Asymptotic Equivalence of Direct Method and Marginal Importance Weighting in Offline Policy Evaluation under Unrealizability

Anonymous Author(s)
Affiliation
Address
email

Abstract

We consider the problem of offline policy evaluation (OPE) with Markov decision processes (MDPs), where the goal is to estimate the utility of given decision-making policies on the basis of static datasets. Recently, theoretical understanding of OPE has been advanced under the (approximate) realizability assumptions. However, such assumptions undermine the applicability of the results since the given environmental models may be completely wrong. We study the behavior of an existing direct estimator with linear function approximation under unrealizability, where the environment of interest may be not at all well approximated with the given function space. Consequently, we obtain an asymptotically exact characterization of the OPE error and find out a novel error-controlling term interpreted as the regression error of marginal density ratio. Leveraging this result, we also establish the nonparametric consistency of the tile-coding estimators under quite mild assumptions.

1 Introduction

We consider the problem of offline data-driven decision optimization, wherein static records of previous interactions between decision makers and the environmental system of interest are given. The possible application areas include autonomous driving vehicles, natural-language dialogue systems, recommender systems, financial portfolio optimization and healthcare treatment optimization.

The framework of offline reinforcement learning (RL) is one of the promising approaches to this task (Levine et al., 2020). In the standard RL, the environment and the decision-making policy are respectively modeled as Markov decision processes (MDPs) $\mathcal{M}$ and conditional distributions of actions π (Sutton and Barto, 2018), where each series of consecutive interactions between $\mathcal{M}$ and π are abstracted as a stochastic sequence of state s, action a and reward r, called an *episode*. The objective of the offline RL is then formalized as the maximization of the policy value $J(\pi)$, the expected value of the total reward obtained from a single episode, given a static dataset of previous interactions.

The crucial part of the problem is that the dataset is static; No additional interaction with the environment is allowed. This constraint poses several unique challenges to the problem. First, the policies we are optimizing, i.e., the target policies, cannot be run in the actual environment. Second, the policies used to generate the dataset, i.e., the behavior policies, are often unknown and may be totally different from the target policies. Consequently, it is even difficult to accurately estimate the value of target policies. This is problematic especially in consideration of real-life applications involving financial costs and healthcare risks.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

To address the issue of policy value estimation, the problem of offline policy evaluation (OPE) have been extensively studied in the literature. A class of OPE algorithms are referred as the direct methods (DMs), in which some characteristics of $\mathcal{M}$ are assumed to be *realizable* under some hypothetical models and $J(\pi)$ is estimated via a direct estimation of such characteristics. For example, in the fitted Q-evaluation (FQE) algorithm (Le et al., 2019), the policy Q function is assumed to be well-approximated with a parametric function class and the OPE is reduced to the estimation of its parameters.

DMs are known to be empirically effective (Fu et al., 2021) if such realizability assumptions are satisfied and, more importantly, *the converse is also true* (Voloshin et al., 2019). However, the theoretical understanding of DMs under unrealizability is still in its active development. For example, several authors have recently studied OPE or offline RL under relatively weak or approximate realizability assumptions (Jin et al., 2020; Xie and Jiang, 2020; Wang et al., 2021) and consequently proposing new algorithms.

In this paper, we approach the problem of unrealizability in the opposite direction; we start with an existing OPE method, study its behavior under complete unrealizability and seek for the possibility of regaining its consistency (i.e., asymptotically achieving zero errors). More specifically, we investigate the properties of a simple DM with linear function approximation, which is equivalent with a number of existing algorithms such as LSTD*Q* (Lagoudakis and Parr, 2003), FQE (Le et al., 2019) with linear function regressors, the marginalized importance sampling estimator (Yin and Wang, 2020) and DualDICE (Nachum et al., 2019) in tabular settings.

In particular, we first characterize the exact asymptotic error of the linear DM under as weak assumptions as possible. It turns out the error is governed by two approximation residuals $\mathcal{R}_B$ and $\mathcal{R}_\chi$,

$$\hat{J}(\pi) - J(\pi) = \mathbb{E}\left[\mathcal{R}_B(s,a)\mathcal{R}_\chi(s,a)\right] + \mathcal{O}\left(1/\sqrt{n}\right), \qquad \text{(informal)}$$

where they are corresponding to the unrealizable components of *the Bellman operator* and *the marginal density ratio*, respectively. While the closedness of the Bellman operator (i.e., $\mathcal{R}_B = 0$) has been well-known as a sufficient condition of the consistency, the hidden factor $\mathcal{R}_\chi$ sheds light on a new interpretation of the linear DMs as marginal density ratio estimators, indicating that the OPE error $\hat{J}(\pi) - J(\pi)$ can be controlled with the error of the density estimation problem. Leveraging the above finding, we also show that a linear DM with the tile-coding function approximation (Section 8.3.2, Sutton and Barto (2018)) is consistent under surprisingly mild conditions with appropriate tile-size scheduling.

The rest of the paper is organized as follows. In Section 2, we formalize the problem setting as well as the definition of the linear direct estimators. In Section 3, we present the main results, i.e., the asymptotic error analysis of the linear direct estimators and a construction of consistent nonparametric estimators as its application. In Section 4, we discuss related works with comparison to our results. Finally, in Section 5, we present concluding remarks, limitations and future directions. All the proofs of the propositions and the theorem are relegated to the appendix. See Section F for the proofs of the propositions. For the theorem, we present a proof sketch and the pointer to the full proof.

