A Note on Batch Value-Function Approximation with Only Realizability

Nai-Chieh, Huang
Department of Computer Science
National Yang Ming Chiao Tung University
naich.cs09@nycu.edu.tw

1 Introduction

Learning the optimal state-action value function Q^* is a crucial problem in the batch reinforcement learning regime. However, it's impossible to learn Q^* without any restrictive conditions. There are several settings are commonly used by the literatures. For example, *low inherent Bellman errors* [Munos and Szepesvári, 2008], *averagers classes* [Gordon, 1995], *state-abstraction* [Li et al., 2006]. In addition, these assumptions are stronger than the reralizability assumption. Moreover, [Chen and Jiang, 2019] provides a conjecture is that efficient learning is impossible with only realizability. Therefore, the following question about batch reinforcement learning remains unelucidated: *Does batch reinforcement learning enjoys efficient sample complexity with only realizability assumption?*

In this paper, they provide an algorithm named *Batch Value-Function Tournament (BVFT)* and analyze its sample complexity. They break the conjecture by the polynomial sample complexity of BVFT by using an exploratory data.

2 Problem Formulation

Markov Decision Processes. Consider a discounted Markov Decision Process $(\mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma, d_0)$, where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a *finite* action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition dynamic of the environment, $R : \mathcal{S} \times \mathcal{A} \to [0, R_{\max}]$ is the bounded reward function, $\gamma \in (0, 1)$ is the discount factor, and $d_0 \in \Delta(\mathcal{S})$ is the initial state distribution. Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ is the unit simplex over $\mathcal{A}$, we define the state-action value function $Q^\pi(\cdot, \cdot)$ as

$$Q^\pi(s, a) \coloneqq \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \Big| s_0 = s, a_0 = a \right]. \tag{1}$$

In addition, we define the total expected reward of a policy π as

$$\mathcal{L}(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \Big| \pi \right], \tag{2}$$

and the above (2) is the objective function. Since we consider a bounded reward function $R \in [0, R_{\max}]$, the objective $\mathcal{L}$ is also bounded in $[0, V_{\max}]$ where $V_{\max} = R_{\max}/(1 - \gamma)$.

We define the optimal policy π^* as the policy that obtains the maximum value in every state and simultaneously maximize the total expected reward $\mathcal{L}$. Also, we define the optimal state-action value function $Q*$ by the Bellman optimality equation: $Q^* = \mathcal{T}Q^*$, where $\mathcal{T}$ is the Bellman optimality operator defined as $(\mathcal{T}f)(s, a) \coloneqq R(s, a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)}[V_f(s)]$, where $V_f(s) \coloneqq \max_a f(s, a)$ for any function f. Also, we define π_Q as the policy that greedily selecting actions according to Q, that is, $\pi_Q(s) = \arg\max_a Q(s, a)$. Note that the optimal policy π^* can be obtained by greedily selecting actions according to Q^*, which is $\pi^*(s) = \pi_{Q^*}(s) = \arg\max_a Q^*(s, a)$.

Batch Data. This paper consider a *batch* reinforcement learning setting, the agent can not directly interact with the MDP. In contrast, it will obtain a batch dataset $\mathcal{D}$ which consists of i.i.d. tuples

(s, a, r, s'), where $(s, a) \sim \mu$, $r = R(s, a)$, and $s' \sim \mathcal{P}(\cdot|s, a)$. Here, $\mu(\cdot, \cdot)$ is a data distribution over the state-action pairs. In this report, I also define $\mu(s)$ and $\mu(a|s)$ as the marginal distribution over state and the conditional probability of action given state, respectively. An exploratory dataset is necessary for efficient learning in batch reinforcement learning, the following is a common technical assumption about the data distribution in batch reinforcement learning literature:

Assumption 1. Let d_t^π be the distribution of (s_t, a_t) that we start from $s_0 \sim d_0$ and interact with the environment through any policy π. We assume that there exists a finite constant C such that $\|d_t^\pi/\mu\|_\infty \leq C$.

However, in the [Xie and Jiang, 2021], which is the main paper we are focus on, they consider a more stringent version as follows:

Assumption 2. Given a data distribution $\mu(s, a) > 0$ for all state-action pairs, we assume the following hold:

- There exists a finite constant $C_\mathcal{A} \geq 1$ such that $\mu(a|s) \geq 1/C_\mathcal{A}$ for all state-action pairs (s, a).
- There exists a finite constant $C_\mathcal{S} \geq 1$ such that $\mathcal{P}(s'|s, a)/\mu(s') \leq C_\mathcal{S}$ for all $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. Also, $d_0(s)/\mu(s) \leq C_\mathcal{S}$ holds.

Throughout this note, we denote $C = C_\mathcal{S} C_\mathcal{A}$ as the concentrability coefficient and suppose Assumption 2 holds. The intuition of the first statement in Assumption 2 is that our data distribution must have enough density for every action in each state, which is equivalent to the exploratoriness of the actions given in the batch $\mathcal{D}$. The second statement is two-fold: (i) It states the exploratoriness of marginal state of the data distribution; (ii) It also gives some constraints to the transition dynamics because the statement contains the $\mathcal{P}$ directly. The point (ii) is undesirable because we would like to consider a general MDP which has general transition dynamics $\mathcal{P}$. However, [Chen and Jiang, 2019] has proven that this result is unavoidable for learning a general function class. Moreover, the stringent Assumption 2 is reasonable not only because the hardness result can be proven if we consider Assumption 1 (see Appendix A of [Xie and Jiang, 2021]), but also the small concentrability coefficient C in low-rank MDP is attainable even with arbitrarily large state space under Assumption 2. Next, we introduce the proposition that directly follows by Assumption 2.

