Table 2. Table of Notations

π_e, π_b	Evaluation policy, Behavior policy
$\{(s_i, a_i, r_i, s_i')\}_{i=1}^n$	Finite sample of data
$(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, \gamma, d_0), \mathcal{X}$	MDP, $\mathcal{X} = \mathcal{S} \times \mathcal{A}$
d_{π_b}	Data distribution over (s, a, r, s') or its marginals
$d_{\pi_e,\gamma}$	Normalized discounted occupancy induced by π_e
$(s, a) \sim d_0 \times \pi_e$	$s \sim d_0$, $a \sim \pi_e(s)$
$\beta_{\pi_e/\pi_b}(a, s)$	Importance weight on action: $\pi_e(a\|s)/\pi_b(a\|s)$
$w_{\pi_e/\pi_b}(s, a)$	$d_{\pi_e,\gamma}(s, a)/d_{\pi_b}(s, a)$
C_w	Bound on $\|w_{\pi_e/\pi_b}\|_\infty$
V^{π_e}	Value function
Q^{π_e}	Q-value function
R_{π_e}	Expected discounted return of π_e
$R_{\mathrm{w}}[\cdot]$	OPE estimator using $(\cdot)$ as the weighting function (population version)
$R_{\mathrm{w},n}[\cdot]$	OPE estimator using $(\cdot)$ as the weighting function (sample-based version)
$R_{\mathrm{q}}[\cdot]$	OPE estimator using $(\cdot)$ as the approximate Q-function
E_n	Empirical approximation
$\mathcal{W}, \mathcal{F}$	Function classes for MWL
$\mathcal{Q}, \mathcal{G}$	Function classes for MQL
$\langle \cdot, \cdot \rangle_{\mathcal{H}_K}$	Inner product of RKHS with a kernel K
$\mathrm{conv}(\cdot)$	convex hull
ν	Uniform measure over the compact space $\mathcal{X}$
$L^2(\mathcal{X}, \nu)$	L^2-space on X with respect to measure ν
$\mathfrak{R}_n(\cdot)$	Rademacher complexity
$\lesssim$	Inequality without constant

A. Proofs and Additional Results of Section 4 (MWL)

We first give the formal version of Lemma 1, which is Lemmas 11 and 12 below, and then provide their proofs.

Lemma 11. *For any function $g(s, a)$, define the map; $g \to \delta(g, s', a')$;*

$$\delta(g, s', a') = \gamma \int P(s'|s, a)\pi_e(a'|s')g(s, a)\mathrm{d}\nu(s, a) - g(s', a') + (1-\gamma)d_0(s')\pi_e(a'|s').$$

Then, $\delta(d_{\pi_e,\gamma}, s', a') = 0 \, \forall (s', a')$.

Proof. We have

$$\begin{aligned}
d_{\pi_e,\gamma}(s', a') &= (1-\gamma)\sum_{t=0}^{\infty} \gamma^t d_{\pi_e,t}(s', a') \\
&= (1-\gamma)\left\{ d_0(s')\pi_e(a'|s') + \sum_{t=1}^{\infty} \gamma^t d_{\pi_e,t}(s', a') \right\} \\
&= (1-\gamma)\left\{ d_0(s')\pi_e(a'|s') + \sum_{t=0}^{\infty} \gamma^{t+1} d_{\pi_e,t+1}(s', a') \right\} \\
&= (1-\gamma)\left\{ d_0(s')\pi_e(a'|s') + \gamma \sum_{t=0}^{\infty} \int P(s'|s, a)\pi_e(a'|s')\gamma^t d_{\pi_e,t}(s, a)\mathrm{d}\nu(s, a) \right\} \\
&= (1-\gamma)d_0(s')\pi_e(a'|s') + \gamma \int P(s'|s, a)\pi_e(a'|s')d_{\pi_e,\gamma}(s, a)\mathrm{d}\nu(s, a).
\end{aligned}$$

This concludes $\delta(d_{\pi_e,\gamma}, s', a') = 0 \, \forall (s', a')$. □

Lemma 12. *$L_{\mathrm{w}}(w_{\pi_e/\pi_b}, f) = 0\,\forall f \in L^2(\mathcal{X}, \nu) := \{f : \int f(s,a)^2 \mathrm{d}\nu < \infty\}$. Moreover, if we further assume that (a) $d_{\pi_b}(s,a) > 0\,\forall(s,a)$, (b) $g(s,a) = d_{\pi_e,\gamma}(s,a)$ if and only if $\delta(g, s', a') = 0\,\forall(s', a')$, then w_{π_e/π_b} is the only function that satisfies such a property.*