2 Preliminary

In Section 2.1, some notational conventions are introduced. The problem of OPE is then formalized in Section 2.2. Then, Section 2.3, 2.4 and 2.5 respectively introduce assumptions and definitions on the data-collecting processes, the environmental models and the class of estimators we will examine.

2.1 Basic Notation

We implicitly assume the spaces we encounter in this paper, such as the state space $\mathcal{S}$ and the action space $\mathcal{A}$, are equipped with respective metrics and base measures, each of which is a compact subset of either a Euclidean space with the Lebesgue measure, a discrete space with the counting measure or a product of those. We denote by $\int_{\mathcal{X}} f(x)\,\mathrm{d}x$ the integration of function f with respect to the base measure of $\mathcal{X}$. The subscript $\mathcal{X}$ may be omitted if it is obvious from the context. This way we can immediately generalize our results to both continuous and discrete spaces. Also, we denote the expectation of function f with respect to probability density p by $\langle f \rangle_p \coloneqq \mathbb{E}_{x\sim p}[f(x)]$.

Let $[m] \coloneqq \{1, 2, \ldots, m\}$ denote the set of integers from 1 to m. Let $\|\cdot\|_p$ denote the ℓ^p-norm for vectors and $\|A\|_{p\to q} \coloneqq \sup_{x\neq 0} \|Ax\|_q / \|x\|_p$ the operator norms for matrices, with the convention $\|A\|_p \coloneqq \|A\|_{p\to p}$, for all $1 \leq p, q \leq \infty$. Also, for any measurable function (including random variables) $f : \mathcal{X} \to \mathbb{R}$, its $L^p(\mathcal{X})$-norm is denoted by $\|f\|_p = [\int |f(x)|^p \, \mathrm{d}x]^{1/p}$ for $1 \leq p < \infty$. The essential supremum is denoted by $\|f\|_\infty$.

2.2 Problem Setting

The goal of OPE is to estimate the *value* of decision-making strategy based on a static dataset of interactions with the environment of interest, without directly knowing its mechanism.

The environment is modeled as a Markov decision process (MDP) $\mathcal{M} \equiv (\mathcal{S}, \mathcal{A}, p_0, p_T, p_r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ the action space, $p_0(s)$ the initial state probability, $p_T(s'|s,a)$ the transition probability and $p_r(r|s,a)$ the $[0,1]$-valued reward probability density function for $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$, $r \in [0,1]$. Here we assume p_T and p_r are unknown. On the other hand, the decision-making strategy is modeled as a *policy*, a state-conditional action distribution $\pi(a|s)$ for $s \in \mathcal{S}$, $a \in \mathcal{A}$.

The value of π is measured with the expected cumulative reward

$$J(\pi) \coloneqq \sum_{h=0}^{\infty} \gamma^h \left\langle P^h \bar{r} \right\rangle_{p_0^\pi}, \tag{1}$$

where $\gamma \in [0,1)$ is the discounting factor, P is the state-transition operator such that $(Pf)(s,a) = \int f(s',a') p_T(s'|s,a)\pi(a'|s') \, \mathrm{d}s' \, \mathrm{d}a'$, $\bar{r}(s,a) \coloneqq \int r \, p_r(r|s,a) \, \mathrm{d}r$ is the expected reward function, and $p_0^\pi(s,a) \coloneqq p_0(s)\pi(a|s)$ is the initial state-action distribution. In particular, $\langle P^h \bar{r} \rangle_{p_0^\pi}$ denotes the expected reward after h transitions starting from p_0^π.

The policy value $J(\pi)$ is estimated based on a collection of transition records $\xi^n \equiv (\xi_1, ..., \xi_n) \in \mathcal{D}^n$ called an *offline* dataset, where $\mathcal{D} \coloneqq \mathcal{S} \times \mathcal{A} \times [0,1] \times \mathcal{S}$ is the space of transition records and $\xi_i \equiv (s_i, a_i, r_i, s_i') \in \mathcal{D}$, $i \in [n]$, is a transition record made of a preceding state-action pair (s_i, a_i), the associated reward r_i, and the state after transition s_i'. The dataset ξ^n is assumed to be an instantiation of the random variables $\Xi^n \equiv (\Xi_1, ..., \Xi_n)$, $\Xi_i \equiv (S_i, A_i, R_i, S_i')$, collected with interactions between the environment $\mathcal{M}$ and a *query distribution* $p_{\mathrm{query}} \equiv \{p_{\mathrm{query}(i)}(s, a|\xi^{i-1})\}_{i\in[n]}$ such that its distribution is given in a conditional fashion,

$$p(\xi_i|\xi^{i-1}) = p_{\mathrm{query}(i)}(s_i, a_i|\xi^{i-1}) \, p_r(r_i|s_i, a_i) \, p_T(s_i'|s_i, a_i), \quad i \in [n].$$

Note that the notion of query distribution is so flexible that it admits ξ^n to be a union of episodes generated with multiple nonstationary policies and even adversaries on the choice of state-action pairs.