Proposition 1. *Let ν be a distribution over $\mathcal{S} \times \mathcal{A}$ and π be any policy. Then, let $\nu' = \mathcal{P}(\nu) \times \pi$ be the distribution $(s', a') \sim \nu'$ obtained by the process: (i) $(s, a) \sim \nu$, (ii) $s' \sim \mathcal{P}(\cdot|s, a)$, (iii) $a' \sim \pi(\cdot|s')$. Under Assumption 2, we have $\|\nu'/\mu\|_\infty \leq C$. Also, the initial distribution holds that $\|(d_0 \times \pi)/\mu\|_\infty \leq C$.*

Value-Function Approximation. To find a well-performed policy π, we often via the state-action value function. Since the state-action space $\mathcal{S} \times \mathcal{A}$ may be extremely large, function approximation is required. Given a function class $\mathcal{F} \subseteq (\mathcal{S} \times \mathcal{A} \to [0, V_{\max}])$ which contains several different Q functions, we care about how well this function class can represent the optimal state-action value function Q^*. Instead of using inherent Bellman errors [Munos and Szepesvári, 2008], they define the approximation error as the follows:

Definition 1. Define $\epsilon_\mathcal{F} \coloneqq \inf_{f \in \mathcal{F}} \|f - Q^*\|_\infty$. Moreover, define f^* as the f attaining the infimum.

Notably, $\epsilon_\mathcal{F} = 0$ only implies the realizability, which means $Q^* \in \mathcal{F}$. The crucial challenge of batch value-function approximation is to generalize the unseen states and actions.

Polynomial Learning. Polynomial learning is the main focus of the reinforcement learning sample complexity literature. The learning goal aims to find a ϵ-optimal policy $\widehat{\pi}$ such that $\mathcal{L}(\pi^*) - \mathcal{L}(\widehat{\pi}) \leq \epsilon \cdot V_{\max}$ with high probability $1 - \delta$ by using a realizable function class $\mathcal{F}$ and a data batch $\mathcal{D}$ with polynomial sample size. Here, the polynomial may depend on the horizon factor $1/(1 - \gamma)$, the inverse of the suboptimality gap $1/\epsilon$ and the error probability $1/\delta$, the concentrability coefficient C, and the statistical complexity of the function class $\log |\mathcal{F}|$. It is worth mentioning that the logarithmic dependence is important because the effective size of the function class is exponential to the number of parameters. Thus, $\log |\mathcal{F}|$ is equivalent to the number of parameters in the given function class. Given the reason above, we consider the exponentially large function class throughout this note.

Notations. Throughout this note, we slightly abuse the notation, we denote $\mathbb{E}_\mu$ as $\mathbb{E}_{(s,a)\sim\mu, r\sim R(s,a), s'\sim\mathcal{P}(\cdot|s,a)}$. Also, we define the μ-weighted ℓ_2-norm as $\|f\|_{2,\mu} \coloneqq [\mathbb{E}_\mu[f^2]]^{1/2}$. Also, we define the empirical approximation of $\|f\|_{2,\mu}$ by the dataset $\mathcal{D}$ as $\|f\|_{2,\mathcal{D}}$.

3 Batch Value-Function Tournament

Their idea is come from the state-abstraction literature: When the function class $\mathcal{F}$ is piecewise constant and realizable, learning Q^* is consistent with Fitted Q-iteration [Antos et al., 2007]. Specifically, piecewise constant function class is stable and has nice property, e.g., the projected Bellman operator is γ-contraction under ℓ_∞, which implies that Q^* is the unique fixed point when the statistical error is ignored, this will be shown in the next section. It is noteworthy that Q^* is always a fixed point under the projected Bellman operator under a realizable function class, however, it is not a *unique* fixed point. Nevertheless, assuming the function class is piecewise constant is too strong for the practical situation. How can the proof leverage this observation? The solution they use is *improper learning*, i.e., augmenting the function class $\mathcal{F}$ to the smallest set that is piecewise constant and contains $\mathcal{F}$ as a subset simultaneously. By the above method, the new function class we obtain inherently retain the realizability.

In practice, to obtain the piecewise constant function class, they discretize the output of each function $f \in \mathcal{F}$ by a discretization error ϵ_{dct}. That is, all the outputs have at most ϵ_{dct} error to the original values. After the discretization, they group the state-action pairs such that the outputs of *all* functions $f \in \mathcal{F}$ are the same over the state-action pairs of the group. We define ϕ as the grouping function, namely, two state-action pairs have the same value of ϕ if and only if they are in the same group in the discretization results. As a consequence, it becomes a piecewise constant function class. However, there is an issue that the resulting function class become too large. The size of the resulting function class can be evaluated by the number of groups, which is $(V_{\max}/\epsilon_{\text{dct}})^{|\mathcal{F}|}$. It is doubly exponential to the complexity that we can afford, which is $\log|\mathcal{F}|$.

To mitigate the doubly exponential size, we notice that if we put the size of $\mathcal{F}$ as a constant, say $|\mathcal{F}| = 2$, the number of groups will decrease for a considerable magnitude. By using this idea, it turns out that the algorithm could execute the pairwise comparisons between $f, f' \in \mathcal{F}$, and then output the result function as Q^* which beats all the other function during the comparison procedure. Pairwise comparison procedure is a tournament-like manner, so the algorithm is called *Batch Value-Function Tournament (BVFT)*. The above intuitions and ideas ignore the errors during the algorithm, we provide their theoretical analyses in the next section. The following Algorithm 1 is the pseudo code of BVFT.

