Bootstrapping Fitted Q-Evaluation for Off-Policy Inference

Botao Hao [1] **Xiang Ji** [2] **Yaqi Duan** [2] **Hao Lu** [2] **Csaba Szepesvári** [1 3] **Mengdi Wang** [1 2]

Abstract

Bootstrapping provides a flexible and effective approach for assessing the quality of batch reinforcement learning, yet its theoretical properties are poorly understood. In this paper, we study the use of bootstrapping in off-policy evaluation (OPE), and in particular, we focus on the fitted Q-evaluation (FQE) that is known to be minimax-optimal in the tabular and linear-model cases. We propose a bootstrapping FQE method for inferring the distribution of the policy evaluation error and show that this method is asymptotically efficient and distributionally consistent for off-policy statistical inference. To overcome the computation limit of bootstrapping, we further adapt a subsampling procedure that improves the runtime by an order of magnitude. We numerically evaluate the bootrapping method in classical RL environments for confidence interval estimation, estimating the variance of off-policy evaluator, and estimating the correlation between multiple off-policy evaluators.

1. Introduction

Off-policy evaluation (OPE) often serves as the starting point of batch reinforcement learning (RL). The objective of OPE is to estimate the value of a target policy based on batch episodes of state-transition trajectories that were generated using a different and possibly unknown behavior policy. In this paper, we investigate statistical inference for OPE. In particular, we analyze the popular fitted Q-evaluation (FQE) method, which is a basic model-free approach that fits unknown value function from data using function approximation and backward dynamic programming (Fonteneau et al., 2013; Munos & Szepesvári, 2008; Le et al., 2019). In practice, FQE has demonstrated robust and satisfying performances on many classical RL tasks under different metrics (Voloshin et al., 2019). A more recent study by Paine et al. (2020) demonstrated surprising scalability and effectiveness of FQE with deep neural nets in a range of complex continuous-state RL tasks. On the theoretical side, FQE was proved to be a minimax-optimal policy evaluator in the tabular and linear-model cases (Yin & Wang, 2020; Duan & Wang, 2020).

The aforementioned research mostly focuses on point estimation for OPE. In practical batch RL applications, a point estimate is far from enough. Statistical inference for OPE is of great interests. For instance, one often hopes to construct tight confidence interval around policy value, estimate the variance of off-policy evaluator, or evaluate multiple policies using the same data and estimate their correlations. Bootstrapping (Efron, 1982), is a conceptually simple and generalizable approach to infer the error distribution based on batch data. Therefore, in this work, we study the use of bootstrapping for *off-policy inference*. We will provide theoretical justifications as well as numerical experiments.

Our main results are summarized below:

- First we analyze the asymptotic distribution of FQE with linear function approximation and show that the policy evaluation error asymptotically follows a normal distribution (Theorem 4.2). The asymptotic variance matches the Cramér–Rao lower bound for OPE (Theorem 4.5) and implies that this estimator is asymptotically efficient.
- We propose a bootstrapping FQE method for estimating the distribution of off-policy evaluation error. We prove that bootstrapping FQE is asymptotically consistent in estimating the distribution of the original FQE (Theorem 5.1) and establish the consistency of bootstrap confidence interval as well as bootstrap variance estimation. Further, we propose a subsampled bootstrap procedure to improve the computational efficiency of bootstrapping FQE.

[1]Deepmind [2]Princeton University [3]University of Alberta. Correspondence to: Botao Hao <haobotao000@gmail.com>, Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 139, 2021.

- We highlight the necessity of *bootstrapping by episodes*, rather than by individual sample transition as considered in previous works; see Kostrikov & Nachum (2020). The reason is that bootstrapping dependent data in general fails to characterize the right error distribution (Remark 2.1 in Singh (1981)). We illustrate this phenomenon via experiments (see Figure 1). All our theoretical analysis applies to episodic dependent data, and we do not require the i.i.d. sample transition assumption commonly made in OPE literatures (Jiang & Huang, 2020; Kostrikov & Nachum, 2020; Dai et al., 2020).

- Finally, we evaluate subsampled bootstrapping FQE in a range of classical RL tasks, including a discrete tabular domain, a continuous control domain and a simulated healthcare example. We test variants of bootstrapping FQE with tabular representation, linear function approximation, and neural networks. We carefully examine the effectiveness and tightness of bootstrap confidence intervals, as well as the accuracy of bootstrapping for estimating the variance and correlation for OPE.

Related Work. Point estimation of OPE receives considerable attentions in recent years. Popular approaches include direct methods (Lagoudakis & Parr, 2003; Ernst et al., 2005; Munos & Szepesvári, 2008; Le et al., 2019), double-robust / importance sampling (Precup et al., 2000; Jiang & Li, 2016; Thomas & Brunskill, 2016), marginalized importance sampling (Hallak & Mannor, 2017; Liu et al., 2018; Xie et al., 2019; Nachum et al., 2019; Uehara & Jiang, 2019; Zhang et al., 2020a;b). On the theoretical side, Uehara & Jiang (2019); Yin & Wang (2020) established asymptotic optimality and efficiency for OPE in the tabular setting and Kallus & Uehara (2020) provided a complete study of semiparametric efficiency in a more general setting. Duan & Wang (2020); Hao et al. (2020b) showed that FQE with linear/sparse lienar function approximation is minimax optimal and Wang et al. (2020) studied the fundamental hardness of OPE with linear function approximation.

Confidence interval estimation of OPE is also important in many high-stake applications. Thomas et al. (2015) proposed a high-confidence OPE based on importance sampling and empirical Bernstein inequality. Kuzborskij et al. (2020) proposed a tighter confidence interval for contextual bandits based on empirical Efron-Stein inequality. However, importance sampling suffers from the curse of horizon (Liu et al., 2018) and concentration-based confidence intervals are typically overly-conservative since they only exploit tail information (Hao et al., 2020a). Another line of recent works formulated the estimation of confidence intervals into an optimization problem (Feng et al., 2020; 2021; Dai et al., 2020). These works are specific to confidence interval construction for OPE, and they do not provide distributional consistency guarantee. Thus, they don't easily generalize to other statistical inference tasks.