Proof. Here, we denote $\beta_{\pi_e/\pi_b}(s,a) = \pi_e(a|s)/\pi_b(a|s)$. Then, we have

$$
\begin{aligned}
&L_{\mathrm{w}}(w, f) \\
&= \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) f(s', \pi_e) - w(s,a) f(s,a)] + (1-\gamma)\mathrm{E}_{d_0 \times \pi_e}[f(s,a)] \\
&= \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) \beta_{\pi_e/\pi_b}(a', s') f(s', a')] - \mathrm{E}_{d_{\pi_b}}[w(s,a) f(s,a)] + (1-\gamma)\mathrm{E}_{d_0 \times \pi_e}[f(s,a)] \\
&= \gamma \int f(s', a') P(s'|s,a) \pi_e(a'|s') d_{\pi_b}(s,a) w(s,a) \mathrm{d}\nu(s,a,s',a') + \\
&- \int w(s', a') f(s', a') d_{\pi_b}(s', a') \mathrm{d}\nu(s', a') + \int (1-\gamma) d_0(s') \pi_e(a'|s') f(s', a') \mathrm{d}\nu(a', s') \\
&= \int \delta(\tilde{g}, s', a') f(s', a') \mathrm{d}\nu(s', a'),
\end{aligned}
$$

where $\tilde{g}(s,a) = d_{\pi_b}(s,a) w(s,a)$. Note that $\mathrm{E}_{d_{\pi_b}}[\cdot]$ means the expectation with respect to $d_{\pi_b}(s,a) P(s'|s,a) \pi_b(a'|s')$.

First statement

We prove that $L_{\mathrm{w}}(w_{\pi_e/\pi_b}, f) = 0\,\forall f \in L^2(\mathcal{X}, \nu)$. This follows because

$$
L_{\mathrm{w}}(w_{\pi_e/\pi_b}, f) = \int \delta(d_{\pi_e,\gamma}, s', a') f(s', a') \mathrm{d}\nu(s', a') = 0.
$$

Here, we have used Lemma 11: $\delta(d_{\pi_e,\gamma}, s', a') = 0\,\forall(s', a')$.

Second statement

We prove the uniqueness part. Assume $L_{\mathrm{w}}(w, f) = 0\,\forall f \in L^2(\mathcal{X}, \nu)$ holds. Noting the

$$
L_{\mathrm{w}}(w, f) = \langle \delta(\tilde{g}, s', a'), f(s', a') \rangle,
$$

where the inner product is for Hilbert space $L^2(\mathcal{X}, \nu)$, the Riesz representative of the functional $f \to L_{\mathrm{w}}(w, f)$ is $\delta(\tilde{g}, s', a')$. From the Riesz representation theorem and the assumption $L_{\mathrm{w}}(w, f) = 0\,\forall f \in L^2(\mathcal{X}, \nu)$, the Riesz representative is uniquely determined as 0, that is, $\delta(\tilde{g}, s', a') = 0$.

From the assumption (b), this implies $\tilde{g} = d_{\pi_e,\gamma}$. From the assumption (a) and the definition of $\tilde{g}$, this implies $w = w_{\pi_e/\pi_b}$. This concludes the proof. □

Proof of Theorem 2. We prove two helper lemmas first.

Lemma 13.

$$
L_{\mathrm{w}}(w, f) = \mathrm{E}_{d_{\pi_b}}[\{w_{\pi_e/\pi_b}(s,a) - w(s,a)\} \prod f(s,a)],
$$

where $\prod f(s,a) = f(s,a) - \gamma \mathrm{E}_{s' \sim P(s,a), a' \sim \pi_e(s')}[f(s', a')]$.

Proof of Lemma 13.

$$
\begin{aligned}
L_{\mathrm{w}}(w, f) &= L_{\mathrm{w}}(w, f) - L_{\mathrm{w}}(w_{\pi_e/\pi_b}, f) \\
&= \mathrm{E}_{d_{\pi_b}}[\{\gamma\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\beta_{\pi_e/\pi_b}(a', s') f(a', s')] \\
&- \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\} f(s,a)] \\
&= \mathrm{E}_{d_{\pi_b}}[\{w_{\pi_e/\pi_b}(s,a) - w(s,a)\} \prod f(s,a)].
\end{aligned}
$$

□

Lemma 14. *Define*

$$f_g(s,a) = \mathrm{E}_{\pi_e}\left[\sum_{t=0}^{\infty}\gamma^t g(s_t,a_t)|s_0=s,a_0=a\right].$$

Here, the expectation is taken with respect to the density $P(s_1|s_0,a_0)\pi_e(a_1|s_1)P(s_2|s_1,a_1)\cdots$. Then, $f=f_g$ is a solution to $g=\prod f$.