Algorithm 1: Batch Value-Function Approximation (BVFT)

Input : Dataset $\mathcal{D}$, Function Class $\mathcal{F}$, Discretization Error ϵ_{dct}.

for $f \in \mathcal{F}$ **do**
  $\bar{f} \leftarrow$ discretize the output of f by ϵ_{dct}.
end
for $f \in \mathcal{F}$ **do**
  for $f' \in \mathcal{F}$ **do**
    Define ϕ s.t. $\phi(s,a) = \phi(s',a')$ iff $\bar{f}(s,a) = \bar{f}(s',a')$ and $\bar{f}'(s,a) = \bar{f}'(s',a')$.
    $\mathcal{E}(f; f') \leftarrow \|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$.
  end
end
$\widehat{f} \leftarrow \arg\min_f \max_{f'} \mathcal{E}(f; f')$.

Output : $\widehat{\pi} = \pi_{\widehat{f}}$.

4 Theoretical Analysis

In this section, I introduce the theoretical guarantees they provided in the paper. Below is the main theorem of their paper.

Theorem 1. *With probability at least* $1-\delta$*, BVFT with* $\epsilon_{dct} = \frac{(1-\gamma)^2\epsilon V_{\max}}{16\sqrt{C}}$ *returns a policy* π *that satisfies*

$$\mathcal{L}(\pi^*) - \mathcal{L}(\pi) \leq \frac{(4+8\sqrt{C})\epsilon_{\mathcal{F}}}{(1-\gamma)^2} + \epsilon \cdot V_{\max}, \tag{3}$$

with a sample complexity

$$|\mathcal{D}| = \tilde{O}\left(\frac{C^2 \ln \frac{|\mathcal{F}|}{\delta}}{\epsilon^4 (1-\gamma)^8}\right). \tag{4}$$

Then, we provide the supporting lemmas and intuition before the proof of Theorem 1. The proof of Theorem 1 is at the bottom of this section. We first state the clear picture of this section. After knowing the intuition by Section 3, the problem we must analyze is: Given a piecewise constant function class $\mathcal{G}_\phi$ (induced by the group result ϕ) with small realizability error $\epsilon_\phi \coloneqq \epsilon_{\mathcal{G}_\phi}$, what is the sample complexity to obtain near-optimal policy? They show that they can compute a statistic for any f and the statistic is a good surrogate to evaluate the value of $\|f - Q^*\|$. This magic statistic is $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$, the definition of $\widehat{\mathcal{T}}_\phi^\mu$ is defined as follows:

Definition 2 (Sample-Based Bellman Operator)**.** Define $\widehat{\mathcal{T}}_\phi^\mu$ as a sample-based Bellman operator with the function class $\mathcal{G}_\phi$, where for any f, we define $\widehat{\mathcal{T}}_\phi^\mu f$ as

$$\widehat{\mathcal{T}}_\phi^\mu f \coloneqq \arg\min_{g \in \mathcal{G}_\phi} \frac{1}{|\mathcal{D}|} \sum_{(s,a,r,s') \in \mathcal{D}} \left[(g(s,a) - r - \gamma V_f(s'))^2\right]. \tag{5}$$

4.1 Warm up

We consider the simple case of $|\mathcal{D}| \to \infty$ and $\epsilon_\phi = 0$ first, and then jump into the general case. With regard to $|\mathcal{D}| \to \infty$, we extend the definition of sample-based Bellman operator in Definition 2:

Definition 3 (Projected Bellman Update)**.** Define $\mathcal{T}_\phi^\mu$ as a projected Bellman update with the function class $\mathcal{G}_\phi$, where for any f, we define $\mathcal{T}_\phi^\mu f$ as

$$\mathcal{T}_\phi^\mu f \coloneqq \arg\min_{g \in \mathcal{G}_\phi} \mathbb{E}_\mu \left[(g(s,a) - r - \gamma V_f(s'))^2\right]. \tag{6}$$

After that, they define a MDP M_ϕ induced by ϕ as the following Definition 4 where the projected Bellman update $\mathcal{T}_\phi^\mu$ coincides the Bellman operator of M_ϕ.

Definition 4 (ϕ-induced MDP)**.** Define an MDP $M_\phi = (\mathcal{S}, \mathcal{A}, \mathcal{P}_\phi, R_\phi, \gamma, d_0)$, where also define $\mathcal{P}_\phi, R_\phi$ as follow:

$$\mathcal{P}_\phi(s'|s,a) \coloneqq \frac{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a}) \cdot \mathcal{P}(s'|\bar{s},\bar{a})}{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a})} \tag{7}$$

$$R(s,a) \coloneqq \frac{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a}) \cdot R(\bar{s},\bar{a})}{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a})} \tag{8}$$

Then, combining the Definitions 3, 4, it immediately follows:

Lemma 1. *$\mathcal{T}_\phi^\mu$ is a Bellman update operator of M_ϕ.*

Proof. Throughout the proof, we define $\mathcal{T}_{M_\phi}$ as the Bellman operator of M_ϕ. We first show the closeness of $\mathcal{T}_{M_\phi}$ under $\mathcal{G}_\phi$, which is for any f, $\mathcal{T}_{M_\phi} f \in \mathcal{G}_\phi$. We directly write down the $(\mathcal{T}_{M_\phi} f)(s,a)$ as

$$(\mathcal{T}_{M_\phi} f)(s,a) = R_\phi(s,a) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}_\phi(\cdot|s,a)}[V_f(s')]. \tag{9}$$