In statistics community, Liao et al. (2019) studied OPE in an infinite-horizon undiscounted MDP and derived the asymptotic distribution of empirical Bellman residual minimization estimator. Their asymptotic variance had a tabular representation and thus didn't show the effect of function approximation. Shi et al. (2020) considered asymptotic confidence interval for policy value but under different model assumption that assumes Q-function is smooth.

Several existing work has investigated the use of bootstrapping in OPE. Thomas et al. (2015); Hanna et al. (2017) constructed confidence intervals by bootstrapping importance sampling estimator or learned models but didn't come with any consistency guarantee. The most related work is Kostrikov & Nachum (2020) that provided the first asymptotic consistency of bootstrap confidence interval for OPE. Our analysis improves their work in the following aspects. First, we study FQE with linear function approximation while Kostrikov & Nachum (2020) only considered the tabular case. Second, we provide distributional consistency of bootstrapping FQE which is stronger than the consistency of confidence interval in Kostrikov & Nachum (2020).

2. Preliminary

Consider an episodic Markov decision process (MDP) that is defined by a tuple $M = (\mathcal{S}, \mathcal{A}, P, r, H)$. Here, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ is the probability of reaching state s' when taking action a in state s, $r : \mathcal{S} \times \mathcal{A} \to [0,1]$ is the reward function, and H is the length of horizon. A policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps states to a distribution over actions. The state-action value function (Q-function) is defined as, for $h = 1, \ldots, H$,

$$Q_h^\pi(s,a) = \mathbb{E}^\pi \left[\sum_{h'=h}^{H} r(s_{h'}, a_{h'}) \,\middle|\, s_h = s, a_h = a \right],$$

where $a_{h'} \sim \pi(\cdot \,|\, s_{h'}), s_{h'+1} \sim P(\cdot \,|\, s_{h'}, a_{h'})$ and $\mathbb{E}^\pi$ denotes expectation over the sample path generated under policy π. The Q-function satisfies the Bellman equation for policy π:

$$Q_{h-1}^\pi(s,a) = r(s,a) + \mathbb{E}\Big[V_h^\pi(s') \Big| s, a \Big],$$

where $s' \sim P(\cdot|s,a)$ and $V_h^\pi : \mathcal{S} \to \mathbb{R}$ is the value function defined as $V_h^\pi(s) = \int_a Q_h^\pi(s,a)\pi(a|s)\mathrm{d}a$.

Let $[n] = \{1, \ldots, n\}$. For a positive semidefinite matrix X, we denote $\lambda_{\min}(X)$ as the minimum eigenvalue of X. Denote $I_d \in \mathbb{R}^{d\times d}$ as a diagonal matrix with 1 as all the diagonal entry and 0 anywhere else.

Off-policy evaluation. Suppose that the batch data $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ consists of K independent episodes collected using an *unknown behavior policy* $\bar{\pi}$. Each episode, denoted as $\mathcal{D}_k = \{(s_h^k, a_h^k, r_h^k)\}_{h\in[H]}$, is a trajectory of H state-transition tuples. It is easy to generalize our analysis to multiple unknown behavior policies since our algorithms do not require the knowledge of the behavior policy. Let $N = KH$ be the total number of sample transitions; and we sometimes write $\mathcal{D} = \{(s_n, a_n, r_n)\}_{n\in[N]}$ for simplicity. The goal of OPE is to estimate the expected cumulative return (i.e., value) of a *target policy* π from a a fixed initial distribution ξ_1, based on the dataset $\mathcal{D}$. The value is defined as

$$v_\pi = \mathbb{E}^\pi\left[\sum_{h=1}^H r(s_h, a_h) \,\middle|\, s_1 \sim \xi_1\right].$$

Fitted Q-evaluation. Fitted Q-evaluation (FQE) is an instance of the fitted Q-iteration method, dated back to Fonteneau et al. (2013); Le et al. (2019). Let $\mathcal{F}$ be a given function class, for examples a linear function class or a neural network class. Set $\widehat{Q}_{H+1}^\pi = 0$. For $h = H, \ldots, 1$, we recursively estimate Q_h^π by regression and function approximation:

$$\widehat{Q}_h^\pi = \operatorname*{argmin}_{f\in\mathcal{F}}\left\{\frac{1}{N}\sum_{n=1}^N \Big(f(s_n,a_n) - y_n\Big)^2 + \lambda\rho(f)\right\},$$

where $y_n = r_n + \int_a \widehat{Q}_{h+1}^\pi(s_{n+1}, a)\pi(a|s_{n+1})\mathrm{d}a$ and $\rho(f)$ is a proper regularizer. The value estimate is

$$\widehat{v}_\pi = \mathbb{E}_{s\sim\xi_1, a\sim\pi(\cdot|s)}\Big[\widehat{Q}_1^\pi(x,a)\Big], \tag{2.1}$$

which can be directly computed based on $\widehat{Q}_1^\pi$. See the full description of FQE in Appendix A.1.

Off-policy inference. Let $\widehat{v}_\pi$ be an off-policy estimator of the target policy value v_π. In addition to the point estimator, we are primarily interested in the distribution of the off-policy evaluation error $\widehat{v}_\pi - v_\pi$. We aim to infer the error distribution of $\widehat{v}_\pi - v_\pi$ in order to conduct statistical inference. Suppose F is an estimated distribution of $\widehat{v}_\pi - v_\pi$. Then we can use F for a range of downstream off-policy inference tasks, for examples:

- *Moment estimation.* With F, we can estimate the p-th moment of $\widehat{v}_\pi - v_\pi$ by $\int x^p dF(x)$. Two important examples are bias estimation and variance estimation.
- *Confidence interval construction.* Define the quantile function of F as $\mathcal{G}(p) = \inf\{x \in \mathbb{R}, p \le F(x)\}$. Specify a confidence level $0 < \delta \le 1$. With F, we can construct the $1-\delta$ confidence interval as $[\widehat{v}_\pi - \mathcal{G}(1-\delta/2), \widehat{v}_\pi - \mathcal{G}(\delta/2)]$. If F is close to the true distribution of $\widehat{v}_\pi - v_\pi$, the above one would be the nearly tightest confidence interval for v_π based on $\widehat{v}_\pi$.
- *Evaluating multiple policies and estimating their correlation.* Suppose there are two target policies π_1, π_2 to evaluate and the corresponding off-policy estimators are $\widehat{v}_{\pi_1}, \widehat{v}_{\pi_2}$. Let F_{12} be the estimated joint distribution of $\widehat{v}_{\pi_1} - v_{\pi_1}$ and $\widehat{v}_{\pi_2} - v_{\pi_2}$. The Pearson correlation coefficient between the two estimators is

  $$\rho(\widehat{v}_{\pi_1}, \widehat{v}_{\pi_2}) = \frac{\mathrm{Cov}(\widehat{v}_{\pi_1}, \widehat{v}_{\pi_2})}{\sqrt{\mathrm{Var}(\widehat{v}_{\pi_1})\mathrm{Var}(\widehat{v}_{\pi_2})}}.$$

  Both the covariance and variance can be estimated from F_{12}, so we can further estimate the correlation between off-policy evaluators.

