Minimax Weight and Q-Function Learning for Off-Policy Evaluation

Masatoshi Uehara[1] **Jiawei Huang**[2] **Nan Jiang**[2]

Abstract

We provide theoretical investigations into off-policy evaluation in reinforcement learning using function approximators for (marginalized) importance weights and value functions. Our contributions include:
(1) A new estimator, MWL, that directly estimates importance ratios over the state-action distributions, removing the reliance on knowledge of the behavior policy as in prior work (Liu et al., 2018).
(2) Another new estimator, MQL, obtained by swapping the roles of importance weights and value-functions in MWL. MQL has an intuitive interpretation of minimizing *average Bellman errors* and can be combined with MWL in a doubly robust manner.
(3) Several additional results that offer further insights, including the sample complexities of MWL and MQL, their asymptotic optimality in the tabular setting, how the learned importance weights depend the choice of the discriminator class, and how our methods provide a unified view of some old and new algorithms in RL.

1. Introduction

In reinforcement learning (RL), off-policy evaluation (OPE) refers to the problem of estimating the performance of a new policy using historical data collected from a different policy, which is of crucial importance to the real-world applications of RL. The problem is genuinely hard as any unbiased estimator has to suffer a variance exponential in horizon in the worst case (Li et al., 2015; Jiang and Li, 2016), known as the *curse of horizon*.

Recently, a new family of estimators based on marginalized importance sampling (MIS) receive significant attention from the community (Liu et al., 2018; Xie et al., 2019), as they overcome the curse of horizon with relatively mild representation assumptions. The basic idea is to learn the marginalized importance weight that converts the state distribution in the data to that induced by the target policy, which sometimes has much smaller variance than the importance weight on action sequences used by standard sequential IS. Among these works, Liu et al. (2018) learn the importance weights by solving a minimax optimization problem defined with the help of a discriminator value-function class.

In this work, we investigate more deeply the space of algorithms that utilize a value-function class and an importance weight class for OPE. Our main contributions are:

- (Section 4) A new estimator, MWL, that directly estimates importance ratios over the state-action distributions, removing the reliance on knowledge of the behavior policy as in prior work (Liu et al., 2018).
- (Section 5) By swapping the roles of importance weights and Q-functions in MWL, we obtain a new estimator that learns a Q-function using importance weights as discriminators. The procedure and the guarantees of MQL exhibit an interesting symmetry w.r.t. MWL. We also combine MWL and MQL in a doubly robust manner and provide their sample complexity guarantees (Section 6).
- (Section 7) We examine the statistical efficiency of MWL and MQL, and show that by modeling state-action functions, MWL and MQL are able to achieve the semiparametric lower bound of OPE in the tabular setting while their state-function variants fail to do so.
- Our work provides a unified view of many old and new algorithms in RL. For example, when both importance weights and value functions are modeled using the same linear class, we recover LSTDQ (Lagoudakis and Parr, 2004) and off-policy LSTD (Bertsekas and Yu, 2009; Dann et al., 2014) as special cases of MWL/MQL and their state-function variants. This gives LSTD algorithms a novel interpretation that is very different from the standard TD intuition. As another example, (tabular) model-based OPE and step-wise importance sampling—two algorithms that are so different that we seldom connect them to each other—are both special cases of MWL.

[1] Harvard University, Massachusetts , Boston, USA [2] University of Illinois at Urbana-Champaign, Champaign, Illinois, USA. Correspondence to: Masatoshi Uehara <uehara-masatoshi136@gmail.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

2. Preliminaries

An infinite-horizon discounted MDP is often specified by a tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \Delta([0, R_{\max}])$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. We also use $\mathcal{X} := \mathcal{S} \times \mathcal{A}$ to denote the space of state-action pairs. Given an MDP, a (stochastic) policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ and a starting state distribution $d_0 \in \Delta(\mathcal{S})$ together determine a distribution over trajectories of the form $s_0, a_0, r_0, s_1, a_1, r_1, \ldots$, where $s_0 \sim d_0$, $a_t \sim \pi(s_t)$, $r_t \sim \mathcal{R}(s_t, a_t)$, and $s_{t+1} \sim P(s_t, a_t)$ for $t \geq 0$. The ultimate measure of the a policy's performance is the (normalized) expected discounted return:

$$R_\pi := (1-\gamma)\mathrm{E}_{d_0,\pi}\left[\textstyle\sum_{t=0}^{\infty} \gamma^t r_t\right], \tag{1}$$

where the expectation is taken over the randomness of the trajectory determined by the initial distribution and the policy on the subscript, and $(1 - \gamma)$ is the normalization factor.

A concept central to this paper is the notion of (normalized) discounted occupancy:

$$d_{\pi,\gamma} := (1-\gamma)\textstyle\sum_{t=0}^{\infty} \gamma^t d_{\pi,t},$$

where $d_{\pi,t} \in \Delta(\mathcal{X})$ is the distribution of (s_t, a_t) under policy π. (The dependence on d_0 is made implicit.) We will sometimes also write $s \sim d_{\pi,\gamma}$ for sampling from its marginal distribution over states. An important property of discounted occupancy, which we will make heavy use of, is

$$R_\pi = \mathrm{E}_{(s,a)\sim d_{\pi,\gamma},\, r\sim\mathcal{R}(s,a)}[r]. \tag{2}$$

It will be useful to define the policy-specific Q-function:

$$Q^\pi(s,a) := \mathrm{E}[\textstyle\sum_{t=0}^{\infty} \gamma^t r_t | s_0 = s, a_0 = a;\, a_t \sim \pi(s_t)\, \forall t > 0].$$

The corresponding state-value function is $V^\pi(s) := Q^\pi(s, \pi)$, where for any function f, $f(s, \pi)$ is the shorthand for $\mathrm{E}_{a\sim\pi(s)}[f(s, a)]$.

Off-Policy Evaluation (OPE) We are concerned with estimating the expected discounted return of an *evaluation policy* π_e under a given initial distribution d_0, using data collected from a different *behavior policy* π_b. For our methods, we will consider the following data generation protocol, where we have a dataset consisting of n i.i.d. tuples (s, a, r, s') generated according to the distribution:

$$s \sim d_{\pi_b}, a \sim \pi_b(s), r \sim \mathcal{R}(s,a), s' \sim P(s,a).$$

Here d_{π_b} is some exploratory state distribution that well covers the state space,[1] and the technical assumptions required on this distribution will be discussed in later sections. With a slight abuse of notation we will also refer to the joint distribution over (s, a, r, s') or its marginal on (s, a) as d_{π_b}, e.g., whenever we write $(s, a, r, s') \sim d_{\pi_b}$ or $(s, a) \sim d_{\pi_b}$, the variables are always distributed according to the above generative process. We will use $\mathrm{E}[\cdot]$ to denote the exact expectation, and use $\mathrm{E}_n[\cdot]$ as its empirical approximation using the n data points.