Definition 1 (OPE problem)**.** *An instance of the offline policy evaluation problem is specified with $\mathcal{P}_{\mathrm{OPE}} \equiv (\mathcal{M}, \pi, \gamma, p_{\mathrm{query}})$, where the goal is to estimate the policy value $J(\pi)$ determined by $(\mathcal{M}, \pi, \gamma)$, given the input data ξ^n generated with $(\mathcal{M}, p_{\mathrm{query}})$, without knowing any of p_T, p_r or p_{query}.*

2.3 Assumptions on Data-Collecting Process

To ensure the existence of reasonable estimators for $\mathcal{P}_{\mathrm{OPE}}$, we pose a couple of conditions on the data-collecting process, i.e., conditions on p_{query}. First, we assume the amounts of mutual dependencies induced by p_{query} between time-distant transition records are bounded.

Assumption 1 (G^*-mixing dataset)**.** *There exists a constant $G^* < \infty$ such that Ξ^n is 'ϕ'-strong mixing with the coefficient $g(h)$ satisfying $1 + 2\sum_{h=1}^{n} \sqrt{g(h)} \leq G^*$.*[1]

See Definition 16 (in the appendix) for the definition of the 'ϕ'-strong mixing coefficients. Typical examples satisfying Assumption 1 include datasets consisting of multiple short episodes and mixing Markov chains induced by stationary behavior policies.

Proposition 1. *The following are sufficient conditions of the G^*-mixing.*

[1]The symbol 'ϕ' in the 'ϕ'-strong mixing has its root in the statistics and completely unrelated to the feature mapping. We denote the mixing coefficient by $g(h)$ to avoid confusion with the feature mappings.

1. *Assume Ξ^n consists of multiple independently collected episodes with length bounded by H, ordered in a consecutive manner. Then we have $G^* \leq 2H - 1$.*

2. *Let $p_{\mathrm{query}(i)}(s, a|\xi^{i-1}) = p_T(s|s_{i-1}, a_{i-1})\pi_b(a|s)$, $1 \leq i \leq n$, for some stationary behavior policy $\pi_b(a|s)$ and assume the resulting Markov chain have a finite mixing time $t_{mix} < \infty$. Then we have $G^* \leq 1 + 7t_{mix}$.*

Note that the definition of G^*-mixing is designed to be more general than these examples. In particular, it is more suitable for our query-distribution framework, which admits adversaries behind the choice of (s_i, a_i)-s or dynamically changing behavior policies.

We also assume the query distribution p_{query} provides sufficient exploration in the state-action space. The degree of exploration is formalized with the notion of the marginal data density.

Definition 2 (Marginal data density)**.** *Let $\mu(s, a)$ be the marginal data density, given by $\mu(s, a) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[p_{\mathrm{query}(i)}(s, a \mid \Xi^{i-1})]$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$.*

Assumption 2 (Sufficient exploration)**.** *Let $c_\mu := \inf_{s\in\mathcal{S}, a\in\mathcal{A}} \mu(s, a)$. Then, we have $c_\mu > 0$.*

The marginal data density quantifies the expected frequency of visitation at each point $(s, a) \in \mathcal{S} \times \mathcal{A}$ made by the querying process. Thus, roughly speaking, Assumption 2 ensures that every location in the space is likely to be sampled in Ξ^n.

2.4 Environmental Model: Linear MDPs

We introduce linear MDPs, a simple class of the environmental models denoted by $\mathcal{H}_\phi$. We also define the projection of $\mathcal{M}$ onto $\mathcal{H}_\phi$ as we are concerned with the unrealizable case, $\mathcal{M} \notin \mathcal{H}_\phi$.

A linear MDP $\mathcal{H}_\phi$ is formally defined via a vector-valued function on the state-action space called a *feature mapping*, denoted by $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^K$, $K \geq 1$. We occasionally make some of the following assumptions on ϕ for technical reasons.

Assumption 3 (Boundedness)**.** $\sup_{s\in\mathcal{S}, a\in\mathcal{A}} \|\phi(s, a)\|_2 \leq 1$.

Assumption 4 (Irreducibility)**.** *$\phi_1, ..., \phi_K$ are linearly independent in $L^2(\mathcal{S} \times \mathcal{A})$.*

Assumption 5 (Alignment)**.** *There exists $v_0 \in \mathbb{R}^K$ such that $v_0^\top \phi(s, a) = 1, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}$.*

Note that if ϕ is not irreducible, one can always reduce the feature dimension to obtain irreducible one without hurting model expressibility. Moreover, if ϕ is not aligned, then we can always make it so by adding an extra 'bias' dimension whose value is always one, and the resulting ϕ is always irreducible. See Proposition 19 (in the appendix) for the rigorous arguments. A typical example of ϕ satisfying Assumption 3, 4 and 5 is the tabular features.

Remark 1. *ϕ is said to be tabular if there exits a K-partition of $\mathcal{S} \times \mathcal{A}$, $\{\mathcal{P}_k\}_{k\in[K]}$, such that*

$$\phi_k(s, a) = \mathbb{I}\{(s, a) \in \mathcal{P}_k\}, \ k \in [K], \ s \in \mathcal{S}, \ a \in \mathcal{A},$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. If ϕ is tabular and has no null (i.e., zero-volume) cell, it is bounded, aligned and irreducible.

Now, the class of ϕ-linear MDPs is defined as follows.