Remark 2.1 (Practical scenarios of estimating correlations)**.** Correlation is a basic statistical metric for comparing two estimators, and we used it as an example to illustrate that bootstrapping can be used for estimating a variety of statistics not limited to confidence intervals. In medical applications, we may have multiple target treatment policies to compare against, where a correlation estimate together with confidence intervals would make physicians better informed to make a fairer comparison.

3. Bootstrapping Fitted Q-Evaluation (FQE)

As shown in Le et al. (2019); Voloshin et al. (2019); Duan & Wang (2020); Paine et al. (2020), FQE not only demonstrates strong empirical performances, but also enjoys provably optimal theoretical guarantees. Thus it is natural to conduct bootstrapping on top of FQE for off-policy inference.

Recall the original dataset $\mathcal{D}$ consists of K episodes. We propose to bootstrap FQE *by episodes*: Draw sample episodes $\mathcal{D}_1^*, \ldots, \mathcal{D}_K^*$ independently with replacement from $\mathcal{D}$. This is the standard Efron's nonparametric bootstrap (Efron, 1982). Then we run FQE on the new bootstrapped set $\mathcal{D}^* = \{\mathcal{D}_1^*, \ldots, \mathcal{D}_K^*\}$ as in Eq. (2.1) and let the output $\widehat{v}_\pi^*$ as the bootstrapping FQE estimator. By repeating the above

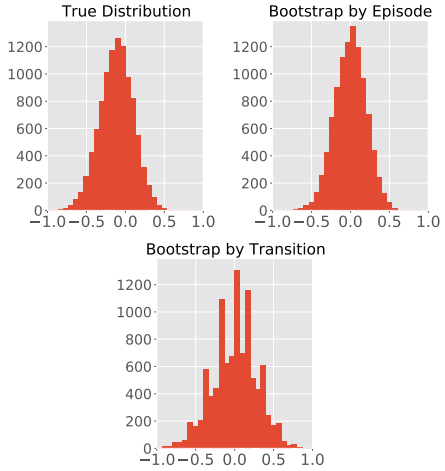

Figure 1. **Bootstrap by episodes vs. by sample transitions.** The first panel is the true FQE error distribution by Monte Carlo approximation. The second panel is the bootstrap distribution by episode while the third one is by sample transitions. Both behavior and target policies are the optimal policy. The number of Monte Carlo and bootstrap samples is 10000.

process, we may obtain multiple samples of $\widehat{v}_\pi^*$, and may use these samples to further conduct off-policy inference (see Section 6.2 for details).

3.1. Bootstrap by episodes *vs.* boostrap by sample transitions

Practitioners may wonder what is the right way to bootstrap a data set. This question is quite well understood in supervised learning when the data points are independent and identically distributed; there the best way to bootstrap is to resample data points directly. However, in episodic RL, although episodes may be generated independently from one another, sample transitions (s_n, a_n, r_n) in the same episode are highly dependent. Therefore, we choose to bootstrap the batch dataset *by episodes*, rather than *by sample transitions* which was commonly done according to previous literatures (Kostrikov & Nachum, 2020). We argue that bootstrapping by sample transitions may fail to correctly characterize the target error distribution of OPE. This is due to the in-episode dependence. To illustrate this phenomenon, we conduct numerical experiments using a toy Cliff Walking environment. We compare the true distribution of FQE error obtained by Monte Carlo sampling with error distributions obtained using bootstrapping FQE. Figure 1 clearly shows that the bootstrap distribution of $\widehat{v}_\pi^* - \widehat{v}_\pi$ (by episodes) closely approximates the true error distribution of $\widehat{v}_\pi - v_\pi$, while the bootstrap distribution by sample transition is highly irregular and incorrect. This validates our belief that it is necessary to bootstrap by episodes and handle dependent data carefully for OPE.

4. Asymptotic Distribution and Optimality of FQE

Before analyzing the use of bootstrap, we first study the asymptotic properties of FQE estimators. For the sake of theoretical abstraction, we focus our analysis on the FQE with linear function approximation, because it is the most basic and universal function approximation. We will show that the FQE error is asympotically normal and its asymptotic variance exactly matches the Cramér–Rao lower bound. All the proofs are deferred to Appendix A.3 and A.4.

Notations. Given a feature map $\phi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^d$, we let $\mathcal{F}$ be a linear function class spanned by ϕ. Without loss of generality, we assume $\|\phi(s,a)\|_\infty \le 1$ for any $(s,a) \in \mathcal{S}\times\mathcal{A}$. Define the Bellman operator for policy π as $\mathcal{P}^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ such that for any $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}, \mathcal{P}^\pi f(s,a) = \mathbb{E}_{s'\sim P(\cdot|s,a), a'\sim\pi(\cdot|s')}[f(s',a')]$. Denote the expected covariance matrix induced by the feature ϕ as $\Sigma = \mathbb{E}[\frac{1}{H}\sum_{h=1}^H \phi(s_h^1, a_h^1)\phi(s_h^1, a_h^1)^\top]$, where $\mathbb{E}$ is the expectation over population distribution generated by the behavior policy.

4.1. Asymptotic normality

We need a representation condition about the function class $\mathcal{F}$, which will ensure sample-efficient policy evaluation via FQE.

Condition 4.1 (Policy completeness). For any $f \in \mathcal{F}$, we assume $\mathcal{P}^\pi f \in \mathcal{F}$, and $r \in \mathcal{F}$.