On the i.i.d. assumption Although we assume i.i.d. data for concreteness and the ease of exposition, the actual requirement on the data is much milder: our method works as long as the empirical expectation (over n data points) concentrates around the exact expectation w.r.t. $(s, a, r, s') \sim d_{\pi_b}$ for *some* d_{π_b}.[2] This holds, for example, when the Markov chain induced by π_b is ergodic, and our data is a single long trajectory generated by π_b without resetting. As long as the induced chain mixes nicely, it is well known that the empirical expectation over the single trajectory will concentrate, and in this case $d_{\pi_b}(s)$ corresponds to the stationary distribution of the Markov chain.[3]

3. Overview of OPE Methods

Direct Methods A straightforward approach to OPE is to estimate an MDP model from data, and then compute the quantity of interest from the estimated model. An alternative but closely related approach is to fit Q^{π_e} directly from data using standard approximate dynamic programming (ADP) techniques, e.g., the policy evaluation analog of Fitted Q-Iteration (Ernst et al., 2005; Le et al., 2019). While these methods overcome the curse of dimensionality and are agnostic to the knowledge of π_b, they often require very strong representation assumptions to succeed: for example, in the case of fitting a Q-value function from data, not only one needs to assume realizability, that the Q-function class (approximately) captures Q^{π_e}, but the class also needs to be closed under Bellman update B^{π_e} (Antos et al., 2008), otherwise ADP can diverge in discounted problems (Tsitsiklis and Van Roy, 1997) or suffer exponential sample complexity in finite-horizon problems (Dann et al., 2018, Theorem 45); we refer the readers to Chen and Jiang (2019) for further discussions on this condition. When the function approximator fails to satisfy these strong assumptions, the estimator can potentially incur a high bias.

Importance Sampling (IS) IS forms an unbiased estimate of the expected return by collecting full-trajectory

[1] Unlike Liu et al. (2018), we do not need to assume that d_{π_b} is π_b's discounted occupancy; see Footnote 15 in Appendix A.5 for the reason, where we also simplify Liu et al. (2018)'s loss so that it does not rely on this assumption.

[2] We assume $a \sim \pi_b(s)$ throughout the paper since this is required by previous methods which we would like to compare to. However, most of our derivations do not require that the data is generated from a single behavior policy (which is a common characteristic of behavior-agnostic OPE methods).

[3] We consider precisely this setting in Appendix C.1 to solidify the claim that we do not really need i.i.d.ness.

	π_b known?	Target object	Func. approx.	Tabular optimality
MSWL (Liu et al., 2018)	Yes	$w^{\mathcal{S}}_{\pi_e/\pi_b}$ (Eq.(3))	$w^{\mathcal{S}}_{\pi_e/\pi_b} \in \mathcal{W}^{\mathcal{S}}$, $V^{\pi_e} \in \mathcal{F}^{\mathcal{S}}$ (*)	No
MWL (Sec 4)	No	w_{π_e/π_b} (Eq.(4))	$w_{\pi_e/\pi_b} \in \mathcal{W}$, $Q^{\pi_e} \in \mathrm{conv}(\mathcal{F})$	Yes
MQL (Sec 5)	No	Q^{π_e}	$Q^{\pi_e} \in \mathcal{Q}$, $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$	Yes
Fitted-Q	No	Q^{π_e}	$\mathcal{Q}$ closed under B^{π_e}	Yes

Table 1. Summary of some of the OPE Methods. For methods that require knowledge of π_b, the policy can be estimated from data to form a "plug-in" estimator (e.g., Hanna et al., 2019). In the function approximation column, we use blue color to mark the conditions for the discriminator classes for minimax-style methods. For Liu et al. (2018), we use $\mathcal{W}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{S}}$ for the function classes to emphasize that their functions are over the state space (ours are over the state-action space). Although they assumed $V^{\pi_e} \in \mathcal{F}^{\mathcal{S}}$ (*), this assumption can also be relaxed to $V^{\pi_e} \in \mathrm{conv}(\mathcal{F}^{\mathcal{S}})$ as in our analyses. Also note that the assumption for the main function classes ($\mathcal{W}^{\mathcal{S}}$, $\mathcal{W}$, and $\mathcal{Q}$) can be relaxed as discussed in Examples 1 and 3, and we put realizability conditions here only for simplicity.

behavioral data and reweighting each trajectory according to its likelihood under π_e over π_b (Precup et al., 2000). Such a ratio can be computed as the cumulative product of the importance weight over action ($\frac{\pi_e(a|s)}{\pi_b(a|s)}$) for each time step, which is the cause of high variance in IS: even if π_e and π_b only has constant divergence per step, the divergence will be amplified over the horizon, causing the cumulative importance weight to have exponential variance, thus the "curse of horizon". Although techniques that combine IS and direct methods can partially reduce the variance, the exponential variance of IS simply cannot be improved when the MDP has significant stochasticity (Jiang and Li, 2016).

Marginalized Importance Sampling (MIS) MIS improves over IS by observing that, if π_b and π_e induces marginal distributions over states that have substantial overlap—which is often the case in many practical scenarios—then reweighting the reward r in each data point (s, a, r, s') with the following ratio

$$w^{\mathcal{S}}_{\pi_e/\pi_b}(s) \cdot \frac{\pi_e(a|s)}{\pi_b(a|s)}, \text{ where } w^{\mathcal{S}}_{\pi_e/\pi_b}(s) := \frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)} \quad (3)$$

can potentially have much lower variance than reweighting the entire trajectory (Liu et al., 2018). The difference between IS and MIS is essentially performing importance sampling using Eq.(1) vs. Eq.(2). However, the weight $w^{\mathcal{S}}_{\pi_e/\pi_b}$ is not directly available and has to be estimated from data. Liu et al. (2018) proposes an estimation procedure that requires two function approximators, one for modeling the weighting function $w^{\mathcal{S}}_{\pi_e/\pi_b}(s)$, and the other for modeling V^{π_b} which is used as a discriminator class for distribution learning. Compared to the direct methods, MIS only requires standard realizability conditions for the two function classes, though it also needs the knowledge of π_b. A related method for finite horizon problems has been developed by Xie et al. (2019).