Since for any $s, a, \bar{s}, \bar{a}$ such that $\phi(s,a) = \phi(\bar{s},\bar{a})$, we have $\mathcal{P}_\phi(\cdot|s,a) = \mathcal{P}_\phi(\cdot|\bar{s},\bar{a})$ and $R_\phi(s,a) = R_\phi(\bar{s},\bar{a})$, we obtain that $\mathcal{T}_{M_\phi} f \in \mathcal{G}_\phi$. Then, we show that $\mathcal{T}_{M_\phi} f(s,a) = R_\phi(s,a) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}_\phi(\cdot|s,a)}[V_f(s')]$ which is exactly the $\arg\min$ of Eq. (6),

$$(\mathcal{T}_\phi^\mu f)(s,a) = \mathbb{E}_{(\bar{s},\bar{a},r,s')}[r + \gamma \cdot V_f(s') | \phi(s,a) = \phi(\bar{s},\bar{a})] \tag{10}$$

$$= \frac{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a}) \cdot (R(\bar{s},\bar{a}) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}(\cdot|\bar{s},\bar{a})}[V_f(s')])}{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a})} \tag{11}$$

$$= R_\phi(s,a) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}_\phi(\cdot|s,a)}[V_f(s')] = (\mathcal{T}_{M_\phi} f)(s,a). \tag{12}$$

We obtain the result by expanding the equation and expectation and then reform the equation by R_ϕ and $\mathcal{P}_\phi$. □

Lemma 1 gives that $\mathcal{T}_\phi^\mu$ is a γ-contraction under ℓ_∞ and thus it has a unique fixed point. Next, we show that the Q^* satisfies the fixed point condition.

Proposition 2. *When $\epsilon_\phi = 0$, Q^* is the unique fixed point under $\mathcal{T}_\phi^\mu$.*

Proof. The existence and uniqueness follow by Lemma 1. We only need to show that $Q^* = \mathcal{T}_\phi^\mu Q^*$. By the Eq. (11), $\mathcal{T}_\phi^\mu$ is a convex combination over $Q^*(\bar{s}, \bar{a})$ where $\phi(s,a) = \phi(\bar{s}, \bar{a})$. Since $\epsilon_\phi = 0$, we have $Q^*(\bar{s}, \bar{a}) = Q^*(s,a)$ if $\phi(s,a) = \phi(\bar{s}, \bar{a})$. Thus, taking convex combination over $Q^*(s,a)$ is $Q^*(s,a)$. Finally, we have $Q^* = \mathcal{T}_\phi^\mu Q^*$ and complete the proof. □

Now, we finish the warm up and are ready to jump into the general case. In the following section, we denote $|\phi|$ as the number of groups induced by ϕ.

4.2 General Case

In general case, they provided the most important proposition for proving Theorem 1 as follows:

Proposition 3. *Fixing any $\epsilon_1, \tilde{\epsilon}$. Suppose the size of dataset satisfies*

$$|\mathcal{D}| \geq \frac{32 V_{\max}^2 |\phi| \ln \frac{8V_{\max}}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2} + \frac{50 V_{\max}^2 |\phi| \ln \frac{80 V_{\max}}{\epsilon_1 \delta}}{\epsilon_1^2}. \tag{13}$$

Then, with probability at least $1-\delta$, for any $\nu \in \Delta(\mathcal{S} \times \mathcal{A})$ such that $\|\nu/\mu\|_\infty \leq C$, the followings hold,

$$\|f - Q^*\|_{2,\nu} \leq \frac{2\epsilon_\phi + \sqrt{C}(\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}} + \epsilon_1 + \tilde{\epsilon})}{1-\gamma}, \tag{14}$$

$$\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}} \leq (1+\gamma)\|f - Q^*\|_\infty + 2\epsilon_\phi + \epsilon_1 + \tilde{\epsilon}. \tag{15}$$

Recall that we said $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$ is a magic statistic as a surrogate of $\|f - Q^*\|$. Proposition 3 provides the bound between $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$ and $\|f - Q^*\|$. We will see why there are two different ϵ, which is come from two different error sources, in the statement of Proposition 3 later in this section. To prove Proposition 3, we need several supporting lemmas.

In the common state-abstraction literature [Jiang, 2018], they often bound $\|f - Q^*\|$ by $\|f - \mathcal{T}_\phi^\mu f\|_\infty$. However, it requires further assumption on the state-action data distribution to bound $\|f - \mathcal{T}_\phi^\mu f\|_\infty$, where in the paper we focus, they do not have this kind of assumption. Thus, this is the challenge they must overcome that they should carefully estimate the error for μ-weighted ℓ_2-norm. This is the reason that they need a more stringent assumption as Assumption 2. Fortunately, Assumption 2 has nice property that it can make the condition of itself holds on M_ϕ for arbitrary ϕ if the original MDP M satisfies Assumption 2, they stated a proposition as follows:

Lemma 2. *The finite concentrability coefficient C which satisfies in the true MDP M also satisfies in the ϕ-induced MDP M_ϕ. Moreover, Proposition 1 also satisfies when we replace $\mathcal{P}$ by $\mathcal{P}_\phi$.*

For the error bound we need, they first provide an error bound of $\|Q^* - \mathcal{T}_\phi^\mu Q^*\|_\infty$, we state it as follows:

Lemma 3. $\|Q^* - \mathcal{T}_\phi^\mu Q^*\|_\infty \leq 2\epsilon_\phi$.