Policy completeness requires the function class $\mathcal{F}$ can well capture the Bellman operator. It is crucial for the estimation consistency of FQE (Le et al., 2019; Duan & Wang, 2020) and implies the realizability condition $Q_h^\pi \in \mathcal{F}$ for $h \in [H]$. Recently, Wang et al. (2020) established a lower bound showing that the condition $Q_h^\pi \in \mathcal{F}$ alone is not enough for sample-efficient OPE. Thus we need the policy completeness condition in order to leverage the generalizability of linear function class.

Next we present our first main result. The theorem presents the asymptotic normality of FQE with linear function approximation. For any $h_1 \in [H], h_2 \in [H]$, define the cross-time-covariance matrix as

$$\Omega_{h_1,h_2} = \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^H \phi(s_{h'}^1, a_{h'}^1)\phi(s_{h'}^1, a_{h'}^1)^\top \varepsilon_{h_1,h'}^1 \varepsilon_{h_2,h'}^1\right],$$

where $\varepsilon^1_{h_1,h'} = Q^\pi_{h_1}(s^1_{h'}, a^1_{h'}) - (r^1_{h'} + V^\pi_{h_1+1}(s^1_{h'+1}))$.

Theorem 4.2 (Asymptotic normality of FQE). Suppose $\lambda_{\min}(\Sigma) > 0$ and Condition 4.1 holds. The FQE with linear function approximation is $\sqrt{N}$-consistent and asymptotically normal:

$$\sqrt{N}\,(\widehat{v}_\pi - v_\pi) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } N \to \infty, \tag{4.1}$$

where $\xrightarrow{d}$ denotes converging in distribution. The asymptotic variance σ^2 is given by

$$\begin{aligned}\sigma^2 =& \sum_{h=1}^{H} (\nu^\pi_h)^\top \Sigma^{-1} \Omega_{h,h} \Sigma^{-1} \nu^\pi_h \\ &+ 2 \sum_{h_1<h_2} (\nu^\pi_{h_1})^\top \Sigma^{-1} \Omega_{h_1,h_2} \Sigma^{-1} \nu^\pi_h,\end{aligned} \tag{4.2}$$

where $\nu^\pi_h = \mathbb{E}^\pi[\phi(s_h, a_h)|s_1 \sim \xi_1]$.

The proof is based on a decomposition of the FQE error $\sqrt{N}(\widehat{v}_\pi - v_\pi)$ into the sum of a primary term, which is a sum of the martingale differences, and two small-order terms that are asymptotically negligible. For the primary term, we utilize classical martingale central limit theorem (McLeish et al., 1974) to prove its asymptotic normality.

Remark 4.3. The second term on the right-hand side of Eq. (4.2) (cross-product term) characterizes the dependency between two different fitted-Q steps. When considering a tabular time-inhomogeneous MDP that was used in Yin & Wang (2020), this cross-product term disappears and the asymptotic variance becomes

$$\sum_{h=1}^{H} \mathbb{E}\left[\frac{\mu^\pi_h(s^1_h, a^1_h)^2}{\bar{\mu}_h(s^1_h, a^1_h)^2} (\varepsilon^1_{h,h})^2\right],$$

where $\bar{\mu}_h$ is the marginal distribution of (s^1_h, a^1_h) and μ^π_h is the marginal distribution of (s_h, a_h) under policy π. This matches the asymptotic variance term in Remark 3.2 of Yin & Wang (2020).

Next, we give a corollary about the joint asymptotic error distribution when evaluating multiple policies. Denote $\Pi = \{\pi_1, \ldots, \pi_L\}$ as a set of target policies to evaluate and denote $\widehat{v}_{\pi_k}$ as the FQE estimator of the policy π_k. For each $\pi_k \in \Pi$, let $\varepsilon^{1,k}_{h_1,h'} = Q^{\pi_k}_{h_1}(s^1_{h'}, a^1_{h'}) - (r^1_{h'} + V^{\pi_k}_{h_1+1}(s^1_{h'+1}))$. For any $h_1 \in [H], h_2 \in [H]$, denote

$$\Omega^{j,k}_{h_1,h_2} = \mathbb{E}\left[\frac{1}{H}\sum_{h'=1}^{H} \phi(s^1_{h'}, a^1_{h'})\phi(s^1_{h'}, a^1_{h'})^\top \varepsilon^{1,j}_{h_1,h'} \varepsilon^{1,k}_{h_2,h'}\right].$$

Corollary 4.4 (Multiple policies). Suppose the conditions in Theorem 4.2 hold. The set of FQE estimators converge in distribution to a multivariate Gaussian distribution:

$$\begin{pmatrix} \sqrt{N}\,(\widehat{v}_{\pi_1} - v_{\pi_1}) \\ \vdots \\ \sqrt{N}\,(\widehat{v}_{\pi_L} - v_{\pi_L}) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Gamma),$$

where the covariance matrix $\Gamma = (\sigma^2_{jk})_{j,k=1}^L \in \mathbb{R}^{L\times L}$ with

$$\begin{aligned}\sigma^2_{jk} =& \sum_{h=1}^{H} (\nu^{\pi_j}_h)^\top \Sigma^{-1} \Omega^{j,k}_{h,h} \Sigma^{-1} \nu^{\pi_k}_h \\ &+ 2 \sum_{h_1<h_2} (\nu^{\pi_j}_{h_1})^\top \Sigma^{-1} \Omega^{j,k}_{h_1,h_2} \Sigma^{-1} \nu^{\pi_k}_h.\end{aligned}$$

4.2. Asymptotic efficiency

An asymptotic efficient estimator has the minimal variance among all the unbiased estimator or its variance matches the Cramér–Rao bound asymptotically.

Theorem 4.5 (Linear Cramér–Rao lower bound). Under Condition 4.1 with linear function class, the variance of any unbiased OPE estimator is lower bounded by σ^2 defined in Eq. (4.2).

The above theorem implies FQE with linear function approximation is asymptotic efficient. Jiang & Li (2016) derived the first Cramér–Rao lower bound for *the tabular MDP* that depends the size of state and action spaces. Our lower bound is stronger in the sense that it only depends on the feature dimension d. Kallus & Uehara (2020) studied more general semiparametric efficiency bound but can not be directly applied to our case since they do not consider the policy completeness assumption.