4. Minimax Weight Learning (MWL)

In this section we propose a simple extension to Liu et al. (2018) that is agnostic to the knowledge of π_b. The estimator in the prior work uses a discriminator class that contains V^{π_e} to learn the marginalized importance weight on state distributions (see Eq.(3)). We show that as long as the discriminator class is slightly more powerful—in particular, it is a Q-function class that realizes Q^{π_e}—then we are able to learn the importance weight over state-action pairs directly:

$$w_{\pi_e/\pi_b}(s, a) := \frac{d_{\pi_e,\gamma}(s, a)}{d_{\pi_b}(s, a)}. \quad (4)$$

We can use it to directly re-weight the rewards without having to know π_b, as $R_{\pi_e} = R_{\mathrm{w}}[w_{\pi_e/\pi_b}] := \mathrm{E}_{\pi_b}[w_{\pi_e/\pi_b}(s, a) \cdot r]$. It will be also useful to define $R_{\mathrm{w},n}[w] := \mathrm{E}_n[w(s, a) \cdot r]$ as the empirical approximation of $R_{\mathrm{w}}[\cdot]$ based on n data points.

Before giving the estimator and its theoretical properties, we start with two assumptions that we will use throughout the paper, most notably that the state-action distribution in data well covers the discounted occupancy induced by π_b.

Assumption 1. Assume $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ is a compact space. Let ν be its Lebesgue measure. [4]

Assumption 2. There exists $C_w < +\infty$ such that $w_{\pi_e/\pi_b}(s, a) \leq C_w \ \forall (s, a) \in \mathcal{X}$.

In the rest of this section, we derive the new estimator and provide its theoretical guarantee. Our derivation (Eqs.(5)–(7)) provides the high-level intuitions for the method while only invoking basic and familiar concepts in MDPs (essentially, just Bellman equations). The estimator of Liu et al. (2018) can be also derived in a similar manner.

Derivation Recall that it suffices to learn $w : \mathcal{X} \to \mathbb{R}$ such that $R_{\mathrm{w}}[w] = R_{\pi_e}$. This is equivalent to

$$\mathbb{E}_{\pi_b}[w(s, a) \cdot r] = (1 - \gamma)\mathbb{E}_{s\sim d_0}[Q^{\pi_e}(s, \pi_e)]. \quad (5)$$

[4] When ν is the counting measure for finite $\mathcal{X}$, all the results hold with minor modifications.

By Bellman equation, we have $\mathbb{E}[r|s,a] = \mathbb{E}[Q^{\pi_e}(s,a) - \gamma Q^{\pi_e}(s',\pi_e)|s,a]$. We use the RHS to replace r in Eq.(5),

$$\mathbb{E}_{\pi_b}[w(s,a)\cdot(Q^{\pi_e}(s,a)-\gamma Q^{\pi_e}(s',\pi_e))] \\ = (1-\gamma)\mathbb{E}_{s\sim d_0}[Q^{\pi_e}(s,\pi_e)]. \tag{6}$$

To recap, it suffices to find any w that satisfies the above equation. Since we do not know Q^{π_e}, we will use a function class $\mathcal{F}$ that (hopefully) captures Q^{π_e}, and find w that minimizes (the absolute value of) the following objective function that measures the violation of Eq.(6) over all $f \in \mathcal{F}$:

$$L_{\mathrm{w}}(w,f) := \mathrm{E}_{(s,a,r,s')\sim d_{\pi_b}}[\{\gamma w(s,a)\cdot f(s',\pi_e) \tag{7}$$

$$-w(s,a)f(s,a)\}] + (1-\gamma)\mathrm{E}_{s\sim d_0}[f(s,\pi_e)]. \tag{8}$$

The loss is always zero when $w = w_{\pi_e/\pi_b}$, so adding functions to $\mathcal{F}$ does not hurt its validity. Such a solution is also unique if we require $L_{\mathrm{w}}(w,f) = 0$ for a rich set of functions and d_{π_b} is supported on the entire $\mathcal{X}$, formalized as the following lemma:[5]

Lemma 1. $L_{\mathrm{w}}(w_{\pi_e/\pi_b}, f) = 0\ \forall f \in L^2(\mathcal{X},\nu) := \{f : \int f(s,a)^2 \mathrm{d}\nu < \infty\}$. *Moreover, under additional technical assumptions,*[6] w_{π_e/π_b} *is the only function that satisfies this.*

This motivates the following estimator, which uses two function classes: a class $\mathcal{W} : \mathcal{X} \to \mathbb{R}$ to model the w_{π_e/π_b} function, and another class $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ to serve as the discriminators:

$$\hat{w}(s,a) = \arg\min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)^2. \tag{9}$$

Note that this is the ideal estimator that assumes exact expectations (or equivalently, infinite amount of data). In reality, we will only have access to a finite sample, and the real estimator replaces $L_{\mathrm{w}}(w,f)$ with its sample-based estimation, defined as

$$L_{\mathrm{w},n}(w,f) := \mathrm{E}_n[\{\gamma w(s,a)f(s',\pi_e) - w(s,a)f(s,a)\}] \\ + (1-\gamma)\mathrm{E}_{d_0}[f(s,\pi_e)]. \tag{10}$$

So the sample-based estimator is $\hat{w}_n(s,a) := \arg\min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} L_{\mathrm{w},n}(w,f)^2$. We call this estimation procedure MWL (minimax weight learning), and provide its main theorem below.

Theorem 2. *For any given* $w : \mathcal{X} \to \mathbb{R}$, *define* $R_{\mathrm{w}}[w] = \mathrm{E}_{d_{\pi_b}}[w(s,a)\cdot r]$. *If* $Q^{\pi_e} \in \mathrm{conv}(\mathcal{F})$, *where* $\mathrm{conv}(\cdot)$ *denotes the convex hull of a function class,*

$$|R_{\pi_e} - R_{\mathrm{w}}[w]| \le \max_{f\in\mathcal{F}} |L_{\mathrm{w}}(w,f)|,$$

$$|R_{\pi_e} - R_{\mathrm{w}}[\hat{w}]| \le \min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} |L_{\mathrm{w}}(w,f)|.$$

A few comments are in order:

1. To guarantee that the estimation is accurate, all we need is $Q^{\pi_e} \in \mathrm{conv}(\mathcal{F})$, and $\min_w \max_f |L_{\mathrm{w}}(w,f)|$ is small. While the latter can be guaranteed by realizability of $\mathcal{W}$, i.e., $w_{\pi_e/\pi_b} \in \mathcal{W}$, we show in an example below that realizability is sufficient but not always necessary: in the extreme case where $\mathcal{F}$ only contains Q^{π_e}, even a constant w function can satisfy $\max_f |L_{\mathrm{w}}(w,f)| = 0$ and hence provide accurate OPE estimation.

   Example 1 (Realizability of $\mathcal{W}$ can be relaxed)**.** When $\mathcal{F} = \{Q^{\pi_e}\}$, as long as $w_0 \in \mathcal{W}$ where w_0 is a constant function that always evaluates to R_{π_e}/R_{π_b}, we have $R_{\mathrm{w}}[\hat{w}] = R_{\pi_e}$. See Appendix A.1 for a detailed proof.