Proof. Let $g^* = \arg\min_{g \in \mathcal{G}_\phi} \|g - Q^*\|_\infty$. Since we discretize the function by the discretization error ϵ_ϕ and obtain $\mathcal{G}_\phi$, we have $\|g^* - Q^*\|_\infty \leq \epsilon_\phi$. Then, for any (s,a), we have

$$|Q^*(s,a) - (\mathcal{T}_\phi^\mu Q^*)(s,a)| \tag{16}$$

$$= \left| Q^*(s,a) - \frac{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a}) \cdot (R(\bar{s},\bar{a}) + \gamma \cdot \mathbb{E}_{s' \sim \mathcal{P}(\cdot|\bar{s},\bar{a})}[V_{Q^*}(s')])}{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a})} \right| \tag{17}$$

$$= \left| Q^*(s,a) - \frac{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a}) \cdot Q^*(\bar{s},\bar{a})}{\sum_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} \mu(\bar{s},\bar{a})} \right| \leq \max_{(\bar{s},\bar{a}):\phi(\bar{s},\bar{a})=\phi(s,a)} |Q^*(s,a) - Q^*(\bar{s},\bar{a})|. \tag{18}$$

Furthermore, for any $(\bar{s}, \bar{a})$ such that $\phi(\bar{s}, \bar{a}) = \phi(s, a)$,

$$
\begin{aligned}
|Q^*(s,a) - Q^*(\bar{s},\bar{a})| &= |Q^*(s,a) - g^*(s,a) + g^*(s,a) - Q^*(\bar{s},\bar{a})| && (19)\\
&= |Q^*(s,a) - g^*(s,a) + g^*(\bar{s},\bar{a}) - Q^*(\bar{s},\bar{a})| && (20)\\
&\leq |Q^*(s,a) - g^*(s,a)| + |g^*(\bar{s},\bar{a}) - Q^*(\bar{s},\bar{a})| && (21)\\
&\leq 2\|g^* - Q^*\|_\infty \leq 2\epsilon_\phi. && (22)
\end{aligned}
$$

□

Lemma 3 inherently gives the result of Proposition 2 when $\epsilon_\phi = 0$. Moreover, we will leverage the result of Lemma 3 in the proof of Proposition 3.

To connect $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$ and $\|f - Q^*\|$, we need two concentration bounds about the error between $\widehat{\mathcal{T}}_\phi^\mu f$ and $\mathcal{T}_\phi^\mu f$ is small and the error between $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}}$ and $\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mu}$ is also small. The two lemmas are as follow:

Lemma 4. *With probability at least* $1 - \delta/2$, $\|\mathcal{T}_\phi^\mu f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mu} \leq \tilde{\epsilon}$, *as long as the size of dataset satisfies,*

$$
|\mathcal{D}| \geq \frac{32 V_{\max}^2 |\phi| \log \frac{8V_{\max}}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2}. \tag{23}
$$

Lemma 5. *With probability at least* $1 - \delta/2, \forall g \in \mathcal{G}_\phi, |\|f - g\|_{2,\mathcal{D}} - \|f - g\|_{2,\mu}| \leq \epsilon_1$, *as long as the size of dataset satisfies,*

$$
|\mathcal{D}| \geq \frac{50 V_{\max}^2 |\phi| \ln \frac{80V_{\max}}{\epsilon_1\delta}}{\epsilon_1^2}. \tag{24}
$$

Then, we are ready to prove Proposition 3.

Proof of Proposition 3

I prove Proposition 3 by myself, so there is a little bit different from them

Throughout the proof, we put $\mathcal{T}$ as the projected Bellman update $\mathcal{T}_\phi^\mu$ induced by ϕ and $\widehat{\mathcal{T}}$ as the sample-based Bellman update $\widehat{\mathcal{T}}_\phi^\mu$. First of all, the dataset size requirement in Eq. (13) is the direct result of Lemma 4 and Lemma 5. In addition, Lemma 4 and Lemma 5 hold both with probability $1 - \delta/2$, by union bound we obtain the probability $1 - \delta$ in the Proposition 3 statement.

Then, We first show the result of Eq. (14). We have

$$
\|f - Q^*\|_{2,\nu} \leq \|Q^* - \mathcal{T}Q^*\|_{2,\nu} + \|\mathcal{T}Q^* - \mathcal{T}f\|_{2,\nu} + \|\mathcal{T}f - f\|_{2,\nu}. \tag{25}
$$