Definition 3 (ϕ-linear MDPs)**.** *We say $\mathcal{M}$ is ϕ-linear if there exist $b \in \mathbb{R}^K$ and $F \in \mathbb{R}^{K\times K}$ such that*

$$\bar{r}(s, a) = b^\top \phi(s, a), \qquad (P\phi)(s, a) = F\phi(s, a) \tag{2}$$

for almost every $s \in \mathcal{S}$ and $a \in \mathcal{A}$. We refer to the set of all the ϕ-linear MDPs as $\mathcal{H}_\phi$.

In other words, $\mathcal{M}$ is ϕ-linear if both the reward distribution and the transition dynamics are linearly predictable in expectation with respect to ϕ. This definition is motivated by the following proposition; if $\mathcal{M}$ is realizable as a member of $\mathcal{H}_\phi$, the problem of OPE is reduced to the estimation of b and F.

Proposition 2. *Under Assumption 3 and 4, if $\mathcal{M} \in \mathcal{H}_\phi$, we have*

$$J(\pi) = b^\top (I - \gamma F)^{-1} x_0. \tag{3}$$

Here, $x_0 := \int \phi(s, a)\, p_0^\pi(s, a)\, \mathrm{d}s\, \mathrm{d}a$.

However, it is not practical to assume we know the mapping ϕ that attains the realizability with the environment of interest $\mathcal{M}$. Instead, we introduce the projection of $\mathcal{M}$ onto $\mathcal{H}_\phi$.

Definition 4 (Discrepancy measure and projection)**.** *Let $D^2(b, F)$ be the parameter discrepancy of the ϕ-linearity, given by*

$$D^2(b, F) \coloneqq \mathbb{E}_{(s,a)\sim\mu}\left[\left|\bar{r}(s,a) - b^\top\phi(s,a)\right|^2 + \left|(P\phi)(s,a) - F\phi(s,a)\right|^2\right]. \tag{4}$$

We refer to its minimizer as the projection of $\mathcal{M}$ onto $\mathcal{H}_\phi$, denoted by $(b^\sharp, F^\sharp)$.

Note that $(b^\sharp, F^\sharp) = (b, F)$ if $\mathcal{M}$ is realizable. Throughout the paper, we consider the general cases in which (b, F) satisfying Definition 3 may not exist, but $(b^\sharp, F^\sharp)$ always does.

2.5 Linear Direct Estimators

We finally introduce the linear direct estimator. The idea of the linear direct estimator is twofold. First, we approximately solve the minimization of (4) based on the sample ξ^n to obtain the estimate of the projection, $(\hat{b}, \hat{F}) \approx (b^\sharp, F^\sharp)$. Then, we plug the estimate into (3) to get a policy value estimate, which seems reasonable if $\mathcal{M}$ is (approximately) realizable.

More precisely, the first step is formalized via the least squares method.

Definition 5. *The least squares solution of $\mathcal{H}_\phi$ with respect to ξ^n is defined as*

$$(\hat{b}, \hat{F}) \coloneqq \operatorname*{argmin}_{b\in\mathbb{R}^K, F\in\mathbb{R}^{K\times K}} \mathcal{C}(b, F; \xi^n),$$

where $\mathcal{C}(b, F; \xi^n)$ is the cost function given by

$$\mathcal{C}(b, F; \xi^n) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left[\left|r_i - b^\top\phi(s_i, a_i)\right|^2 + \left|\psi_\pi(s_i') - F\phi(s_i, a_i)\right|^2\right].$$

Here, $\psi_\pi(s) \coloneqq \int \phi(s,a)\pi(a|s)\,\mathrm{d}a$ is the state-marginal feature mapping.

This definition is justified as follows.

Proposition 3. *For all $b \in \mathbb{R}^K$ and $F \in \mathbb{R}^{K\times K}$, $\nabla_{b,F}\mathbb{E}[\mathcal{C}(b, F; \Xi^n)] = \nabla_{b,F} D^2(b, F)$.*

In other words, the gradient of the cost function coincides with that of the parameter discrepancy function in expectation and thus one can expect $(b, F) \to (b^\sharp, F^\sharp)$ in the large sample limit.

We have a closed form of the least squares solution.

Proposition 4. *Let $(\Phi, \Psi_\pi, \hat{r})$ be given by*

$$\Phi \coloneqq [\phi(s_1, a_1), ..., \phi(s_n, a_n)]^\top, \qquad \Psi_\pi \coloneqq [\psi_\pi(s_1'), ..., \psi_\pi(s_n')]^\top, \qquad \hat{r} \coloneqq [r_1, ..., r_n]^\top.$$

Then, the least squares solution is given by

$$\hat{b} = \frac{1}{n}\hat{\Sigma}^{-1}\Phi^\top\hat{r}, \qquad \hat{F} = \frac{1}{n}\Psi_\pi^\top\Phi\hat{\Sigma}^{-1}, \tag{5}$$

where $\hat{\Sigma} \coloneqq \frac{1}{n}\Phi^\top\Phi$ is the empirical covariance matrix.

The whole procedure is summarized in Algorithm 1. Note that it has no valid output if $\hat{\Sigma}$ or $I - \gamma\hat{F}$ is singular. Also $\psi_\pi(s)$ in Ψ_π and x_0 is not necessarily tractable in a closed form. One can always resort to Monte-Carlo estimates $\psi_\pi(s) \approx \frac{1}{n_\psi}\sum_{\ell=1}^{n_\psi}\phi(s, a_\ell)$, where $a_\ell \sim \pi(a|s)$, $\ell \in [n_\psi]$, are i.i.d. samples. Proposition 3 still holds under this approximation.