2. For the condition that $Q^{\pi_e} \in \mathrm{conv}(\mathcal{F})$, we can further relax the convex hull to the linear span, though we will need to pay the ℓ_1 norm of the combination coefficients in the later sample complexity analysis. It is also straightforward to incorporate approximation errors (see Remark 2 in Appendix A) and we do not further consider these relaxations for simplicity.

3. Although Eq.(9) uses $L_{\mathrm{w}}(w,f)^2$ in the objective function, the square is mostly for optimization convenience and is not vital in determining the statistical properties of the estimator. In later sample complexity analysis, it will be much more convenient to work with the equivalent objective function that uses $|L_{\mathrm{w}}(w,f)|$ instead.

4. When the behavior policy π_b is known, we can incorporate this knowledge by setting $\mathcal{W} = \{s \mapsto w(s)\frac{\pi_e(a|s)}{\pi_b(a|s)} : w \in \mathcal{W}^{\mathcal{S}}\}$, where $\mathcal{W}^{\mathcal{S}}$ is some function class over the state space. The resulting estimator is still different from (Liu et al., 2018) since our discriminator class is still over the state-action space.

4.1. Case Studies

The estimator in Eq.(10) requires solving a minimax optimization problem, which can be computationally challenging. Following Liu et al. (2018) we show that the inner maximization has a closed form solution when we choose $\mathcal{F}$ to correspond to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$ be a RKHS associated with kernel $K(\cdot,\cdot)$. We include an informal statement below and defer the detailed expression to Appendix A.2 due to space limit.

Lemma 3 (Informal)**.** *When* $\mathcal{F} = \{f \in (\mathcal{X} \to \mathbb{R}) : \langle f, f\rangle_{\mathcal{H}_K} \le 1\}$, *the term* $\max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)^2$ *has a closed form expression.*

As a further special case when both $\mathcal{W}$ and $\mathcal{F}$ are linear classes under the same state-action features $\phi : \mathcal{X} \to \mathbb{R}^d$. The resulting algorithm has a close connection to LSTDQ

[5] All proofs of this paper can be found in the appendices.

[6] As we will see, the identifiability of w_{π_e/π_b} is not crucial to the OPE goal, so we defer the technical assumptions to a formal version of the lemma in Appendix A, where we also show that the same statement holds when $\mathcal{F}$ is an ISPD kernel; see Theorem 15.

(Lagoudakis and Parr, 2004), which we will discuss more explicitly later in Section 5, Example 7.

Example 2. Let $w(s,a;\alpha) = \phi(s,a)^\top \alpha$ where $\phi(s,a) \in \mathbb{R}^d$ is some basis function and α is the parameters. If we use the same linear function space as $\mathcal{F}$, i.e., $\mathcal{F} = \{(s,a) \mapsto \phi(s,a)^\top \beta : \beta \in \mathbb{R}^d\}$, then the estimation of α given by MWL is (assuming the matrix being inverted is full-rank)

$$\hat{\alpha} = \mathrm{E}_n[-\gamma\phi(s',\pi_e)\phi(s,a)^\top + \phi(s,a)\phi(s,a)^\top]^{-1} \\ (1-\gamma)\mathrm{E}_{s\sim d_0}[\phi(s,\pi_e)]. \tag{11}$$

The sample-based estimator for the OPE problem is therefore $R_{\mathrm{w},n}[\hat{w}_n] = \mathrm{E}_n[r\phi(s,a)^\top]\hat{\alpha}$; see Appendix A.3 for a full derivation.

Just as our method corresponds to LSTDQ in the linear setting, it is worth pointing out that the method of Liu et al. (2018)—which we will call MSWL (minimax state weight learning) for distinction and easy reference—corresponds to off-policy LSTD (Bertsekas and Yu, 2009; Dann et al., 2014); see Appendix A.5 for details.

4.2. Connections to related work

Nachum et al. (2019a) has recently proposed a version of MIS with a similar goal of being agnostic to the knowledge of π_b. In fact, their estimator and ours have an interesting connection, as our Lemma 1 can be obtained by taking the functional derivative of their loss function; we refer interested readers to Appendix A.4 for details. That said, there are also important differences between our methods. First, our loss function can be reduced to single-stage optimization when using an RKHS discriminator, just as in Liu et al. (2018). In comparison, the estimator of Nachum et al. (2019a) cannot avoid two-stage optimization. Second, they do not directly estimate $w_{\pi_e/\pi_b}(s,a)$, and instead estimate $\nu^*(s,a)$ such that $\nu^*(s,a) - \gamma\mathrm{E}_{s'\sim P(s,a)\, a'\sim\pi_e(s')}[\nu^*(s',a')] = w_{\pi_e/\pi_b}(s,a)$, which is more indirect.

In the special case of $\gamma = 0$, i.e., when the problem is a contextual bandit, our method essentially becomes kernel mean matching when using an RKHS discriminator (Gretton et al., 2012), so MWL can be viewed as a natural extension of kernel mean matching in MDPs.

5. Minimax Q-Function Learning (MQL)

In Section 4, we show how to use value-function class as discriminators to learn the importance weight function. In this section, by swapping the roles of w and f, we derive a new estimator that learns Q^{π_e} from data using importance weights as discriminators. The resulting objective function has an intuitive interpretation of *average Bellman errors*, which has many nice properties and interesting connections to prior works in other areas of RL.

Setup We assume that we have a class of state-action importance weighting functions $\mathcal{G} \subset (\mathcal{X} \to \mathbb{R})$ and a class of state-action value functions $\mathcal{Q} \subset (\mathcal{X} \to \mathbb{R})$. To avoid confusion we do not reuse the symbols $\mathcal{W}$ and $\mathcal{F}$ in Section 4, but when we apply both estimators on the same dataset (and possibly combine them via doubly robust), it can be reasonable to choose $\mathcal{Q} = \mathcal{F}$ and $\mathcal{G} = \mathcal{W}$. For now it will be instructive to assume that Q^{π_e} is captured by $\mathcal{Q}$ (we will relax this assumption later), and the goal is to find $q \in \mathcal{Q}$ such that

$$R_{\mathrm{q}}[q] := (1-\gamma)\mathrm{E}_{s\sim d_0}[q(s,\pi_e)] \tag{12}$$

(i.e., the estimation of R_{π_e} as if q were Q^{π_e}) is an accurate estimate of R_{π_e}.[7]

Loss Function The loss function of MQL is

$$L_{\mathrm{q}}(q,g) = \mathrm{E}_{d_{\pi_b}}[g(s,a)(r + \gamma q(s',\pi_e) - q(s,a))].$$