Proof of Lemma 14.

$$\begin{aligned}
&\prod f_g(s,a)\\
&= f_g(s,a) - \gamma \mathrm{E}_{s'\sim P(s,a),a'\sim\pi_e(s')}[f_g(s',a')]\\
&= \mathrm{E}_{\pi_e}\left[\sum_{t=0}^{\infty}\gamma^t g(s_t,a_t)|s_0=s,a_0=a\right] - \mathrm{E}_{\pi_e}\left[\sum_{t=0}^{\infty}\gamma^{t+1} g(s_{t+1},a_{t+1})|s_0=s,a_0=a\right]\\
&= \mathrm{E}_{\pi_e}[g(s_0,a_0)|s_0=s,a_0=a] = g(s,a).
\end{aligned}$$

□

We continue with the main proof of Theorem 2. Here, we have $L_{\mathrm{w}}(w,Q^{\pi_e}) = R_{\pi_e} - R_{\mathrm{w}}[w]$ since

$$\begin{aligned}
L_{\mathrm{w}}(w,Q^{\pi_e}) &= \mathrm{E}_{d_{\pi_b}}[\{w_{\pi_e/\pi_b}(s,a) - w(s,a)\}\prod Q^{\pi_e}(s,a)]\\
&= \mathrm{E}_{d_{\pi_b}}[\{w_{\pi_e/\pi_b}(s,a) - w(s,a)\}\mathrm{E}[r|s,a]]\\
&= \mathrm{E}_{d_{\pi_b}}[\{w_{\pi_e/\pi_b}(s,a) - w(s,a)\}r] = R_{\pi_e} - R_{\mathrm{w}}[w].
\end{aligned}$$

In the first line, we have used Lemma 13. From the first line to the second line, we have used Lemma 14.

Therefore, if $Q^{\pi_e}\in\mathrm{conv}(\mathcal{F})$, for any w,

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

Here, we have used a fact that $\max_{f\in\mathrm{conv}(\mathcal{F})}|L_{\mathrm{w}}(w,f)| = \max_{f\in\mathcal{F}}|L_{\mathrm{w}}(w,f)|$. This is proved as follows. First, $\max_{f\in\mathrm{conv}(\mathcal{F})}|L_{\mathrm{w}}(w,f)|$ is equal to $|L_{\mathrm{w}}(w,\tilde{f})|$ where $\tilde{f} = \sum\lambda_i f_i$ and $\sum\lambda_i = 1$. Since $|L_{\mathrm{w}}(w,\tilde{f})| \le \sum\lambda_i|L_{\mathrm{w}}(w,f_i)| \le \max_{f\in\mathcal{F}}|L_{\mathrm{w}}(w,f)|$, we have

$$\max_{f\in\mathrm{conv}(\mathcal{F})}|L_{\mathrm{w}}(w,f)| \le \max_{f\in\mathcal{F}}|L_{\mathrm{w}}(w,f)|.$$

The reverse direction is obvious.

Finally, from the definition of $\hat{w}$,

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

□

Remark 2. When $Q^{\pi_e}\in\mathrm{conv}(\mathcal{F})$ is only approximately satisfied, it is straightforward to incorporate the approximation errors into Theorem 2, where the theorem statement becomes:

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

For completeness we include the proof below:

Proof. For any w and for any $q^\dagger\in\mathcal{F}$,

$$\begin{aligned}
|R_{\pi_e} - R_{\mathrm{w}}[w]| = |L_{\mathrm{w}}(w,Q^{\pi_e})| &\le |L_{\mathrm{w}}(w,q^\dagger)| + |L_{\mathrm{w}}(w,Q^{\pi_e} - q^\dagger)|\\
&\le \max_{f\in\mathcal{F}}|L_{\mathrm{w}}(w,f)| + |L_{\mathrm{w}}(w,Q^{\pi_e} - q^\dagger)|.
\end{aligned}$$

Therefore, for any w,

$$|R_{\pi_e} - R_{\mathrm{w}}[w]| \leq \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)| + \min_{q^\dagger \in \mathcal{F}} |L_{\mathrm{w}}(w, Q^{\pi_e} - q^\dagger)|.$$

Then, for any w,

$$|R_{\pi_e} - R_{\mathrm{w}}[w]| \leq \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)| + \max_{w^\dagger \in \mathcal{W}} \min_{q^\dagger \in \mathcal{F}} |L_{\mathrm{w}}(w^\dagger, Q^{\pi_e} - q^\dagger)|.$$