5. Distributional Consistency of Bootstrapping FQE

In this section, we show that the bootstrapping FQE method is distributionally consistent. More precisely, we prove that, the bootstrap distribution of $\sqrt{N}(\widehat{v}^*_\pi - \widehat{v}_\pi)$, conditioned on data $\mathcal{D}$, asymptotically imitates the true error distribution $\sqrt{N}(\widehat{v}_\pi - v_\pi)$. Consequently, we may use the method to construct confidence regions with asymptotically correct and tight coverage. All the proofs are deferred to Appendix A.5 and A.6.

Suppose that the batch dataset $\mathcal{D}$ is generated from a probability space $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\mathcal{D})$, and the bootstrap weight W^* is from an independent probability space $(\mathcal{W}, \Omega, \mathbb{P}_W)$. Their joint probability measure is $\mathbb{P}_{\mathcal{D}W^*} = \mathbb{P}_\mathcal{D} \times \mathbb{P}_{W^*}$. Let $\mathbb{P}_{W^*|\mathcal{D}}$ denote the conditional distribution once the dataset $\mathcal{D}$ is given.

Theorem 5.1 (Distributional consistency). Suppose the same assumptions in Theorem 4.2 hold. Conditioned on $\mathcal{D}$, we have

$$\sqrt{N}\big(\widehat{v}_\pi^* - \widehat{v}_\pi\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \text{ as } N \to \infty, \tag{5.1}$$

where σ^2 is defined in Eq. (4.2). Consequently, it implies

$$\sup_{\alpha\in(0,1)} \Big|\mathbb{P}_{W^*|\mathcal{D}}\Big(\sqrt{N}\big(\widehat{v}_\pi^* - \widehat{v}_\pi\big) \le \alpha\Big) - \mathbb{P}_{\mathcal{D}}\Big(\sqrt{N}(\widehat{v}_\pi - v_\pi) \le \alpha\Big)\Big| \to 0.$$

Note that the convergence in distribution result applies to the sequence of probability measures $\mathbb{P}_{W^*|\mathcal{D}}$ where datasize grows to infinity. The proof of Theorem 5.1 uses techniques that are different from classical analysis of supervised learning. This is because FQE is a fixed-point iteration type algorithm and it has no objective function to minimize directly. This poses some difficulties to apply conventional bootstrap analysis. Thus, our proof utilizes the equivalence between FQE and a model-based plug-in estimator described in Appendix A.2, together with the Mallows metric (Bickel & Freedman, 1981; Freedman et al., 1981) and the multivariate delta theorem.

Theorem 5.1 sets the theoretical foundation for using bootstrapping for off-policy inference. Eq. (4.1) and Eq. (5.1) together show that the bootstrap error distribution converges to the same limit as the target error distribution of FQE, which are both asymptotically efficient and match the Cramér–Rao lower bound.

By using the distributional consistency of bootstrapping FQE, we may further construct consistent confidence intervals. Denote the lower δth quantile of bootstrap error distribution $q_\delta^\pi = \inf\{t : \mathbb{P}_{W^*|\mathcal{D}}(\widehat{v}_\pi^* - \widehat{v}_\pi \le t) \ge \delta\}$. Then we construct the $1-\delta$ confidence interval of the policy value by: $\text{CI}(\delta) = [\widehat{v}_\pi - q_{1-\delta/2}^\pi, \widehat{v}_\pi - q_{\delta/2}^\pi]$.

We next establish that the coverage probability of the percentile bootstrap confidence interval for v_π converges to the nominal level as a consequence of Theorem 5.1 and the consistency of bootstrap moment estimation.

Corollary 5.2 (Consistency of the coverage probability). Under the assumptions in Theorem 5.1, we have as $N \to \infty$, $\mathbb{P}_{\mathcal{D}W^*}(v_\pi \in \text{CI}(\delta)) \to 1 - \delta$.

Remark 5.3. Kostrikov & Nachum (2020) proved the consistency of bootstrap confidence interval in the tabular case. In contrast, our result is more general. We establish the distributional consistency for OPE with function approximation.

Corollary 5.4 (Consistency of the moment estimation). Suppose the assumptions in Theorem 5.1 holds and $\limsup_{N\to\infty} \mathbb{E}_{W^*|\mathcal{D}}[(\sqrt{N}(\widehat{v}_\pi^* - \widehat{v}_\pi))^q] < \infty$ for some $q > 2$. Then we have for any $1 \le r < q$,

$$\mathbb{E}_{W^*|\mathcal{D}}\Big[\big(\sqrt{N}(\widehat{v}_\pi^* - \widehat{v}_\pi)\big)^r\Big] \to \int t^r \mathrm{d}\mu(t),$$

where $\mu(\cdot)$ is the distribution of $\mathcal{N}(0, \sigma^2)$.

The consistency of bootstrap variance estimate is immediately implied by setting $r = 2$.

6. Subsampled Bootstrapping FQE

Computing bootstrap-based quantities can be prohibitively demanding as the data size grows. Inspired by recent developments from statistics community (Kleiner et al., 2014; Sengupta et al., 2016), we adapt a simple subsampled bootstrap procedure for FQE to accelerate the computation.

6.1. Subsampled bootstrap

Let the original dataset be $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$. For any dataset $\widetilde{\mathcal{D}}$, we denote by $\widehat{v}_\pi(\widetilde{\mathcal{D}})$ the FQE estimator based on dataset $\widetilde{\mathcal{D}}$ and B as the number of bootstrap samples. The subsampled bootstrap includes the following three steps. For each $b \in [B]$, we first construct a random subset $\mathcal{D}_{K,s}^{(b)}$ of s episodes where each sample episode is drawn independently *without replacement* from dataset $\mathcal{D}$. Typically $s = K^\gamma$ for some $0 < \gamma \le 1$.Then we generate a resample set $\mathcal{D}_{K,s}^{(b)*}$ of K episodes where each sample episode is drawn independently *with replacement* from $\mathcal{D}_{K,s}^{(b)}$. Note that when $s = K$, $\mathcal{D}_{K,s}^{(b)}$ is always equal to $\mathcal{D}$ such that the subsampled bootstrap reduces to vanilla bootstrap. In the end, we compute $\varepsilon^{(b)} = \widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)*}) - \widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)})$. Algorithm 1 gives the full description.