As we alluded to earlier, if g is the importance weight that converts the data distribution (over (s,a)) d_{π_b} to some other distribution μ, then the loss becomes $\mathrm{E}_\mu[r + \gamma q(s',\pi_e) - q(s,a)]$, which is essentially the average Bellman error defined by Jiang et al. (2017). An important property of this quantity is that, if $\mu = d_{\pi_e,\gamma}$, then by (a variant of) Lemma 1 of Jiang et al. (2017), we immediately have $R_{\pi_e} - R_{\mathrm{q}}[q] =$

$$\mathrm{E}_{d_{\pi_e,\gamma}}[r + \gamma q(s',\pi_e) - q(s,a)] (= L_{\mathrm{q}}(q, w_{\pi_e/\pi_b})).$$

Similar to the situation of MWL, we can use a rich function class $\mathcal{G}$ to model w_{π_e/π_b}, and find q that minimizes the RHS of the above equation for all $g \in \mathcal{G}$, which gives rise to the following estimator:

$$\hat{q} = \arg\min_{q\in\mathcal{Q}} \max_{g\in\mathcal{G}} L_{\mathrm{q}}(q,g)^2.$$

We call this method MQL (minimax Q-function learning). Similar to Section 4, we use $\hat{q}_n$ to denote the estimator based on a finite sample of size n (which replaces $L_{\mathrm{q}}(q,g)$ with its empirical approximation $L_{\mathrm{q},n}(q,g)$), and develop the formal results that parallel those in Section 4 for MWL. All proofs and additional results can be found in Appendix B.

Lemma 4. $L_{\mathrm{q}}(Q^{\pi_e}, g) = 0$ *for* $\forall g \in L^2(\mathcal{X},\nu)$. *Moreover, if we further assume that* $d_{\pi_b}(s,a) > 0\ \forall(s,a)$, *then* Q^{π_e} *is the only function that satisfies such a property.*

Similar to the case of MWL, we show that under certain representation conditions, the estimator will provide accurate estimation to R_{π_e}.

[7] Note that $R_{\mathrm{q}}[\cdot]$ only requires knowledge of d_0 and can be computed directly. This is different from the situation in MWL, where $R_{\mathrm{w}}[\cdot]$ still requires knowledge of d_{π_b} even if the importance weights are known, and the actual estimator needs to use the empirical approximation $R_{\mathrm{w},n}[\cdot]$.

Theorem 5. *The following holds if $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$:*

$$|R_{\pi_e} - R_{\mathrm{q}}[q]| \le \max_{g\in\mathcal{G}} |L_{\mathrm{q}}(q,g)|,$$
$$|R_{\pi_e} - R_{\mathrm{q}}[\hat{q}]| \le \min_{q\in\mathcal{Q}} \max_{g\in\mathcal{G}} |L_{\mathrm{q}}(q,g)|.$$

5.1. Case Studies

We proceed to give several special cases of this estimator corresponding to different choices of $\mathcal{G}$ to illustrate its properties. In the first example, we show the analogy of Example 1 for MWL, which demonstrates that requiring $\min_q \max_g L_{\mathrm{q}}(q,g) = 0$ is weaker than realizability $Q^{\pi_e} \in \mathcal{Q}$:

Example 3 (Realizability of $\mathcal{Q}$ can be relaxed)**.** When $\mathcal{G} = \{w_{\pi_e/\pi_b}\}$, as long as $q_0 \in \mathcal{Q}$, where q_0 is a constant function that always evaluates to $R_{\pi_e}/(1-\gamma)$, we have $R_{\mathrm{q}}[\hat{q}] = R_{\pi_e}$. See Appendix B.1 for details.

Next, we show a simple and intuitive example where $w_{\pi_e/\pi_b} \notin \mathcal{G}$ but $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$, i.e., there are cases where relaxing $\mathcal{G}$ to its convex hull yields stronger representation power and the corresponding theoretical results provide a better description of the algorithm's behavior.

Example 4. Suppose $\mathcal{X}$ is finite. Let $\mathcal{Q}$ be the tabular function class, and $\mathcal{G}$ is the set of state-action indicator functions.[8] Then $w_{\pi_e/\pi_b} \notin \mathcal{G}$ but $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$, and $R_{\mathrm{q}}[\hat{q}] = 0$. Furthermore, the sample-based estimator $\hat{q}_n$ coincides with the model-based solution, as $L_{\mathrm{q},n}(q,g) = 0$ for each g is essentially the Bellman equation on the corresponding state-action pair in the estimated MDP model. (In fact, the solution remains the same if we replace $\mathcal{G}$ with the tabular function class.)

In the next example, we choose $\mathcal{G}$ to be a rich L^2-class with bounded norm, and recover the usual (squared) Bellman error as a special case. A similar example has been given by Feng et al. (2019).

Example 5 (L^2-class)**.** When $\mathcal{G} = \{g : \mathrm{E}_{d_{\pi_b}}[g^2] \le 1\}$,

$$\max_{g\in\mathcal{G}} L_{\mathrm{q}}(q,g)^2 = \mathrm{E}_{d_{\pi_b}}[((B^{\pi_e}q)(s,a) - q(s,a))^2],$$

where B^{π} is the Bellman update operator $(B^{\pi}q)(s,a) := \mathrm{E}_{r\sim\mathcal{R}(s,a),s'\sim P(s,a)}[r + \gamma q(s',\pi)]$.

Note that the standard Bellman error cannot be directly estimated from data when the state space is large, even if the $\mathcal{Q}$ class is realizable (Szepesvari and Munos, 2005; Sutton and Barto, 2018; Chen and Jiang, 2019). From our perspective, this difficulty can be explained by the fact that squared Bellman error corresponds to an overly rich discriminator class that demands an unaffordable sample complexity.

The next example is RKHS class which yields a closed-form solution to the inner maximization as usual.

Example 6 (RKHS class)**.** When $\mathcal{G} = \{g(s,a); \langle g,g\rangle_{\mathcal{H}_K} \le 1\}$, we have the following:

Lemma 6. *Let* $\mathcal{G} = \{g(s,a); \langle g,g\rangle_{\mathcal{H}_K} \le 1\}$. *Then,* $\max_{g\in\mathcal{G}} L_{\mathrm{q}}(q,g)^2 = \mathrm{E}_{d_{\pi_b}}[\Delta^q(q;s,a,r,s')\Delta^q(q;\tilde{s},\tilde{a},\tilde{r},\tilde{s}')K((s,a),(\tilde{s},\tilde{a}))]$, *where* $\Delta^q(q;s,a,r,s') = g(s,a)(r + \gamma q(s',\pi_e) - q(s,a))$.

Finally, the linear case.