Algorithm 1 is equivalent to the LSTDQ (Lagoudakis and Parr, 2003) algorithm and a number of equivalence relationships to recent OPE estimators are drawn in Duan et al. (2020). For the completeness, we show Algorithm 1 is equivalent to the limit of Fitted Q-Evaluation (Le et al., 2019) with linear function approximators, shown in Algorithm 2.

Proposition 5. *The output of Algorithm 1, $\hat{J}(\pi)$, is identical to the limit of that of Algorithm 2, $\lim_{H\to\infty}\hat{J}_H(\pi)$, if both exist.*

Algorithm 1 Linear Direct OPE

Input: Initial distribution p_0, target policy π, data ξ^n, feature mapping ϕ

Output: Policy value estimate $\hat{J}(\pi)$

1: Compute the least squares solution $(\hat{b}, \hat{F})$ according to Proposition 4.

2: Compute $\hat{J}(\pi) = \hat{b}^\top (I - \gamma \hat{F})^{-1} x_0$, where $x_0 \coloneqq \int \psi_\pi(s) p_0(s) \, \mathrm{d}s$.

Algorithm 2 Linear Fitted Q-Evaluation

Input: Initial distribution p_0, target policy π, data ξ^n, feature mapping ϕ, iteration number H

Output: Policy value estimate $\hat{J}_H(\pi)$

1: Let $q_0 \coloneqq 0 \in \mathbb{R}^K$.

2: **for** $h = 1, 2, ..., H$ **do**

3: Find $q_h \coloneqq \operatorname{argmin}_{q \in \mathbb{R}^K} \frac{1}{n} \sum_{i=1}^n \left| r_i + \gamma q_{h-1}^\top \psi_\pi(s_i') - q^\top \phi(s_i, a_i) \right|^2$.

4: **end for**

5: Compute $\hat{J}_H(\pi) = q_H^\top x_0$, where $x_0 \coloneqq \int \psi_\pi(s) p_0(s) \, \mathrm{d}s$.

3 Main Results

First, we give an asymptotic characterization of the error $\hat{J}(\pi) - J(\pi)$, which sheds light on a hidden error-controlling factor related to the smoothness of a density ratio function. Second, leveraging the first result, we show novel consistency properties of a simple tile-coding estimator.

3.1 Asymptotic Bias of Linear Direct Estimators

As will be shown later, the dominant term of the OPE error is written as an inner product of two functions, namely the χ-residual function $\mathcal{R}_\chi$ and the Bellman residual function $\mathcal{R}_B$. To introduce these residual functions, we begin with the definitions of the ϕ-spanned function space and the marginal target density.

Definition 6 (ϕ-spanned function space)**.** *We denote by $\mathcal{F}_\phi$ the function space spanned by $\phi_1, ..., \phi_K$, i.e., $\mathcal{F}_\phi \coloneqq \left\{ (v^\top \phi) : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \;\middle|\; v \in \mathbb{R}^K \right\}$.*

Definition 7 (Marginal target density)**.** *Let $\nu(s, a)$ be the marginal target density, given by $\nu(s, a) \coloneqq (1 - \gamma) \sum_{h=0}^\infty \gamma^h (P^{\dagger h} p_0^\pi)(s, a)$, where $P^\dagger$ is the adjoint operator of P.*

Note that $(P^{\dagger h} p_0^\pi)(s, a)$ denotes the state-action density after h transitions starting from p_0. Thus, ν can be thought of as the relative frequency of the state-action visitations in the target episode with horizon-dependent multiplicative weights γ^h.

Then, two residual functions are defined as follows.

Definition 8 (χ-residual function)**.** *The χ-residual function is defined as*

$$\mathcal{R}_\chi(s, a) \coloneqq \frac{\nu(s, a)}{\mu(s, a)} - f^*(s, a),$$

where $f^* = \operatorname{argmin}_{f \in \mathcal{F}_\phi} \langle (\frac{\nu}{\mu} - f)^2 \rangle_\mu$.

Definition 9 (Bellman residual function)**.** *The Bellman residual function is defined as*

$$\mathcal{R}_B(s, a) \coloneqq \bar{r}(s, a) + \gamma (P Q^\sharp)(s, a) - f^*(s, a),$$

where $Q^\sharp(s, a) \coloneqq b^{\sharp\top} (I - \gamma F^\sharp)^{-1} \phi(s, a)$ is the action value function induced by $(b^\sharp, F^\sharp)$ and $f^* = \operatorname{argmin}_{f \in \mathcal{F}_\phi} \langle (\bar{r} + \gamma P Q^\sharp - f)^2 \rangle_\mu$.

Note that these functions are 'residual' since they are the remainders of the projection of some functions onto $\mathcal{F}_\phi$. $\mathcal{R}_\chi$ is the residual of the density ratio ν/μ and $\mathcal{R}_B$ is the residual of the Bellman-operated action value function $\bar{r} + \gamma P Q^\sharp$.

Now we are ready to state our first result.