Thus,

$$|R_{\pi_e} - R_{\mathrm{w}}[\hat{w}]| \leq \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)| + \max_{w^\dagger \in \mathcal{W}} \min_{q^\dagger \in \mathcal{F}} |L_{\mathrm{w}}(w^\dagger, Q^{\pi_e} - q^\dagger)|.$$

where $\hat{w} = \arg\min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)|$. Notice that the additional term $\max_{w^\dagger \in \mathcal{W}} \min_{q^\dagger \in \mathcal{F}} |L_{\mathrm{w}}(w^\dagger, Q^{\pi_e} - q^\dagger)|$ becomes 0 when $Q^{\pi_e} \in \mathrm{conv}(\mathcal{F})$, in which case we recover Theorem 2. □

Proof of Lemma 17. We have

$$\begin{aligned}
&L_{\mathrm{w}}(w, f)^2 \\
&= \left\{ \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) \mathrm{E}_{a' \sim \pi_e(s')}[f(s', a')] - w(s,a) f(s,a)] + (1-\gamma) \mathrm{E}_{d_0 \times \pi_e}[f(s,a)] \right\}^2 && (15) \\
&= \{ \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) \mathrm{E}_{a' \sim \pi_e(s')}[\langle f, K((s', a'), \cdot) \rangle_{\mathcal{H}_K}] - w(s,a) \langle f, K((s,a), \cdot) \rangle_{\mathcal{H}_K}] && (16) \\
&+ (1-\gamma) \mathrm{E}_{d_0 \times \pi_e}[\langle f, K((s,a), \cdot) \rangle_{\mathcal{H}_K}] \}^2 \\
&= \langle f, f^* \rangle_{\mathcal{H}_K}^2, && (17)
\end{aligned}$$

where

$$f^*(\cdot) = \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) \mathrm{E}_{a' \sim \pi_e(s')}[K((s', a'), \cdot)] - w(s,a) K((s,a), \cdot)] + (1-\gamma) \mathrm{E}_{d_0 \times \pi_e}[K((s,a), \cdot)].$$

Here, from (15) to (16), we have used a reproducing property of RKHS; $f(s,a) = \langle f(\cdot), K((s,a), \cdot) \rangle$. From (16) to (17), we have used a linear property of the inner product.

Therefore,

$$\max_{f \in \mathcal{F}} L_{\mathrm{w}}(w, f)^2 = \max_{f \in \mathcal{F}} \langle f, f^* \rangle_{\mathcal{H}_H}^2 = \langle f^*, f^* \rangle_{\mathcal{H}_H}^2.$$

from Cauchy—Schwarz inequality. This is equal to

$$\begin{aligned}
\max_{f \in \mathcal{F}} L_{\mathrm{w}}(w, f)^2 = \mathrm{E}_{d_{\pi_b}}[&\gamma^2 w(s,a) w(\tilde{s}, \tilde{a}) \mathrm{E}_{a' \sim \pi_e(s'), \tilde{a}' \sim \pi_e(\tilde{s}')}[K((s', a'), (\tilde{s}', \tilde{a}'))]] + \\
&+ \mathrm{E}_{d_{\pi_b}}[w(s,a) w(\tilde{s}, \tilde{a}) K((s,a), (\tilde{s}, \tilde{a}))] \\
&+ (1-\gamma)^2 \mathrm{E}_{d_0 \times \pi_e}[K((s,a), (\tilde{s}, \tilde{a}))] \\
&- 2 \mathrm{E}_{d_{\pi_b}}[\gamma w(s,a) w(\tilde{s}, \tilde{a}) \mathrm{E}_{a' \sim \pi_e(s')}[K((s', a'), (\tilde{s}, \tilde{a}))]] \\
&+ 2\gamma(1-\gamma) \mathrm{E}_{(s,a) \sim d_{\pi_b}, (\tilde{s}, \tilde{a}) \sim d_0 \times \pi_e}[\gamma w(s,a) \mathrm{E}_{a' \sim \pi_e(s')}[K((s', a'), (\tilde{s}, \tilde{a}))]] \\
&- 2(1-\gamma) \mathrm{E}_{(s,a) \sim d_{\pi_b}, (\tilde{s}, \tilde{a}) \sim d_0 \times \pi_e}[w(s,a) w(\tilde{s}, \tilde{a}) K((s,a), (\tilde{s}, \tilde{a}))],
\end{aligned}$$

where the first expectation is taken with respect to the density $d_{\pi_b}(s, a, s') d_{\pi_b}(\tilde{s}, \tilde{a}, \tilde{s}')$.