Remark 6.1 (Computational benefit). In Algorithm 1, although each run of FQE is still over a dataset of K episodes, only s of them are distinct. As a result, the runtime of running FQE on a bootstrapped set can be substantially reduced. With linear function approximation, one run of FQE requires solving H least square problems. Thus the total runtime complexity of the subsampled bootstrapping FQE is $O(B(K^{2\gamma}H^3d + Hd^3))$, where $0 < \gamma < 1$ controls the subsample size. When γ is small, we achieve significant speedup by an order of magnitude improvements.

Algorithm 1 Subsampled Bootstrapping FQE

input Dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, target policy π, subset size s, number of bootstrap samples B.

1: Compute FQE estimator $\widehat{v}_\pi(\mathcal{D})$ (Algorithm 2).
2: **for** $b = 1, \ldots, B$ **do**
3: Build a random subset $\mathcal{D}_{K,s}^{(b)}$.
4: Feed $\mathcal{D}_{K,s}^{(b)}$ to FQE and compute $\widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)})$.
5: Generate a resample set $\mathcal{D}_{K,s}^{(b)*}$.
6: Compute $\widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)*})$.
7: Compute $\varepsilon^{(b)} = \widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)*}) - \widehat{v}_\pi(\mathcal{D}_{K,s}^{(b)})$.
8: **end for**

output $\{\varepsilon^{(1)}, \ldots, \varepsilon^{(B)}\}$.

6.2. Off-policy inference via bootstrapping FQE

We describe how to conduct off-policy inference based on the output of Algorithm 1.

- **Bootstrap variance estimation.** To estimate the variance of FQE estimators, we calculate the bootstrap sample variance as

$$\widehat{\mathrm{Var}}(\widehat{v}_\pi(\mathcal{D})) = \frac{1}{B-1}\sum_{b=1}^{B}(\varepsilon^{(b)} - \bar{\varepsilon})^2,$$

  where $\bar{\varepsilon} = \frac{1}{B}\sum_{b=1}^{B}\varepsilon^{(b)}$.

- **Bootstrap confidence interval.** Compute the $\delta/2$ and $1-\delta/2$ quantile of the empirical distribution $\{\varepsilon^{(1)}, \ldots, \varepsilon^{(B)}\}$, denoted as $\widehat{q}^{\pi}_{\delta/2}, \widehat{q}^{\pi}_{1-\delta/2}$ respectively. The percentile bootstrap confidence interval is $[\widehat{v}_\pi(\mathcal{D}) - \widehat{q}^{\pi}_{1-\delta/2}, \widehat{v}_\pi(\mathcal{D}) - \widehat{q}^{\pi}_{\delta/2}]$.

- **Bootstrap correlation estimation.** For any of two target policies π_1 and π_2, we want to estimate the Pearson correlation coefficient between their FQE estimators. The bootstrap sample correlation can be computed as $\widehat{\rho}(\widehat{v}_{\pi_1}(\mathcal{D}), \widehat{v}_{\pi_2}(\mathcal{D})) =$

$$\frac{\sum_{b=1}^{B}(\varepsilon_1^{(b)} - \bar{\varepsilon}_1)(\varepsilon_2^{(b)} - \bar{\varepsilon}_2)}{\sqrt{\sum_{b=1}^{B}(\varepsilon_1^{(b)} - \bar{\varepsilon}_1)^2}\sqrt{\sum_{b=1}^{B}(\varepsilon_2^{(b)} - \bar{\varepsilon}_2)^2}}.$$

7. Experiments

In this section, we numerically evaluate the proposed bootstrapping FQE method in several RL environments. For constructing confidence intervals, we fix the confidence level at $\delta = 0.1$. For estimating variance and correlations, we average the results over 200 trials. More details about the experiment are given in Appendix C.

7.1. Experiment with tabular discrete environment

We first consider the Cliff Walking environment (Sutton & Barto, 2018), with artificially added randomness to create stochastic transitions (see Appendix C for details). The target policy is chosen to be a near-optimal policy, trained using Q-learning. Consider three choices of the behavior policy: the same as the target policy (on-policy), 0.1 ϵ-greedy policy and soft-max policy with temperature 1.0 based on the learned optimal Q-function. The results for soft-max policy and correlation estimation are deferred to Appendix C.

We test three different methods. The first two methods are subsampled bootstraping FQE with subsample sizes $s = K$ (the vanilla bootstrap) and $s = K^{0.5}$ (the computational-efficient version), where $B = 100$. The third method is the high-confidence off-policy evaluation (HCOPE) (Thomas et al., 2015), which we use as a baseline for comparison. HCOPE is a method for constructing off-policy confident interval for tabular MDP, and it is based concentration inequalities and has provable coverage guarantee. We also compare these methods with the oracle confidence interval (which is the true distribution's quantile obtained by Monte Carlo simulation).

Coverage and tightness of off-policy confidence interval (CI). We study the empirical coverage probability and interval width with different number of episodes. Figure 2 shows the result under different behavior policies. In the left panel of Figure 3, we report the effect of the number of bootstrap samples on empirical coverage probability (ϵ-greedy behavior policy, $K = 100$). It is clear that the empirical coverage of our confidence interval based on bootstrapping FQE becomes increasingly close to the expected coverage ($= 1 - \delta$) as the number of episodes increases. The width of bootstrapping-FQE confidence interval is significantly tighter than that of the HCOPE and very close to the oracle one. It is worth noting that, even in the on-policy case, our bootstrap-based confidence interval still has a clear advantage over the concentration-based confidence interval. The advantage of our method comes from that it fully exploits the distribution information. However, bootstrap confidence interval tends to be under-estimate when the number of episodes is extremely small ($K = 10$). Thus we suggest the practitioner to use bootstrap methods when the sample size is moderately large ($K > 50$).