For the first term in the RHS of Eq. (25), we have $\|\cdot\|_{2,\nu} \leq \|\cdot\|_\infty$ and by Lemma 3, we have

$$
\|f - Q^*\|_{2,\nu} \leq \|f - Q^*\|_\infty \leq 2\epsilon_\phi. \tag{26}
$$

Next, for the second term, we define a policy $\pi_{Q^*,f}(s) = \arg\max_a |Q^*(s,a) - f(s,a)|$. We have

$$
\begin{aligned}
\|\mathcal{T}Q^* - \mathcal{T}f\|_{2,\nu}^2 &= \mathbb{E}_{(s,a)\sim\nu}[((\mathcal{T}Q^*)(s,a) - (\mathcal{T}f)(s,a))^2] && \text{(Def. of norm)}\\
&= \gamma^2 \mathbb{E}_{(s,a)\sim\nu}[(\mathbb{E}_{s'\sim\mathcal{P}_\phi(\cdot|s,a)}[(V_{Q^*}(s') - V_f(s'))])^2] && \text{(Def. of } \mathcal{T})\\
&\leq \gamma^2 \mathbb{E}_{(s,a)\sim\nu, s'\sim\mathcal{P}_\phi(\cdot|s,a)}[(V_{Q^*}(s') - V_f(s'))^2] && \text{(Jensen's Inequality)}\\
&= \gamma^2 \cdot \sum_{(s,a)} \sum_{s'} \nu(s,a) \cdot \mathcal{P}_\phi(s'|s,a)(V_{Q^*}(s') - V_f(s'))^2 && (27)\\
&= \gamma^2 \cdot \sum_{(s,a)} \sum_{s'} \nu(s,a) \cdot \mathcal{P}_\phi(s'|s,a)(\max_a Q^*(s',a) - \max_{a'} f(s',a'))^2 && (28)\\
&\leq \gamma^2 \cdot \sum_{(s,a)} \sum_{s'} \nu(s,a) \cdot \mathcal{P}_\phi(s'|s,a) \max_a (Q^*(s',a) - f(s',a))^2 && (29)\\
&\leq \gamma^2 \cdot \sum_{(s,a)} \sum_{s'} \nu(s,a) \cdot \mathcal{P}_\phi(s'|s,a)(Q^*(s',\pi_{Q^*,f}(s')) - f(s',\pi_{Q^*,f}(s')))^2 && (30)\\
&= \gamma^2 \cdot \|Q^* - f\|_{2,\mathcal{P}_\phi(\nu)\times\pi_{Q^*,f}(s)}^2. && (31)
\end{aligned}
$$

Last, we take the square root from the both sides to obtain the bound of the second term. For the third term, since we have $\|\nu/\mu\|_\infty \leq C$, we have

$$\|\mathcal{T}f - f\|_{2,\nu} = \left[\sum_{s,a} \nu(s,a)(\mathcal{T}f - f)^2\right]^{1/2} \tag{32}$$

$$\leq \left[\sum_{s,a} C \cdot \mu(s,a)(\mathcal{T}f - f)^2\right]^{1/2} \tag{33}$$

$$= \sqrt{C}\|\mathcal{T}f - f\|_{2,\mu}. \tag{34}$$

Thus, we have

$$\|f - Q^*\|_{2,\nu} \leq 2\epsilon_\phi + \gamma\|Q^* - f\|_{2,\mathcal{P}_\phi(\nu)\times\pi_{Q^*,f}(s)} + \sqrt{C}\|\mathcal{T}f - f\|_{2,\mu}. \tag{35}$$

By Lemma 2, $\mathcal{P}_\phi(\nu) \times \pi_{Q^*,f}(s)$ also satisfies $\|(\cdot)/\mu\|_\infty \leq C$, so we consider it as one of the ν from the LHS. Thus, we have

$$\sup_{\nu:\|\nu/\mu\|_\infty\leq C} \|f - Q^*\|_{2,\nu} \leq \gamma \sup_{\nu:\|\nu/\mu\|_\infty\leq C} \|f - Q^*\|_{2,\mathcal{P}_\phi(\nu)\times\pi_{Q^*,f}(s)} + 2\epsilon_\phi + \sqrt{C}\|\mathcal{T}f - f\|_{2,\mu} \tag{36}$$

$$\leq \gamma \sup_{\nu:\|\nu/\mu\|_\infty\leq C} \|f - Q^*\|_{2,\nu} + 2\epsilon_\phi + \sqrt{C}\|\mathcal{T}f - f\|_{2,\mu}. \tag{37}$$

We rearrange the terms and obtain,

$$\|f - Q^*\|_{2,\nu} \leq \frac{2\epsilon_\phi + \sqrt{C}\|\mathcal{T}f - f\|_{2,\mu}}{1-\gamma}. \tag{38}$$

Last, we leverage Lemma 4 and Lemma 5 to bound $\|\mathcal{T}f - f\|_{2,\mu}$ by $\|\mathcal{T}f - f\|_{2,\mathcal{D}}$ and obtain the result of Eq. (14), we have

$$\|\mathcal{T}f - f\|_{2,\mu} \leq \|\mathcal{T}f - \widehat{\mathcal{T}}f\|_{2,\mu} + \|\widehat{\mathcal{T}}f - f\|_{2,\mu} \tag{39}$$

$$\leq \tilde{\epsilon} + \|\widehat{\mathcal{T}}f - f\|_{2,\mathcal{D}} + \epsilon_1. \tag{40}$$

To sum up, for any fixed f, we have

$$\|f - Q^*\|_{2,\nu} \leq \frac{2\epsilon_\phi + \sqrt{C}(\|f - \widehat{\mathcal{T}}_\phi^\mu f\|_{2,\mathcal{D}} + \epsilon_1 + \tilde{\epsilon})}{1-\gamma}. \tag{41}$$

Next, we prove for the Eq. (15).