For example, the term $(1-\gamma)^2 \mathrm{E}_{d_0 \times \pi_e}[K((s,a), (\tilde{s}, \tilde{a}))]$ is derived by

$$\begin{aligned}
&\langle (1-\gamma) \mathrm{E}_{d_0 \times \pi_e}[K((s,a), \cdot)], (1-\gamma) \mathrm{E}_{d_0 \times \pi_e}[K((s,a), \cdot)] \rangle_{\mathcal{H}_K} \\
&= (1-\gamma)^2 \left\langle \int K((s,a), \cdot) d_0(s) \pi(a|s) \nu(a, s), \int K((\tilde{s}, \tilde{a}), \cdot) d_0(\tilde{s}) \pi(\tilde{a}|\tilde{s}) \nu(\tilde{a}, \tilde{s}) \right\rangle_{\mathcal{H}_K} \\
&= (1-\gamma)^2 \int \langle K((s,a), \cdot), K((\tilde{s}, \tilde{a}), \cdot) \rangle_{\mathcal{H}_K} d_0(s) \pi(a|s) d_0(\tilde{s}) \pi(\tilde{a}|\tilde{s}) \mathrm{d}\nu(\tilde{a}, \tilde{s}, \tilde{a}', \tilde{s}') \\
&= (1-\gamma)^2 \int K((s,a), (\tilde{s}, \tilde{a})) d_0(s) \pi(a|s) d_0(\tilde{s}) \pi(\tilde{a}|\tilde{s}) \mathrm{d}\nu(\tilde{a}, \tilde{s}, \tilde{a}', \tilde{s}').
\end{aligned}$$

Other term are derived in a similar manner. Here, we have used a kernel property

$$\langle K((s,a),\cdot), K((\tilde{s},\tilde{a}),\cdot)\rangle_{\mathcal{H}_K} = K((s,a),(\tilde{s},\tilde{a})). \qquad \square$$

Next we show the result mentioned in the main text, that the minimizer of $\max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)$ is unique when $\mathcal{F}$ corresponds to an ISPD kernel.

Theorem 15. *Assume $\mathcal{W}$ is realizable, i.e., $w_{\pi_e/\pi_b} \in \mathcal{W}$ and conditions (a), (b) in Lemma 12. Then for $\mathcal{F} = L^2(\mathcal{X},\nu)$, $\hat{w}(s,a) = w_{\pi_e/\pi_b}(s,a)$ is the unique minimizer of $\max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)$. The same result holds when $\mathcal{F}$ is a RKHS associated with a integrally strictly positive definite (ISPD) kernel $K(\cdot,\cdot)$ (Sriperumbudur et al., 2010).*

Proof. The first statement is obvious from Lemma 12 and the proof is omitted, and here we prove the second statement on the ISPD kernel case. If we can prove Lemma 12 when replacing $L^2(\mathcal{X},\nu)$ with RKHS associated with an ISPD kernel $K(\cdot,\cdot)$, the statement is concluded. More specifically, what we have to prove is

Lemma 16. *Assume conditions in Theorem 15. Then, $L_{\mathrm{w}}(w;f) = 0$, $\forall f \in \mathcal{F}$ holds if and only if $w(s,a) = w_{\pi_e/\pi_b}(s,a)$.*

This is proved by Mercer's theorem (Mohri et al., 2012). From Mercer's theorem, there exist an orthnormal basis $(\phi_j)_{j=1}^{\infty}$ of $L^2(\mathcal{X},\nu)$ such that RKHS is represented as

$$\mathcal{F} = \left\{ f = \sum_{j=1}^{\infty} b_j\phi_j ; (b_j)_{j=1}^{\infty} \in l^2(\mathbb{N}) \text{ with } \sum_{j=1}^{\infty} \frac{\beta_j^2}{\mu_j} < \infty \right\},$$

where each μ_j is a positive value since kernel is ISPD. Suppose there exists $w(s,a) \neq w_{\pi_e/\pi_b}(s,a)$ in $w(s,a) \in \mathcal{W}$ satisfying $L_{\mathrm{w}}(w,f) = 0, \forall f \in \mathcal{F}$. Then, by taking $b_j = 1\,(j = j')$, $b_j = 1\,(j \neq j')$, for any $j' \in \mathbb{N}$, we have $L_{\mathrm{w}}(w,\phi_{j'}) = 0$. This implies $L_{\mathrm{w}}(w,f) = 0$, $\forall f \in L^2(\mathcal{X},\nu) = 0$. This contradicts the original Lemma 12. Then, Lemma 16 is concluded. $\square$