Theorem 6. *Suppose Assumption 1, 2, 3 and 4 holds. Let $F_\gamma^\sharp := (I - \gamma F^\sharp)^{-1}$ and assume its existence. Then, we have the almost-sure convergence*

$$\hat{J}(\pi) - J(\pi) \overset{n\to\infty}{\longrightarrow} -\frac{1}{1-\gamma} \langle \mathcal{R}_B \mathcal{R}_\chi \rangle_\mu \tag{6}$$

in a $\mathrm{poly}(\frac{1}{1-\gamma}, \frac{1}{n}, G^*, \frac{1}{c_\mu})$ *rate uniformly with respect to the choice of* p_0*. Moreover, the convergence is also uniform with respect to the choice of* π*, if* $\sup_\pi \|F_\gamma^\sharp\|_2 < \infty$.

Proof. (Sketch.) It is directly derived from the non-asymptotic bound (Theorem 11, in the appendix), whose proof strategy is to decompose the error by $\hat{J}(\pi) - J(\pi) = (J^\sharp(\pi) - J(\pi)) + (\hat{J}(\pi) - J^\sharp(\pi))$, where $J^\sharp(\pi) := \langle Q^\sharp \rangle_{p_0^\pi}$ is the policy value induced by the projection $(F^\sharp, b^\sharp)$, and evaluate these terms separately. The limit (6) is obtained by evaluating the first term, $J^\sharp(\pi) - J(\pi)$, and the second term vanishes in a rate of $\mathcal{O}(1/\sqrt{n})$. In particular, the key step in the evaluation of the first term is the following series of identities,

$$J^\sharp(\pi) - J(\pi) = ... = -\langle \mathcal{R}_B \rangle_\nu = -\left\langle \mathcal{R}_B \frac{\nu}{\mu} \right\rangle_\mu = -\left\langle \mathcal{R}_B \left(\frac{\nu}{\mu} - f \right) \right\rangle_\mu, \quad \forall f \in \mathcal{F}_\phi,$$

where the last identity is what allows us to fit arbitrary function in $\mathcal{F}_\phi$ away from ν/μ, which is made possible with $\mathcal{R}_B$ being in the orthogonal complement of $\mathcal{F}_\phi$. The full proof is deferred to Section A. □

3.1.1 Interpretation

Theorem 6 characterizes when the linear estimators are consistent, $\langle \mathcal{R}_B \mathcal{R}_\chi \rangle_\mu = 0$. In particular, if the Bellman operator $f \mapsto \bar{r} + \gamma P f$ is closed under $\mathcal{F}_\phi$, we have $\mathcal{R}_B = 0$ and thus $\hat{J}(\pi)$ is consistent, which recovers the known results on realizable settings.

One important implication of Theorem 6 is another route to achieve the (near) consistency, namely $\mathcal{R}_\chi \approx 0$. It suggests we have small biases when the density ratio is well-approximated with $\mathcal{F}_\phi$, even if the Bellman-operator is not at all closed under $\mathcal{F}_\phi$.

Also note that $\mathcal{R}_\chi$ is easier to control via the complexity of $\mathcal{F}_\phi$. By definition, it is guaranteed to be monotonically non-increasing in the $L^2(\mu)$-norm as $\mathcal{F}_\phi$ grows, i.e., $\mathcal{F}_{\phi^{(1)}} \subset \mathcal{F}_{\phi^{(2)}} \subset ...$, whereas $\mathcal{R}_B$ is not. This is because the projection target ν/μ is independent of ϕ, while $\bar{r} + \gamma(PQ^\sharp)$ is dependent on ϕ through $Q^\sharp$.

3.1.2 On Boundedness of $F_\gamma^\sharp$

In a practical sense, the existence of $F_\gamma^\sharp$ is a necessary (not just sufficient) condition for the asymptotic convergence; if $I - \gamma F^\sharp$ has a zero singular value, the smallest singular value of $I - \gamma \hat{F}$ approaches to zero and hence the inverse is unstable. This is seen from the concentration of $\hat{F}$ (see Proposition 17 in the appendix) and the continuity of singular values.

The existence and the uniform boundedness of $F_\gamma^\sharp$ are properties associated with each problem instance $\mathcal{P}_{\mathrm{OPE}}$. The following proposition formalizes a sufficient condition for the uniform boundedness which depends only on μ and ϕ.

Proposition 7. *Under Assumption 4 and 5, if* $\phi(s,a)\mathbb{E}[\hat{\Sigma}]^{-1}\phi(s',a') \geq 0$ *for all* $s, s' \in \mathcal{S}$ *and* $a, a' \in \mathcal{A}$*, we have* $\sup_\pi \|F_\gamma^\sharp\|_2 < \infty$.

Though it has not been proved, we conjecture that this is the weakest condition for the π-uniform convergence that does not constrain π nor p_T. As a special case, any tabular feature ϕ satisfies the condition of Proposition 7 with any choice of μ.

For general ϕ other than the tabular ones, it is possible to compute a high-probability upper bound on $\|F_\gamma^\sharp\|_2$ (*cf.* Proposition 17 and 18 in the appendix) and combine it with the non-asymptotic bound to obtain a data-dependent concentration bound. If the resulting concentration rate is not acceptable, then one may change the feature mapping or fall back on the tabular one.

3.2 Consistency of Tile-Coding Estimators under Unrealizability

As is seen from Theorem 6, not surprisingly, a linear estimator with fixed ϕ is not consistent in general under unrealizability. This motivates us to investigate alternative methods that adaptively selects feature mappings. Such an estimation method is formally defined as follows.