Example 7. Let $q(s,a;\alpha) = \phi(s,a)^\top\alpha$ where $\phi(s,a) \in \mathbb{R}^d$ is some basis function and α is the parameters. If we use the same linear function space as $\mathcal{G}$, i.e., $\mathcal{G} = \{(s,a) \mapsto \phi(s,a)^\top\beta : \beta \in \mathbb{R}^d\}$, then MQL yields $\hat{\alpha}$:

$$\mathrm{E}_n[-\gamma\phi(s,a)\phi(s',\pi_e)^\top + \phi(s,a)\phi(s,a)^\top]^{-1}\mathrm{E}_n[r\phi(s,a)].$$

The derivation is similar to that of Example 2 (Appendix A.3) and omitted.)The resulting $q(s,a;\hat{\alpha})$ as an estimation of Q^{π_e} is precisely LSTDQ (Lagoudakis and Parr, 2004). In addition, the final OPE estimator $R_{\mathrm{q}}[\hat{q}_n] = (1-\gamma)\mathrm{E}_{s\sim d_0}[q(s,\pi_e;\hat{\alpha})]$ is the same as $R_{\mathrm{w},n}[\hat{w}_n]$ when $\mathcal{W}$ and $\mathcal{F}$ are the same linear class (Example 2).

5.2. Connection to Kernel Loss (Feng et al., 2019)

Feng et al. (2019) has recently proposed a method for value-based RL. By some transformations, we may rewrite their loss over state-value function $v : \mathcal{S} \to \mathbb{R}$ as

$$\max_{g\in\mathcal{G}^{\mathcal{S}}} \left(\mathrm{E}_{\pi_e}[\{r + \gamma v(s') - v(s)\}g(s)]\right)^2, \qquad (13)$$

where $\mathcal{G}^{\mathcal{S}}$ is an RKHS over the state space. While their method is very similar to MQL when written as the above expression, they focus on learning a state-value function and need to be on-policy for policy evaluation. In contrast, our goal is OPE (i.e., estimating the expected return instead of the value function), and we learn a Q-function as an intermediate object and hence are able to learn from off-policy data. More importantly, the importance weight interpretation of g has eluded their paper and they interpret this loss purely from a kernel perspective. In contrast, by leveraging the importance weight interpretation, we are able to establish approximation error bounds based on representation assumptions that are fully expressed in quantities directly defined in the MDP. We also note that their loss for policy optimization can be similarly interpreted as minimizing average Bellman errors under a set of distributions.

Furthermore, it is easy to extend their estimator to the OPE task using knowledge of π_b, which we call MVL; see Appendix B.2 for details. Again, just as we discussed in Appendix A.5 on MSWL, when we use linear classes for both value functions and importance weights, these two estimators become two variants of off-policy LSTD (Dann et al.,

[8] Strictly speaking we need to multiply these indicator functions by C_w to guarantee $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$; see the comment on linear span after Theorem 2.

2014; Bertsekas and Yu, 2009) and coincide with MSWL and its variant.

6. Doubly Robust Extension and Sample Complexity of MWL & MQL

In the previous sections we have seen two different ways of using a value-function class and an importance-weight class for OPE. Which one should we choose?

In this section we show that there is no need to make a choice. In fact, we can combine the two estimates naturally through the doubly robust trick (Kallus and Uehara, 2019b) (see also (Tang et al., 2020)), whose population version is:

$$\begin{aligned} R[w,q] &= (1-\gamma)\mathrm{E}_{d_0}[q(s,\pi_e)] \\ &\quad + \mathrm{E}_{d_{\pi_b}}[w(s,a)\{r+\gamma q(s',\pi_e)-q(s,a)\}]. \end{aligned} \tag{14}$$

As before, we write $R_n[w,q]$ as the empirical analogue of $R[w,q]$. While w and q are supposed to be the MWL and MQL estimators in practice, in this section we will sometimes treat w and q as arbitrary functions from the $\mathcal{W}$ and $\mathcal{Q}$ classes to keep our results general. By combining the two estimators, we obtain the usual doubly robust property, that when either $w = w_{\pi_e/\pi_b}$ or $q = Q^{\pi_e}$, we have $R[w,q] = R_{\pi_e}$, that is, as long as either one of the models works well, the final estimator behaves well.[9]

Besides being useful as an estimator, Eq.(14) also provides a unified framework to analyze the previous estimators, which are all its special cases: Note that $R[w,\mathbf{0}] = R_{\mathrm{w}}[w]$ and $R[\mathbf{0},q] = R_{\mathrm{q}}[q]$, where $\mathbf{0}$ means a constant function that always evaluates to 0. Below we first prove a set of results that unify and generalize the results in Sections 4 and 5, and then state the sample complexity guarantees for the proposed estimators.

Lemma 7. $R[w,q] - R_{\pi_e} = \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{\gamma V^{\pi_e}(s') - \gamma v(s') + q(s,a) - Q^{\pi_e}(s,a)\}]$.

Theorem 8. *Fixing any $q' \in \mathcal{Q}$, if $[Q^{\pi_e} - q'] \in \mathrm{conv}(\mathcal{F})$,*

$$|R[w,q'] - R_{\pi_e}| \le \max_{f\in\mathcal{F}} |L_{\mathrm{w}}(w,f)|,$$

$$|R[\hat{w},q'] - R_{\pi_e}| \le \min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} |L_{\mathrm{w}}(w,f)|.$$

Similarly, fixing any $w' \in \mathcal{W}$, if $[w_{\pi_e/\pi_b} - w'] \in \mathrm{conv}(\mathcal{G})$,

$$|R[w',q] - R_{\pi_e}| \le \max_{g\in\mathcal{G}} |L_{\mathrm{q}}(q,g)|,$$

$$|R[w',\hat{q}] - R_{\pi_e}| \le \min_{q\in\mathcal{Q}} \max_{g\in\mathcal{G}} |L_{\mathrm{q}}(q,g)|.$$

Remark 1. When $q' = \mathbf{0}$, the first statement is reduced to Theorem 2. When $w' = \mathbf{0}$, the second statement is reduced to Theorem 5.

Theorem 9 (Double robust inequality for discriminators (i.i.d case))**.** *Recall that*

$$\hat{w}_n = \arg\min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} L_{\mathrm{w},n}(w,f)^2,$$

$$\hat{q}_n = \arg\min_{q\in\mathcal{Q}} \max_{g\in\mathcal{G}} L_{\mathrm{q},n}(q,g)^2,$$

where $L_{\mathrm{w},n}$ and $L_{\mathrm{q},n}$ are the empirical losses based on a set of n i.i.d samples. We have the following two statements.