$$\|\widehat{\mathcal{T}}f - f\|_{2,\mathcal{D}} \leq \epsilon_1 + \|\widehat{\mathcal{T}}f - f\|_{2,\mu} \qquad \text{(Lemma 5)}$$

$$\leq \epsilon_1 + \|\mathcal{T}f - \widehat{\mathcal{T}}f\|_{2,\mu} + \|\mathcal{T}f - f\|_{2,\mu} \tag{42}$$

$$\leq \epsilon_1 + \tilde{\epsilon} + \|\mathcal{T}f - f\|_{2,\mu} \qquad \text{(Lemma 4)}$$

$$\leq \epsilon_1 + \tilde{\epsilon} + \|\mathcal{T}f - f\|_\infty \tag{43}$$

$$\leq \epsilon_1 + \tilde{\epsilon} + \|\mathcal{T}f - \mathcal{T}Q^*\|_\infty + \|\mathcal{T}Q^* - Q^*\|_\infty + \|Q^* - f\|_\infty \tag{44}$$

$$\leq \epsilon_1 + \tilde{\epsilon} + \|\mathcal{T}f - \mathcal{T}Q^*\|_\infty + 2\epsilon_\phi + \|Q^* - f\|_\infty \qquad \text{(Lemma 3)}$$

$$\leq \epsilon_1 + \tilde{\epsilon} + 2\epsilon_\phi + (1+\gamma)\|Q^* - f\|_\infty \qquad (\gamma\text{-contraction of } \mathcal{T})$$

(45)

□

At the bottom of this section, we prove the main theorem, which is Theorem 1. We first provide their intuition of the proof. In the proof, they only consider the comparisons between f, f' such that $Q^* \in \{f, f'\}$. Since the target is to find the Q^*, while Q^* never participates in those comparisons. Thus, in their proof that we will see later, they only consider the $2|\mathcal{F}|$ pairs where $f^* = Q^* \in \{f, f'\}$.

Before we begin the proof, we must state the last supporting lemma that is required, which is connecting the approximation error ϵ_ϕ of $\mathcal{G}_\phi$ to the original one $\epsilon_\mathcal{F}$ of $\mathcal{F}$.

Lemma 6. *The number of group induced by ϕ satisfies $|\phi| \leq (V_{\max}/\epsilon_{dct})^2$. Moreover, when $f^* \in \{f, f'\}$, we have $\epsilon_\phi \leq \epsilon_{\mathcal{F}} + \epsilon_{dct}$.*

Proof. Since there are at most $V_{\max}/\epsilon_{\text{dct}}$ number of different values of a function f. Thus, the number of group induced by ϕ can be upper bounded by all the combinations of the output of two functions, which is $(V_{\max}/\epsilon_{\text{dct}})^2$.

For the second statement, by the definition of ϵ_ϕ, we have $\epsilon_\phi = \min_{g \in \mathcal{G}_\phi} \|g - Q^*\|_\infty$. We consider the function $\bar{f}^*$ which is the result after discretization of f^*. We have

$$\begin{aligned}
\epsilon_\phi &= \min_{g \in \mathcal{G}_\phi} \|g - Q^*\|_\infty && (46)\\
&\leq \|\bar{f}^* - Q^*\|_\infty && (47)\\
&\leq \|f^* - Q^*\|_\infty + \|\bar{f}^* - f^*\|_\infty && (48)\\
&\leq \epsilon_{\mathcal{F}} + \epsilon_{\text{dct}}. && (49)
\end{aligned}$$

□

Proof of Theorem 1

We only consider the comparisons between f, f' such that $f^* \in \{f, f'\}$. There are $2|\mathcal{F}|$ pairs of comparisons. We require that those comparisons must hold the properties of Proposition 3 with probability $1 - \delta$, so we replace the δ by $\delta/(2|\mathcal{F}|)$ for union bound. In addition, we let $\epsilon_1 = \tilde{\epsilon}$ for simplicity. Then, by Lemma 6, the dataset size requirement is

$$|\mathcal{D}| \geq \frac{32V_{\max}^2|\phi| \ln \frac{8V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta} + 50V_{\max}^2|\phi| \ln \frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2}. \tag{50}$$

We find a simple upper bound for the RHS of the above Eq. (50). We have

$$\begin{aligned}
\frac{32V_{\max}^2|\phi| \ln \frac{8V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta} + 50V_{\max}^2|\phi| \ln \frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2} &\leq V_{\max}^4 \frac{32 \ln \frac{8V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta} + 50 \ln \frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2} && (51)\\
&\leq \frac{V_{\max}^4}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2} \ln \left(\left(\frac{8V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}\right)^{32} \cdot \left(\frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}\right)^{50} \right) && (52)\\
&\leq \frac{V_{\max}^4}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2} \ln \left(\left(\frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}\right)^{32} \cdot \left(\frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}\right)^{50} \right) && (53)\\
&\leq \frac{V_{\max}^4}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2} \ln \left(\frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}\right)^{82} && (54)\\
&\leq \frac{82V_{\max}^4 \ln \frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2} && \\
& \qquad \text{(Their proof use 160 instead of my 80.)} && (55)
\end{aligned}$$

Thus, the dataset size requirement becomes:

$$|\mathcal{D}| \geq \frac{82V_{\max}^4 \ln \frac{80V_{\max}|\mathcal{F}|}{\tilde{\epsilon}\delta}}{\tilde{\epsilon}^2\epsilon_{\text{dct}}^2}. \tag{56}$$