Definition 10 (Nonparametric estimator)**.** *We refer to $(\boldsymbol{\phi}, \hat{m})$ as a nonparametric estimator if $\boldsymbol{\phi} = \{\phi^{(m)}\}_{m=1}^{\infty}$ is a sequence of feature mappings such that $\phi^{(m)} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{K_m}$, $K_m \geq 1$, and $\hat{m} : \mathbb{N} \to \mathbb{N}$ is a model-selecting function such that $\lim_{n\to\infty} \hat{m}(n) = \infty$. The output of the estimator is given as $\hat{J}(\pi; \boldsymbol{\phi}, \hat{m}) \coloneqq \hat{J}(\pi; \phi^{(\hat{m}(n))})$, where $\hat{J}(\pi; \phi)$ denotes the linear direct estimate given by Algorithm 1 with a feature mapping ϕ.*

In principle, if $\hat{m}(n)$ diverges slowly compared to the convergence (6), the error of $(\boldsymbol{\phi}, \hat{m})$ is still characterized by Theorem 6. Thus, if $\boldsymbol{\phi}$ is such that the asymptotic bias given by (6) goes to zero as $m \to \infty$, there exists a consistent nonparametric estimator $(\boldsymbol{\phi}, \hat{m})$ with sufficiently slowly diverging function $\hat{m}(n)$.

Below, we show a typical instance of such consistent nonparametric estimators, namely the refining tile-coding estimator. Henceforth, we assume for simplicity $\mathcal{S} \times \mathcal{A} = [0, 1]^d$, $d \geq 1$, i.e., both states and actions are continuous. [2]

Definition 11 (Refining tile-coding sequence)**.** *We call $\boldsymbol{\phi}$ as the refining tile-coding sequence if, for all $m \geq 1$, $\phi^{(m)}$ is tabular with respect to the m^d-partition $\{\mathcal{P}_{\mathbf{k}}^{(m)}\}_{\mathbf{k}\in[m]^d}$ such that $\mathcal{P}_{\mathbf{k}}^{(m)} = \prod_{j=1}^{d} \mathcal{I}_{k_j}^{(m)}$ for all $\mathbf{k} = (k_1, ..., k_d) \in [m]^d$, where $\mathcal{I}_k^{(m)} \coloneqq [\frac{k-1}{m}, \frac{k}{m})$ for all $1 \leq k < m$ and $\mathcal{I}_m^{(m)} \coloneqq [1 - \frac{1}{m}, 1]$.*

Leveraging Theorem 11, the non-asymptotic version of Theorem 6, it is shown the nonparametric estimation using the tile-coding scheme is consistent under very mild assumptions.

Proposition 8. *For the refining tile-coding sequence $\boldsymbol{\phi}$, we have $\hat{J}(\pi; \boldsymbol{\phi}, \hat{m}) \overset{n\to\infty}{\longrightarrow} J(\pi)$ a.s. if*

1. *Ξ^n is G^*-mixing for some $G^* < \infty$ (Assumption 1).*
2. *$0 < c_\mu \leq C_\mu \coloneqq \sup_{s\in\mathcal{S}, a\in\mathcal{A}} \mu(s, a) < \infty$ (*cf. *Assumption 2).*
3. *$\hat{m}(n)^d/\sqrt{n} \to 0$ as $n \to \infty$.*

Note that Proposition 8 assumes nothing on p_r, p_T, p_0 and π other than the implicit well-definedness of their density functions. Because of this, the rate of convergence may be arbitrarily slow. We have a stronger guarantee if some regularities of the density ratio ν/μ is given. A typical example of such regularity is the Lipschitz continuity.

Proposition 9. *Under the assumptions of Proposition 8, if ν/μ is Lipschitz continuous on $\mathcal{S} \times \mathcal{A}$, then we have $|\hat{J}(\pi; \boldsymbol{\phi}, \hat{m}) - J(\pi)| = \mathcal{O}(n^{-\frac{1}{2d+2}})$ with $\hat{m}(n) = \Theta(n^{\frac{1}{2d+1}})$.*

Note that the stronger convergence result is obtained still without any explicit conditions on the reward and the transition dynamics. This matches the implication of Theorem 6; the bias can be controlled with the norm of $\mathcal{R}_\chi$, which measures the regularity of ν/μ, without forcing explicit regularities on the Bellman operator.

4 Related Work

The offline policy evaluation is closely related to the off-policy policy evaluation in the bandit and RL literatures (Precup, 2000; Dudík et al., 2011; Sutton and Barto, 2018), where behavior policies are often assumed to be known. Recently, a number of researchers are focusing on more 'offline' settings (Levine et al., 2020) featuring unknown data-collecting policies and relatively large distribution shifts.

The theory of OPE has been often studied under a number of different types of realizability conditions. A common type of realizability is the Bellman-operator realizability Yin and Wang (2020); Duan et al. (2020). Jin et al. (2020) also considered the same, but approximate realizability and presented

[2] It is possible to extend it to allow states and actions being discrete partially or entirely.

an upper bound with respect to the violation of the realizability. Xie and Jiang (2020) studied offline RL under the Q-function realizability, a relaxation of the Bellman-operator realizability, also allowing small realizability-violation in their analysis. In comparison, we focus on exactly evaluating (dominating term of) the OPE error to understand more about the phenomenon under unrealizability, not just bounding it from above, which allows us straightforward analyses on the nonparametric scenario.