Further, the subsampled bootstrapping FQE demonstrates a competitive performance as well as significantly reduced computation time. The saving in computation time becomes

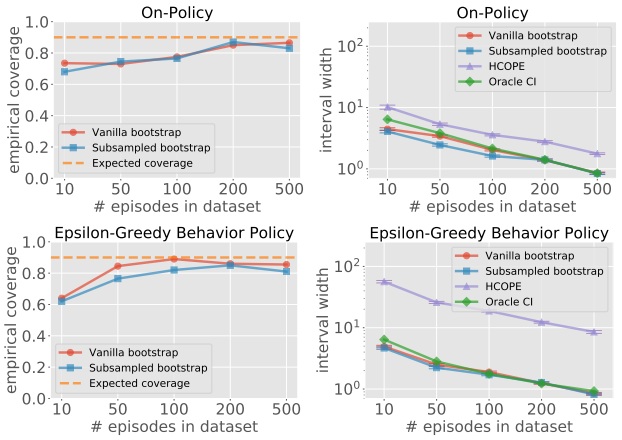

Figure 2. **Off-policy CI for Cliff Walking.** Left: Empirical coverage probability of CI; Right: CI width under different behavior policies. Boostrapping-FQE confidence interval method demonstrates better and tighter coverage of the groundtruth. It closely resembles the oracle confidence interval which comes from the true error distribution.

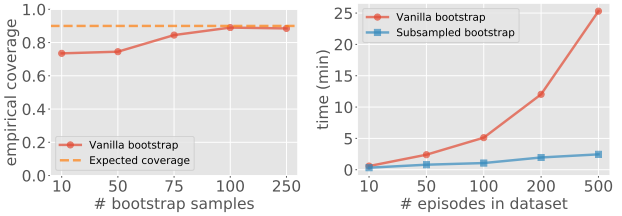

Figure 3. **Sample and time efficiency of bootstrapping FQE.** Left: Empirical coverage of bootstrapping-FQE CI, as #bootstrap samples increases. Right: Runtime of bootstrapping FQE, as data-size increases (with subsample size $s = K^{0.5}$).

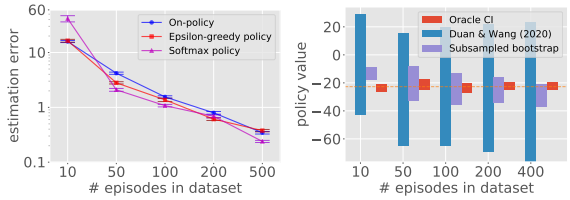

Figure 4. **Bootstrapping for variance estimation and with function approximation** Left: Error of variance estimates for tabular case, as data size increases. Right: Confidence interval constructed using bootstrapping FQE with linear function approximation.

increasingly substantial as the data gets big; see the right panel of Figure 3.

Bootstrapping FQE for variance estimation. We study the performance of variance estimation using subsampled bootstrapping FQE under three different behavior policies. We vary the number of episodes and the true $\mathrm{Var}(\widehat{v}_\pi(\mathcal{D}))$ is computed through Monte Carlo method. We report the estimation error of $\widehat{\mathrm{Var}}(\widehat{v}_\pi(\mathcal{D})) - \mathrm{Var}(\widehat{v}_\pi(\mathcal{D}))$ across 200 trials in the left panel of Figure 4.

7.2. Experiment with Mountain Car using linear function approximation

Next we test the methods on the classical Mountain Car environment (Moore, 1990) with linear function approximation. We artificially added a Gaussian random force to the car's dynamics to create stochastic transitions. For the linear function approximation, we choose 400 radial basis functions (RBF) as the feature map. The target policy is chosen as the optimal policy trained by Q-learning; and the behavior policy is chosen to be the 0.1 ϵ-greedy policy based on the learned optimal Q-function.

For comparison, we compute an empirical Bernstein-inequality-based confidence interval (Duan & Wang, 2020), which to our best knowledge is the only provable CI based on FQE with function approximation (see Appendix C for its detailed form). We also compute the oracle CI using Monte Carlo simulation. Figure 4 right give all the results. According to the results, our method demonstrates good coverage of the groundtruth and is much tighter than the concentration-based CI, even both of them use linear function approximation.

7.3. Experiment with septic management using neural nets for function approximation

Lastly, we consider a real-world healthcare problem for treating sepsis in the intensive care unit (ICU). We use the septic management simulator by Oberst & Sontag (2019) for our study. It simulates a patient's vital signs, e.g. the heart rate, blood pressure, oxygen concentration, and glucose levels, with three treatment actions (antibiotics, vasopressors, and mechanical ventilation) to choosen from at each time step. The reward is +1 when a patient is discharged and -1 if the patient reaches a life critical state.

We apply the bootstrapping FQE using neural network function approximator with three fully connected layers, where the first layer uses 256 units and a Relu activation function, the second layer uses 32 units and a Selu activation function, and the last layer uses Softsign. The network takes as input the state-action pair (a 11-dim vector) and outputs a Q-value estimate. Let the behavior policy be the 0.15 ϵ-greedy policy.

We evaluate two policies based on the same set of data. This is very common in healthcare problem since we may have multiples treatments by the doctor. One target policy is fixed

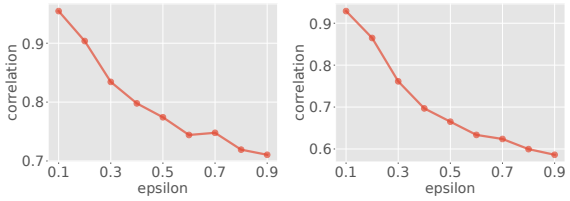

Figure 5. **Bootstrapping FQE with neural nets for estimating the correlation between two FQE estimators.** The left panel is using 300 episodes, while the right panel is using 500 episodes.

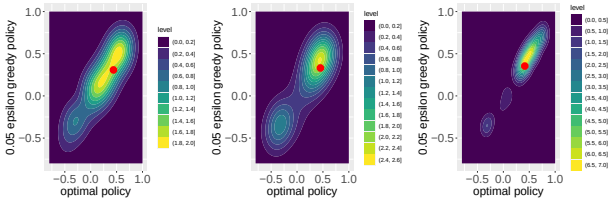

Figure 6. **Estimated confidence region for evaluating two policies using bootstrapping FQE with neural networks.** Two target policies are optimal policy and 0.15 ϵ-greedy policy. Red point are true values of those two target policies. From left to right, the sample sizes are $K = 100, 300, 500$. Here we use the geom density() function to illustrate the density of 2-D bootstrap estimators.

to be the optimal policy while we vary the other one with different ϵ-greedy noise. We expect the correlation decreases as the difference between two target policies increases. Figure 5 is well aligned with our expectation. In Figure 6, we plot the confidence region of two target policies obtained by bootstrapping FQE using neural networks. According to Figures 5 and 6, the bootstrapping FQE method can effectively construct confidence regions and correlation estimates, even when using neural networks for function approximation. These results suggest that the proposed bootstrapping FQE method reliably achieves off-policy inference, with more general function approximators.