Next, we consider the comparison between $\widehat{f}, f^*$. By the Eq. (14) in Proposition 3, for any ν satisfies $\|\nu/\mu\|_\infty \leq C$, we have

$$\begin{aligned}
\|\widehat{f} - Q^*\|_{2,\nu} &\leq \frac{2\epsilon_\phi + \sqrt{C}(\|\widehat{f} - \widehat{\mathcal{T}}_\phi^\mu \widehat{f}\|_{2,\mathcal{D}} + 2\tilde{\epsilon})}{1-\gamma} && (57)\\
&= \frac{2\epsilon_\phi + \sqrt{C}(\mathcal{E}(\widehat{f}; f^*) + 2\tilde{\epsilon})}{1-\gamma} && (58)\\
&\leq \frac{2\epsilon_\phi + \sqrt{C}(\max_{f'} \mathcal{E}(\widehat{f}; f') + 2\tilde{\epsilon})}{1-\gamma}. && (59)
\end{aligned}$$

According to the above RHS, we need to bound the $\max_{f'} \mathcal{E}(\widehat{f}; f')$. By Algorithm 1, we have

$$\max_{f'} \mathcal{E}(\widehat{f}; f') = \min_{f} \max_{f'} \mathcal{E}(f; f') \leq \max_{f'} \mathcal{E}(f^*; f'). \tag{60}$$

Define ϕ' as the group induced by f^* and f', we leverage the Eq. (15), we obtain

$$\begin{aligned}
\max_{f'} \mathcal{E}(f^*; f') &= \|f^* - \widehat{\mathcal{T}}^{\mu}_{\phi'} f^*\|_{2,\mathcal{D}} && \text{(Def. of } \mathcal{E}) \\
&\leq (1+\gamma)\|f^* - Q^*\|_{\infty} + 2\epsilon_{\phi'} + 2\tilde{\epsilon} && \text{(Eq. 15)} \\
&\leq 2\epsilon_{\mathcal{F}} + 2(\epsilon_{\mathcal{F}} + \epsilon_{\text{dct}}) + 2\tilde{\epsilon} && (\gamma < 1 \text{ and Lemma 6}) \\
&= 4\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + 2\tilde{\epsilon}. && (61)
\end{aligned}$$

Thus, rearranging the term in Eq. (59), we have

$$\|\widehat{f} - Q^*\|_{2,\nu} \leq \frac{2\epsilon_{\phi} + \sqrt{C}(4\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + 2\tilde{\epsilon} + 2\tilde{\epsilon})}{1-\gamma} \tag{62}$$

$$\leq \frac{2(\epsilon_{\mathcal{F}} + \epsilon_{\text{dct}}) + \sqrt{C}(4\epsilon_{\mathcal{F}} + 2\epsilon_{\text{dct}} + 4\tilde{\epsilon})}{1-\gamma} \tag{63}$$

$$\leq \frac{(4\sqrt{C}+2)\epsilon_{\mathcal{F}} + 4\sqrt{C}(\epsilon_{\text{dct}} + \tilde{\epsilon})}{1-\gamma}. \tag{64}$$

Last, we consider the suboptimality gap, by leveraging Lemma 13 of [Chen and Jiang, 2019] (should updated to be self-contained), since any state-action distribution with any policy satisfies the Proposition 1, we have

$$\mathcal{L}(\pi^*) - \mathcal{L}(\widehat{\pi}) \leq \frac{2}{1-\gamma} \sup_{\nu : \|\nu/\mu\|_{\infty} \leq C} \|\widehat{f} - Q^*\|_{2,\nu} \tag{65}$$

$$\leq \frac{(8\sqrt{C}+4)\epsilon_{\mathcal{F}} + 8\sqrt{C}(\epsilon_{\text{dct}} + \tilde{\epsilon})}{(1-\gamma)^2}. \tag{66}$$

Finally, we let

$$\epsilon_{\text{dct}} = \tilde{\epsilon} = \frac{(1-\gamma)^2 \epsilon V_{\max}}{16\sqrt{C}}. \tag{67}$$

to guarantee the inequality $\frac{8\sqrt{C}(\epsilon_{\text{dct}}+\tilde{\epsilon})}{(1-\gamma)^2} \leq \epsilon \cdot V_{\max}$ holds. We substitute the values of ϵ_{dct} and $\tilde{\epsilon}$ to the dataset size requirement Eq. (56) and obtain the result. □

5 Concluding Remarks

The algorithm BVFT we provided can analyze the batch value-function approximation with only realizability. The sample complexity is with the $1/\epsilon^4$ rate. After the analysis, the rate has two sources; (i) The statistical complexity of pairwise comparison, which is $1/\epsilon^2$; (ii) The discretization error causes by the improper learning has also $1/\epsilon^2$ complexity. Therefore, their results obtain a overall $1/\epsilon^4$ rate.

The limitation of this work is that they consider a more stringent assumption on the distribution of the batch data. Recently, [Foster et al., 2021] has proved that the sample efficient batch learning requires either assumptions about data distributions (*data coverage*) or assumptions about function class representation. In addition, another limitation is the pairwise comparison of BVFT, which will cause $O(|\mathcal{F}|^2)$ computational complexity, which is computationally expensive.

References

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(27):815–857, 2008. URL `http://jmlr.org/papers/v9/munos08a.html`.

Geoffrey J Gordon. Stable function approximation in dynamic programming. In *Machine learning proceedings 1995*, pages 261–268. Elsevier, 1995.

Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *ISAIM*, 4(5):9, 2006.

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.

András Antos, Csaba Szepesvári, and Rémi Munos. Fitted q-iteration in continuous action-space mdps. *Advances in neural information processing systems*, 20, 2007.

Nan Jiang. Notes on state abstractions, 2018.

Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: Fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.