(1) Assume $[Q^{\pi_e} - q'] \in \mathrm{conv}(\mathcal{F})$ for some q', and $\forall f \in \mathcal{F}, \|f\|_\infty < C_f$. Then, with probability at least $1-\delta$,

$$\begin{aligned} |R[\hat{w}_n,q'] - R_{\pi_e}| &\lessapprox \min_{w\in\mathcal{W}} \max_{f\in\mathcal{F}} |L_{\mathrm{w}}(w,f)| \\ &+ \mathfrak{R}_n(\mathcal{F},\mathcal{W}) + C_f C_w \sqrt{\frac{\log(1/\delta)}{n}} \end{aligned}$$

where $\mathfrak{R}_n(\mathcal{W},\mathcal{F})$ is the Rademacher complexity [10] of the function class $\{(s,a,s') \mapsto w(s,a)(\gamma f(s',\pi_e) - f(s,a)) : w \in \mathcal{W}, f \in \mathcal{F}\}$.

(2) Assume $[w_{\pi_e/\pi_b} - w'] \in \mathrm{conv}(\mathcal{G})$ for some w', and $\forall g \in \mathcal{G}, \|g\|_\infty < C_g$. Then, with probability at least $1-\delta$,

$$\begin{aligned} |R[w',\hat{q}_n] - R_{\pi_e}| &\lessapprox \min_{q\in\mathcal{Q}} \max_{g\in\mathcal{G}} |L_{\mathrm{q}}(q,g)| \\ &+ \mathfrak{R}_n(\mathcal{Q},\mathcal{G}) + C_g \frac{R_{\max}}{(1-\gamma)} \sqrt{\frac{\log(1/\delta)}{n}}, \end{aligned}$$

where $\mathfrak{R}_n(\mathcal{Q},\mathcal{G})$ is the Rademacher complexity of the function class $\{(s,a,r,s') \mapsto g(s,a)\{r + \gamma q(s',\pi_e) - q(s,a)\} : q \in \mathcal{Q}, g \in \mathcal{G}\}$.

Here $A \lessapprox B$ means inequality without an (absolute) constant. Note that we can immediately extract the sample complexity guarantees for the MWL and the MQL estimators as the corollaries of this general guarantee by letting $q' = \mathbf{0}$ and $w' = \mathbf{0}$.[11] In Appendix C.1 we also extend the analysis to the non-i.i.d. case and show that similar results can be established for β-mixing data.

7. Statistical Efficiency in the Tabular Setting

As we have discussed earlier, both MWL and MQL are equivalent to LSTDQ when we use the same linear class for all function approximators. Here we show that in the tabular setting, which is a special case of the linear setting, MWL and MQL can achieve the semiparametric lower bound of OPE (Kallus and Uehara, 2019a), because they coincide

[9] See Kallus and Uehara (2019b, Theorem 11,12) for formal statements.

[10] See Bartlett and Mendelson (2003) for the definition.

[11] Strictly speaking, when $q' = \mathbf{0}$, $R[\hat{w}_n, q'] = R_{\mathrm{w}}[\hat{w}_n]$ is very close to but slightly different from the sample-based MWL estimator $R_{\mathrm{w},n}[\hat{w}_n]$, but their difference can be bounded by a uniform deviation bound over the $\mathcal{W}$ class in a straightforward manner. The MQL analysis does not have this issue as $R_{\mathrm{q}}[\cdot]$ does not require empirical approximation.

with the model-based solution. This is a desired property that many OPE estimators fail to obtain, including MSWL and MVL.

Theorem 10. *Assume the whole data set $\{(s,a,r,s')\}$ is geometrically ergodic* [12]. *Then, in the tabular setting, $\sqrt{n}(R_{\mathrm{w},n}[\hat{w}_n]-R_{\pi_e})$ and $\sqrt{n}(R_{\mathrm{q}}[\hat{q}_n]-R_{\pi_e})$ weakly converge to the normal distribution with mean 0 and variance*

$$\mathrm{E}_{d_{\pi_b}}[w^2_{\pi_e/\pi_b}(s,a)(r+\gamma V^{\pi_e}(s')-Q^{\pi_e}(s,a))^2].$$

This variance matches the semiparametric lower bound for OPE given by Kallus and Uehara (2019a, Theorem 5).

The details of this theorem and further discussions can be found in Appendix D, where we also show that MSWL and MVL have an asymptotic variance greater than this lower bound. To back up this theoretical finding, we also conduct experiments in the Taxi environment (Dietterich, 2000) following Liu et al. (2018, Section 5), and show that MWL performs significantly better than MSWL in the tabular setting; see Appendix D.3 for details. It should be noted, however, that our optimal claim is asymptotic, whereas explicit importance weighting of MSWL and MVL may provide strong regularization effects and hence preferred in the regime of insufficient data; we leave the investigation to future work.

8. Experiments

We empirically demonstrate the effectiveness of our methods and compare them to baseline algorithms in CartPole with function approximation. We compare MWL & MQL to MSWL (Liu et al., 2018, with estimated behavior policy) and DualDICE (Nachum et al., 2019a). We use neural networks with 2 hidden layers as function approximators for the main function classes for all methods, and use an RBF kernel for the discriminator classes (except for DualDICE); due to space limit we defer the detailed settings to Appendix E. Figure 1 shows the log MSE of relative errors of different methods, where MQL appears to the best among all methods. Despite that these methods require different function approximation capabilities and it is difficult to compare them apple-to-apple, the results still show that MWL/MQL can achieve similar performance to related algorithms and sometimes outperform them significantly.

9. Discussions

We conclude the paper with further discussions.

On the dependence of $\hat{w}$ on $\mathcal{F}$ In Example 1, we have shown that with some special choice of the discriminator class, MWL can pick up very simple weighting functions—such as constant functions—that are very different from the “true” w_{π_e/π_b} and nevertheless produce accurate OPE estimates with very low variances. Therefore, the function w that satisfies $L_\mathrm{w}(w,f)=0\,\forall f\in\mathcal{F}$ may not be unique, and the set of feasible w highly depends on the choice of $\mathcal{F}$. This leads to several further questions, such as how to choose $\mathcal{F}$ to allow for these simple solutions, and how regularization can yield simple functions for better bias-variance trade-off. [13] In case it is not clear that the set of feasible w generally depends on $\mathcal{F}$, we provide two additional examples which are also of independent interest themselves. The first example shows that standard step-wise IS can be viewed as a special case of MWL, when we choose a very rich discriminator class of history-dependent functions.

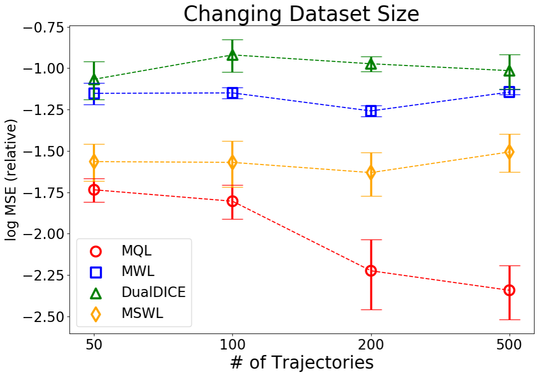

Figure 1. Accuracy of OPE methods as a function of sample size. Error bars show 95% confidence intervals.