A.1. Proof for Example 1

Suppose $w_0(s,a) = C\ \forall(s,a)$. Then, for $f = Q^{\pi_e}$ we have

$$\begin{aligned} L_{\mathrm{w}}(w_0,f) &= C\mathrm{E}_{d_{\pi_b}}[\{\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)\}] + (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[Q^{\pi_e}(s,a)] \\ &= CR_{\pi_b} - R_{\pi_e}. \end{aligned}$$

Hence $C = R_{\pi_e}/R_{\pi_b}$ satisfies that $L_{\mathrm{w}}(w_0,f) = 0, \forall f \in \mathcal{F}$. From Theorem 2, $|R_{\pi_e} - R_{\mathrm{w}}[\hat{w}]| \leq \min_w \max_f |L_{\mathrm{w}}(w,f)| \leq \max_f |L_{\mathrm{w}}(w_0,f)| = 0$.

A.2. The RKHS Result

Lemma 17 (Formal version of Lemma 3)**.** *Let $\langle\cdot,\cdot\rangle_{\mathcal{H}_K}$ be the inner product of $\mathcal{H}_K$, which satisfies the reproducing property $f(x) = \langle f, K(\cdot,x)\rangle_{\mathcal{H}_K}$. When $\mathcal{F} = \{f \in (\mathcal{X}\to\mathbb{R}) : \langle f,f\rangle_{\mathcal{H}_K} \leq 1\}$, the term $\max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)^2$ has a closed form expression:*

$$\begin{aligned} &\max_{f\in\mathcal{F}} L_{\mathrm{w}}(w,f)^2 \\ &= \mathrm{E}_{d_{\pi_b}}[\gamma^2 w(s,a)w(\tilde{s},\tilde{a})\mathrm{E}_{a'\sim\pi_e(s'),\tilde{a}'\sim\pi_e(\tilde{s}')}[K((s',a'),(\tilde{s}',\tilde{a}'))]] + \\ &+ \mathrm{E}_{d_{\pi_b}}[w(s,a)w(\tilde{s},\tilde{a})K((s,a),(\tilde{s},\tilde{a}))] + (1-\gamma)^2\mathrm{E}_{d_0\times\pi_e}[K((s,a),(\tilde{s},\tilde{a}))] + \\ &- 2\mathrm{E}_{d_{\pi_b}}[\gamma w(s,a)w(\tilde{s},\tilde{a})\mathrm{E}_{a'\sim\pi_e(s')}[K((s',a'),(\tilde{s},\tilde{a}))]] \\ &+ 2\gamma(1-\gamma)\mathrm{E}_{(s,a)\sim d_{\pi_b},(\tilde{s},\tilde{a})\sim d_0\times\pi_e}[\gamma w(s,a)\mathrm{E}_{a'\sim\pi_e(s')}[K((s',a'),(\tilde{s},\tilde{a}))]] \\ &- 2(1-\gamma)\mathrm{E}_{(s,a)\sim d_{\pi_b},(\tilde{s},\tilde{a})\sim d_0\times\pi_e}[w(s,a)w(\tilde{s},\tilde{a})K((s,a),(\tilde{s},\tilde{a}))]. \end{aligned}$$

In the above expression, $(s,a)\sim d_0\times\pi_e$ means $s\sim d_0$, $a\sim\pi_e(s)$. All a' and $\tilde{a}'$ terms are marginalized out in the inner expectations, and when they appear together they are always independent. Similarly, in the first 3 lines when (s,a,s') and $(s,a,\tilde{s}')$ appear together in the outer expectation, they are i.i.d. following the distribution specified in the subscript.

A.3. Details of Example 2 (LSTDQ)

Here we provide the detailed derivation of Eq.(11), that is, the closed-form solution of MWL when both $\mathcal{W}$ and $\mathcal{F}$ are set to the same linear class. We assume that the matrix being inverted in Eq.(11) is non-singular.

We show that Eq.(11) is a solution to the objective function of MWL by showing that $L_{\mathrm{w},n}(w(\cdot;\hat{\alpha}), f) = 0$ for any f in the linear class. This suffices because for any w, the loss $L_{\mathrm{w},n}(w(\cdot;\hat{\alpha}), f)^2$ is non-negative and at least 0, so any w that achieves 0 for all $f \in \mathcal{F}$ is an argmin of the loss.

