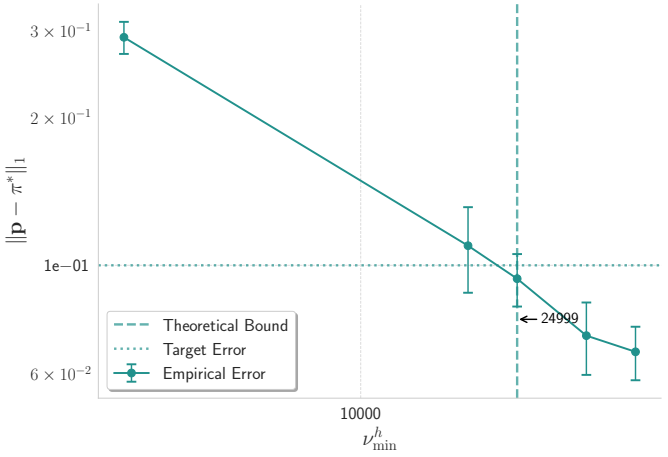

Fig. 3. Sample complexity analysis for $\epsilon = 0.1$, showing increased sample requirements for lower error tolerance.

VIII. Conclusion and Future Directions

This theoretical analysis of sample complexity in offline RL using SM approaches reveals critical insights into the relationship between dataset composition and learning effectiveness. By identifying distinct small-data and large-data regimes separated by a critical transition point, the study challenges the notion that simply increasing dataset size is sufficient for improved performance. Instead, it emphasizes the importance of balanced data collection strategies that ensure adequate coverage of high-return contexts. The revealed trade-off between context coverage breadth and sampling depth suggests that adaptive sampling methods may be more effective than uniform strategies. These findings provide a foundation for developing more efficient offline RL algorithms and data collection strategies, potentially leading to improved performance in real-world applications where data collection is costly or constrained. Future work should focus on validating these theoretical results using realistic transformer architectures on standard offline RL benchmarks, while exploring their practical implications for algorithm design, particularly in developing efficient data selection and weighting strategies that bridge the gap between theoretical insights and real-world applications.

References

[1] Scott Fujimoto, David Meger, and Doina Precup, "Off-policy deep reinforcement learning without exploration," 2019.

[2] Sascha Lange, Thomas Gabel, and Martin Riedmiller, *Batch Reinforcement Learning*, pp. 45–73, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018.

[4] Scott Fujimoto, Herke Hoof, and David Meger, "Addressing function approximation error in actor-critic methods," in *International conference on machine learning*. PMLR, 2018, pp. 1587–1596.

[5] Aviral Kumar, Justin Fu, George Tucker, and Sergey Levine, "Stabilizing off-policy q-learning via bootstrapping error reduction," 2019.

[6] Scott Fujimoto and Shixiang Shane Gu, "A minimalist approach to offline reinforcement learning," *Advances in neural information processing systems*, vol. 34, pp. 20132–20145, 2021.

[7] Ilya Kostrikov, Ashvin Nair, and Sergey Levine, "Offline reinforcement learning with implicit q-learning," 2021.

[8] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine, "Conservative q-learning for offline reinforcement learning," 2020.

[9] Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine, "Rvs: What is essential for offline rl via supervised learning?," *arXiv preprint arXiv:2112.10751*, 2021.

[10] Anurag Ajay, Yilun Du, Abhi Gupta, Joshua B. Tenenbaum, Tommi S. Jaakkola, and Pulkit Agrawal, "Is conditional generative modeling all you need for decision making?," in *The Eleventh International Conference on Learning Representations*, 2023.

[11] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15084–15097, 2021.

[12] Michael Janner, Qiyang Li, and Sergey Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al., "Improving language understanding by generative pre-training," 2018.

[15] Sham Machandranath Kakade, *On the sample complexity of reinforcement learning*, University of London, University College London (United Kingdom), 2003.

[16] Alexander L Strehl, Lihong Li, and Michael L Littman, "Reinforcement learning in finite mdps: Pac analysis.," *Journal of Machine Learning Research*, vol. 10, no. 11, 2009.

[17] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu, "Offline reinforcement learning: Tutorial, review, and perspectives on open problems," *arXiv preprint arXiv:2005.01643*, 2020.

[18] Michael Janner, Qiyang Li, and Sergey Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.

[19] Vladimir Naumovich Vapnik, Vlamimir Vapnik, et al., "Statistical learning theory," 1998.

[20] Mohamed El Amine Seddik, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah, "How bad is training on synthetic data? a statistical analysis of language model collapse," 2024.

[21] Shi Fu, Sen Zhang, Yingjie Wang, Xinmei Tian, and Dacheng Tao, "Towards theoretical understandings of self-consuming generative models," 2024.

[22] Shengyi Jiang, Jingcheng Pang, and Yang Yu, "Offline imitation learning with a misspecified simulator," *Advances in neural information processing systems*, vol. 33, pp. 8510–8520, 2020.