Example 8 (Step-wise IS as a Special Case of MWL)**.** Every episodic MDP can be viewed as an equivalent MDP whose state is the *history* of the original MDP. The marginal density ratio in this history-based MDP is essentially the cumulative product of importance weight used in step-wise IS, and from Lemma 11 we know that such a function is the unique minimizer of MWL's population loss if $\mathcal{F}$ is chosen to be a sufficiently rich class of functions over histories. See Appendix F for more details on this example.

Example 9 (Bisimulation)**.** Let ϕ be a bisimulation state abstraction (see Li et al. (2006) for definition). If π_e and π_b only depend on s through $\phi(s)$, and $\mathcal{F}$ only contains functions that are piece-wise constant under ϕ, then $w(s,a)=\frac{d_{\pi_e,\gamma}(\phi(s),a)}{d_{\pi_b}(\phi(s),a)}$ also satisfies $L_\mathrm{w}(w,f)=0,\ \forall f\in\mathcal{F}$.

Duality between MWL and MQL From Sections 4 and 5, one can observe an obvious symmetry between MWL and MQL from the estimation procedures to the guarantees, which reminds us a lot about the duality between value functions and distributions in linear programming for MDPs. Formalizing this intuition is an interesting direction.[14]

[12] Regarding the definition, refer to (Meyn and Tweedie, 2009)

[13] Such a trade-off is relatively well understood in contextual bandits (Kallus, 2020; Hirshberg and Wager, 2017), though extension to sequential decision-making is not obvious.

[14] See the parallel work by Nachum et al. (2019) and the follow-

Acknowledgements

We would like to thank the anonymous reviewers for their insightful comments and suggestions.

Masatoshi Uehara was supported in part by MASASON Foundation.

References

Antos, A., C. Szepesvári, and R. Munos (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning 71*, 89–129.

Bartlett, P. L. and S. Mendelson (2003). Rademacher and Gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research 3*, 463–482.

Bertsekas, D. P. (2012). *Dynamic programming and optimal control* (4th ed. ed.). Athena Scientific optimization and computation series. Belmont, Mass: Athena Scientific.

Bertsekas, D. P. and H. Yu (2009). Projected equation methods for approximate solution of large linear systems. *Journal of Computational and Applied Mathematics 227*, 27–50.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba (2016). OpenAI gym. *arXiv preprint arXiv:1606.01540*.

Chen, J. and N. Jiang (2019). Information-theoretic considerations in batch reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, Volume 97, pp. 1042–1051.

Dann, C., N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire (2018). On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems 31*, pp. 1422–1432.

Dann, C., G. Neumann, and J. Peters (2014). Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research 15*, 809–883.

Dietterich, T. G. (2000). Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of Artificial Intelligence Research 13*, 227–303.

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

Feng, Y., L. Li, and Q. Liu (2019). A kernel loss for solving the bellman equation. In *Advances in Neural Information Processing Systems 32*, pp. 15430–15441.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *Journal of Machine Learning Research 13*, 723–773.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica 66*, 315–331.

Hanna, J., S. Niekum, and P. Stone (2019). Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2605–2613.

Henmi, M. and S. Eguchi (2004). A paradox concerning nuisance parameters and projected estimating functions. *Biometrika 91*, 929–941.

Hirano, K., G. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

Hirshberg, D. A. and S. Wager (2017). Augmented minimax linear estimation. *arXiv preprint arXiv:1712.00038*.

Jiang, N. (2019). On value functions and the agent-environment boundary. *arXiv preprint arXiv:1905.13341*.

Jiang, N. and J. Huang (2020). Minimax confidence interval for off-policy evaluation and policy optimization. *arXiv preprint arXiv:2002.02081*.

Jiang, N., A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire (2017). Contextual Decision Processes with low Bellman rank are PAC-learnable. In *Proceedings of the 34th International Conference on Machine Learning*, Volume 70, pp. 1704–1713.

Jiang, N. and L. Li (2016). Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, Volume 48, pp. 652–661.

Kallus, N. (2020). Generalized optimal matching methods for causal inference. *JMLR (To appear)*.

Kallus, N. and M. Uehara (2019a). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *arXiv preprint arXiv:1908.08526*.

Kallus, N. and M. Uehara (2019b). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

up work of Jiang and Huang (2020) for some intriguing discussions on this matter.

Lagoudakis, M. and R. Parr (2004). Least-squares policy iteration. *Journal of Machine Learning Research 4*, 1107–1149.

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

Li, L., R. Munos, and C. Szepesvari (2015). Toward minimax off-policy value estimation. *In Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 608–616.

Li, L., T. J. Walsh, and M. L. Littman (2006). Towards a unified theory of state abstraction for MDPs. In *Proceedings of the 9th International Symposium on Artificial Intelligence and Mathematics*, pp. 531–539.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371.

Meyn, S. and R. L. Tweedie (2009). *Markov Chains and Stochastic Stability* (2nd ed. ed.). New York: Cambridge University Press.

Mohri, M. and A. Rostamizadeh (2009). Rademacher complexity bounds for non-i.i.d. processes. In *Advances in Neural Information Processing Systems 21*, pp. 1097–1104.

Mohri, M., A. Rostamizadeh, and A. Talwalkar (2012). *Foundations of machine learning*. MIT press.

Nachum, O., Y. Chow, B. Dai, and L. Li (2019a). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32*.

Nachum, O., Y. Chow, B. Dai, and L. Li (2019b). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32*, pp. 2315–2325.

Nachum, O., B. Dai, I. Kostrikov, Y. Chow, L. Li, and D. Schuurmans (2019). Algaedice: Policy gradient from arbitrary experience. *arXiv preprint arXiv:1912.02074*.

Newey, W. K. and D. L. Mcfadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics IV*, 2113–2245.

Precup, D., R. S. Sutton, and S. P. Singh (2000). Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pp. 759–766.

Sriperumbudur, B. K., A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research 11*, 1517–1561.

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

Szepesvari, C. and R. Munos (2005). Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd international conference on machine learning*, pp. 880–887.

Tang, Z., Y. Feng, L. Li, D. Zhou, and Q. Liu (2020). Harnessing infinite-horizon off-policy evaluation: Double robustness via duality. *ICLR 2020(To appear)*.

Tsitsiklis, J. N. and B. Van Roy (1997). An analysis of temporal-difference learning with function approximation. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL 42*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.

Xie, T., Y. Ma, and Y.-X. Wang (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pp. 9665–9675.