8. Conclusion

This paper studies bootstrapping FQE for statistical off-policy inference and establishes its asymptotic distributional consistency as a theoretical benchmark. Our experiments suggest that bootstrapping FQE is effective and efficient in a range of tasks, from tabular problems to continuous problems, with linear and neural network approximation.

Acknowledgements

Csaba Szepesvári gratefully acknowledges funding from the Canada CIFAR AI Chairs Program, Amii and NSERC. Mengdi Wang gratefully acknowledges funding from the U.S. National Science Foundation (NSF) grant CMMI1653435, Air Force Office of Scientific Research (AFOSR) grant FA9550-19-1-020, and C3.ai DTI.

References

Bickel, P. J. and Freedman, D. A. Some asymptotic theory for the bootstrap. *The annals of statistics*, pp. 1196–1217, 1981.

Dai, B., Nachum, O., Chow, Y., Li, L., Szepesvári, C., and Schuurmans, D. Coindice: Off-policy confidence interval estimation. *arXiv preprint arXiv:2010.11652*, 2020.

Duan, Y. and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. *Internation Conference on Machine Learning*, 2020.

Eck, D. J. Bootstrapping for multivariate linear regression models. *Statistics & Probability Letters*, 134:141–149, 2018.

Efron, B. *The jackknife, the bootstrap and other resampling plans*. SIAM, 1982.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. Accountable off-policy evaluation with kernel bellman statistics. *Proceedings of the International Conference on Machine Learning*, 2020.

Feng, Y., Tang, Z., Zhang, N., and Liu, Q. Non-asymptotic confidence intervals of off-policy evaluation: Primal and dual bounds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dKg5D1Z1Lm.

Fonteneau, R., Murphy, S. A., Wehenkel, L., and Ernst, D. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208(1):383–416, 2013.

Freedman, D. A. et al. Bootstrapping regression models. *The Annals of Statistics*, 9(6):1218–1228, 1981.

Hallak, A. and Mannor, S. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1372–1383. JMLR. org, 2017.

Hanna, J. P., Stone, P., and Niekum, S. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Hao, B., Abbasi-Yadkori, Y., Wen, Z., and Cheng, G. Bootstrapping upper confidence bound. *Thirty-fourth Annual Conference on Neural Information Processing Systems*, 2020a.

Hao, B., Duan, Y., Lattimore, T., Szepesvári, C., and Wang, M. Sparse feature selection makes batch reinforcement learning more sample efficient. *arXiv preprint arXiv:2011.04019*, 2020b.

Jiang, N. and Huang, J. Minimax value interval for off-policy evaluation and policy optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167): 1–63, 2020.

Kato, K. A note on moment convergence of bootstrap m-estimators. *Statistics & Risk Modeling*, 28(1):51–61, 2011.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pp. 795–816, 2014.

Kostrikov, I. and Nachum, O. Statistical bootstrapping for uncertainty estimation in off-policy evaluation. *arXiv preprint arXiv:2007.13609*, 2020.

Kuzborskij, I., Vernade, C., György, A., and Szepesvári, C. Confident off-policy evaluation and selection through self-normalized importance weighting. *arXiv preprint arXiv:2006.10460*, 2020.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *Journal of machine learning research*, 4(Dec): 1107–1149, 2003.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *Proceedings of Machine Learning Research*, 97:3703–3712, 2019.

Liao, P., Klasnja, P., and Murphy, S. Off-policy estimation of long-term average outcomes with applications to mobile health. *arXiv preprint arXiv:1912.13088*, 2019.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

McLeish, D. L. et al. Dependent central limit theorems and invariance principles. *the Annals of Probability*, 2(4): 620–628, 1974.

Moore, A. W. Efficient memory-based learning for robot control. 1990.

Munos, R. and Szepesvári, C. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9 (5), 2008.

Nachum, O., Chow, Y., Dai, B., and Li, L. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems*, pp. 2315–2325, 2019.

Oberst, M. and Sontag, D. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pp. 4881–4890, 2019.

Paine, T. L., Paduraru, C., Michi, A., Gulcehre, C., Zolna, K., Novikov, A., Wang, Z., and de Freitas, N. Hyperparameter selection for offline reinforcement learning. *arXiv preprint arXiv:2007.09055*, 2020.

Petersen, K. and Pedersen, M. The matrix cookbook. technical university of denmark. *Technical Manual*, 2008.

Precup, D., Sutton, R. S., and Singh, S. Eligibility traces for off-policy policy evaluation. In *ICML'00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

Sengupta, S., Volgushev, S., and Shao, X. A subsampled double bootstrap for massive data. *Journal of the American Statistical Association*, 111(515):1222–1232, 2016.

Shi, C., Zhang, S., Lu, W., and Song, R. Statistical inference of the value function for reinforcement learning in infinite horizon settings. *arXiv preprint arXiv:2001.04515*, 2020.

Singh, K. On the asymptotic accuracy of efron's bootstrap. *The Annals of Statistics*, pp. 1187–1195, 1981.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148, 2016.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388. PMLR, 2015.

Uehara, M. and Jiang, N. Minimax weight and Q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

Xie, T., Ma, Y., and Wang, Y.-X. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems*, pp. 9665–9675, 2019.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. *International Conference on Artificial Intelligence and Statistics*, 2020.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. GenDICE: Generalized offline estimation of stationary values. *arXiv preprint arXiv:2002.09072*, 2020a.

Zhang, S., Liu, B., and Whiteson, S. GradientDICE: Rethinking generalized offline estimation of stationary values. *arXiv preprint arXiv:2001.11113*, 2020b.