Consider $w \in \mathcal{W}$ whose parameter is α and $f \in \mathcal{F}$ whose parameter is β. Then

$$\begin{aligned} L_{\mathrm{w},n}(w, f) &= \mathrm{E}_n[\gamma\alpha^\top\phi(s,a)\phi(s',\pi_e)^\top\beta - \alpha^\top\phi(s,a)\phi(s,a)^\top\beta] + (1-\gamma)\mathrm{E}_{d_0}[\phi(s,\pi_e)^\top\beta] \\ &= \left(\alpha^\top\mathrm{E}_n[\gamma\phi(s,a)\phi(s',\pi_e)^\top - \phi(s,a)\phi(s,a)^\top] + (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[\phi(s,a)^\top]\right)\beta. \end{aligned}$$

Since $L_{\mathrm{w},n}(w, f)$ is linear in β, to achieve $\max_f L_{\mathrm{w},n}(w, f)^2 = 0$ it suffices to satisfy

$$\alpha^\top\mathrm{E}_n[\gamma\phi(s,a)\phi(s',\pi_e)^\top - \phi(s,a)\phi(s,a)^\top] + (1-\gamma)\mathrm{E}_{d_0}[\phi(s,\pi_e)^\top] = \mathbf{0}. \tag{18}$$

Note that this is just a set of linear equations where the number of unknowns is the same as the number of equations, and the $\hat{\alpha}$ in Eq.(11) is precisely the solution to Eq.(18) when the matrix multiplied by $\alpha^\top$ is non-singular.

A.4. Connection to Dual DICE (Nachum et al., 2019a)

Nachum et al. (2019a) proposes an extension of Liu et al. (2018) without the knowledge of the behavior policy, which shares the same goal with our MWL in Section 4. In fact, there is an interesting connection between our work and theirs, as our key lemma (12) can be obtained if we take the functional gradient of their loss function. (Ideal) DualDICE with the chi-squared divergence $f(x) = 0.5x^2$ is described as follows;

- Estimate $\nu(s,a)$;

$$\min_\nu 0.5\mathrm{E}_{d_{\pi_b}}[\{\nu(s,a) - (\mathcal{B}\nu)(s,a)\}^2] - (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[\nu(s,a)], \tag{19}$$

  where $(\mathcal{B}\nu)(s,a) = \gamma\mathrm{E}_{s'\sim P(s,a),a'\sim\pi_e(s')}[\nu(s',a')]$.

- Estimate the ratio as $\nu(s,a) - (\mathcal{B}\nu)(s,a)$.

Because this objective function includes an integral in $(\mathcal{B}\nu)(s,a)$, the Monte-Carlo approximation is required. However, even if we take an Monte-Carlo sample for the approximation, it is biased. Therefore, they further modify this objective function into a more complex minimax form. See (11) in (Nachum et al., 2019a).

Here, we take a functional derivative of (19)(Gateaux derivative) with respect to ν. The functional derivative at $\nu(s,a)$ is

$$f(s,a) \to \mathrm{E}_{d_{\pi_b}}[\{\nu(s,a) - (\mathcal{B}\nu)(s,a)\}\{f(s,a) - (\mathcal{B}f)(s,a)\}] - (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[f(s,a)].$$

The first order condition exactly corresponds to our Lemma 12:

$$\begin{aligned} -L_{\mathrm{w}}(w, f) &= \mathrm{E}_{d_{\pi_b}}[w(s,a)\{f(s,a) - \gamma\mathrm{E}_{s'\sim P(s,a),a'\sim\pi_e(s')}[f(s',a')]\}] + (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[f(s,a)] = 0 \\ &\iff \mathrm{E}_{d_{\pi_b}}[w(s,a)\{f(s,a) - \gamma\mathrm{E}_{a'\sim\pi_e(s')}[f(s',a')]\}] + (1-\gamma)\mathrm{E}_{d_0\times\pi_e}[f(s,a)] = 0. \end{aligned}$$

where $w(s,a) = \nu(s,a) - (\mathcal{B}\nu)(s,a)$. Our proposed method with RKHS enables us to directly estimate $w(s,a)$ in one step, and in contrast their approach requires two additional steps: estimating $(\mathcal{B}\nu)(s,a)$ in the loss function, estimating $\nu(s,a)$ by minimizing the loss function, and taking the difference $\nu(s,a) - (\mathcal{B}\nu)(s,a)$.