In the context of the Q-function realizability, Wang et al. (2021) showed that there is a hard instance with a specific construction of ϕ. Though there are minor differences in our problem setting (e.g., finite vs infinite horizons), we eliminate this kind of hard instances by bounding $\|F_\gamma^\sharp\|_2$, e.g., we chose ϕ to be tabular.

Our analysis bears new connection between the fixed-point iteration estimators (Le et al., 2019) and the marginal importance sampling (MIS) estimators Xie et al. (2019); Yin and Wang (2020); Nachum et al. (2019); Dai et al. (2020). Some of these results on the MIS estimator are derived without assuming realizability, but the effect of the function approximation error is not investigated analytically enough to be compared with our results.

Theories of the kernel-based RL naturally takes into account the function approximation error. The difference between kernel-based methods and linear methods is subtle; Although kernel methods tends to be more expensive in computation, they can be approximated with a finite linear basis (Rahimi et al., 2007). Ormoneit and Sen (2002) shows a kernel-smoothing approach yields a nonparametric convergence rate of $\mathcal{O}(n^{-\frac{1}{2d+4}})$ under different regularity assumptions, including the continuity of $\bar{r}$. (Feng et al., 2020) studied a kernel-based OPE with confidence error bound, but the error bound is algorithmically determined and not studied analytically.

5 Concluding Remarks

We have derived an asymptotically exact characterization of the error of the linear direct estimators, one of the most simple and basic OPE methods, under a completely unrealizable setting. As a consequence, we have found that the error converges to an inner product of two residual functions, each of which measures the unrealizability of the Bellman operator and the marginal density ratio, respectively. We have further investigated as its application the error of nonparametric estimators. To the best of our knowledge, the present study is the first to report the density-ratio-estimation interpretation of linear direct methods in unrealizable settings.

Limitation. One limitation of the present work is that the speed of the concentration (6) relies on the boundedness of $\|F_\gamma^\sharp\|$, although it seems inevitable if one employs Algorithm 1 (Section 3.1.2) In particular, this makes it difficult to construct consistent nonparametric estimators with general feature sequences $\boldsymbol{\phi}$ other than the tabular ones since the norm $\|F_\gamma^\sharp\|$ changes along with ϕ and ϕ changes along with n in the nonparametric setting. Moreover, as discussed in Section 4, controlling $\|F_\gamma^\sharp\|$ is at least equivalent of classifying 'hard instances' away from easy instances in the sense of Wang et al. (2021). Thus, we speculate it could be a key question for better understanding of the hardness of OPE problems, rather than just a technical difficulty.

Another limitation is that all the results in this paper are derived under the implicit assumption that the marginal target density ν is well defined with respect to the base measure of $\mathcal{S} \times \mathcal{A}$. In particular, generalized functions like Dirac's delta density function cannot be handled straightforwardly in our framework. This may be problematic in practice if, for example, the trajectory of the target episode quickly concentrates at a nontrivial point of the continuous space $\mathcal{S} \times \mathcal{A}$.

Feature Work. Studies on the feature mappings ϕ with higher-order smoothness such as Gaussian radial basis functions, spline functions and neural tangent kernels (Jacot et al., 2018) is promising for investigating the faster convergence capability of the linear direct estimators. Another promising direction is to extend the current analysis towards adaptive state-abstraction methods, e.g., Whiteson (2007).

References

Bradley, R. C. (2005). Basic properties of strong mixing conditions. a survey and some open questions. *arXiv preprint math/0511078*.

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

Duan, Y., Jia, Z., and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR.

Dudík, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Feng, Y., Ren, T., Tang, Z., and Liu, Q. (2020). Accountable off-policy evaluation with kernel bellman statistics. In *International Conference on Machine Learning*, pages 3102–3111. PMLR.

Freedman, D. A. (1975). On tail probabilities for martingales. *the Annals of Probability*, pages 100–118.

Fu, J., Norouzi, M., Nachum, O., Tucker, G., Wang, Z., Novikov, A., Yang, M., Zhang, M. R., Chen, Y., Kumar, A., et al. (2021). Benchmarks for deep off-policy evaluation. *arXiv preprint arXiv:2103.16596*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *arXiv preprint arXiv:1806.07572*.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR.

Lagoudakis, M. G. and Parr, R. (2003). Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149.

Le, H., Voloshin, C., and Yue, Y. (2019). Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Levine, S., Kumar, A., Tucker, G., and Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

Nachum, O., Chow, Y., Dai, B., and Li, L. (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*.

Ormoneit, D. and Sen, Ś. (2002). Kernel-based reinforcement learning. *Machine learning*, 49(2):161–178.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Rahimi, A., Recht, B., et al. (2007). Random features for large-scale kernel machines. In *NIPS*, volume 3, page 5. Citeseer.

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

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434.

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

Wang, R., Foster, D., and Kakade, S. M. (2021). What are the statistical limits of offline RL with linear function approximation? In *International Conference on Learning Representations*.

Whiteson, S. (2007). Adaptive tile coding for value function approximation. Technical report.

Xie, T. and Jiang, N. (2020). Batch value-function approximation with only realizability. *arXiv preprint arXiv:2008.04990*.

Xie, T., Ma, Y., and Wang, Y.-X. (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *arXiv preprint arXiv:1906.03393*.

Yin, M. and Wang, Y.-X. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3948–3958. PMLR.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]