A.5. Connection between MSWL (Liu et al., 2018) and Off-policy LSTD

Just as our methods become LSTDQ using linear function classes (Examples 2 and 7), here we show that the method of Liu et al. (2018) (which we call MSWL for easy reference) corresponds to off-policy LSTD. Under our notations, their method

is[15]

$$\arg\min_{w\in\mathcal{W}^{\mathcal{S}}}\max_{f\in\mathcal{F}^{\mathcal{S}}}\left\{\mathrm{E}_{d_{\pi_b}}\left[\left(\gamma w(s)f(s')\frac{\pi_e(a|s)}{\pi_b(a|s)}-w(s)f(s)\right)\right]+(1-\gamma)\mathrm{E}_{d_0}[f(s)]\right\}^2. \tag{20}$$

We call this method MSWL (minimax state weight learning) for easy reference. A slightly different but closely related estimator is

$$\arg\min_{w\in\mathcal{W}^{\mathcal{S}}}\max_{f\in\mathcal{F}^{\mathcal{S}}}\left\{\mathrm{E}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\left(\gamma w(s)f(s')-w(s)f(s)\right)\right]+(1-\gamma)\mathrm{E}_{d_0}[f(s)]\right\}^2. \tag{21}$$

Although the two objectives are equal in expectation, under empirical approximations the two estimators are different. In fact, Eq.(21) corresponds to the most common form of off-policy LSTD (Bertsekas and Yu, 2009) when both $\mathcal{W}^{\mathcal{S}}$ and $\mathcal{F}^{\mathcal{S}}$ are linear (similar to Example 2). In the same linear setting, Eq.(20) corresponds to another type of off-policy LSTD discussed by Dann et al. (2014). In the tabular setting, we show that cannot achieve the semiparametric lower bound in Appendix D.

B. Proofs and Additional Results of Section 5 (MQL)

Proof for Lemma 4. **First statement**

We prove $L_{\mathrm{q}}(Q^{\pi_e},g)=0\,\forall g\in L^2(\mathcal{X},\nu)$. The function Q^{π_e} satisfies the following Bellman equation;

$$\mathrm{E}_{r\sim\mathcal{R}(s,a),\,s'\sim P(s,a)}[r+\gamma Q^{\pi_e}(s',\pi_e)-Q^{\pi_e}(s,a)]=0. \tag{22}$$

Then, $\forall g\in L^2(\mathcal{X},\nu)$,

$$\begin{aligned}0&=\int\left\{\mathrm{E}_{r\sim\mathcal{R}(s,a),\,s'\sim P(s,a)}[r+\gamma Q^{\pi_e}(s',\pi_e)-Q^{\pi_e}(s,a)]\right\}g(s,a)d_{\pi_b}(s,a)\mathrm{d}\nu(s,a)\\&=L_{\mathrm{q}}(Q^{\pi_e},g).\end{aligned}$$

Second statement

We prove the uniqueness part. Recall that Q^{π_e} is uniquely characterized as (Bertsekas, 2012): $\forall(s,a)$,

$$\mathrm{E}_{r\sim\mathcal{R}(s,a),\,s'\sim P(s,a)}[r+\gamma q(s',\pi_e)-q(s,a)]=0. \tag{23}$$

Assume

$$\mathrm{E}_{(s,a,r,s')\sim d_{\pi_b}}[\{r+\gamma q(s',\pi_e)-q(s,a)\}g(s,a)]=0,\forall g(s,a)\in L^2(\mathcal{X},\nu) \tag{24}$$

Note that the left hand side term is seen as

$$L_{\mathrm{q}}(q,g)=\langle\{\mathrm{E}_{r\sim\mathcal{R}(s,a)}[r]+\mathrm{E}_{s'\sim P(s,a)}[\gamma q(s',\pi_e)]-q(s,a)\}d_\pi(s,b),g(s,a)\rangle$$

where the inner product $\langle\cdot,\cdot\rangle$ is for Hilbert space $L^2(\mathcal{X},\nu)$ and the representator of the functional $g\to L_{\mathrm{q}}(q,g)$ is $\{\mathrm{E}[r|s,a]+\gamma q(s',\pi_e)-q(s,a)\}d_{\pi_b}(s,a)$. From Riesz representation theorem and the assumption (24), the representator of the linear bounded functional $g\to L_{\mathrm{q}}(q,g)$ is uniquely determined as 0. Since we also assume $d_{\pi_b}(s,a)>0\,\forall(s,a)$, we have $\forall(s,a)$,

$$\mathrm{E}_{r\sim\mathcal{R}(s,a),\,s'\sim P(s,a)}[r+\gamma q(s',\pi_e)-q(s,a)]=0. \tag{25}$$

From (23), such q is Q^{π_e}. □

[15] Here we have simplified their loss for the discounted case (i.e., their Eq.(15)) with an identity $\mathrm{E}_{d_{\pi_b}}[w(s)f(s)]=\gamma\mathrm{E}_{d_{\pi_b}}[w(s')f(s')]+(1-\gamma)\mathrm{E}_{d_0}[w(s)f(s)]$ that holds when d_{π_b} is the discounted occupancy of the behavior policy (which Liu et al. (2018) assume but we do not). Replacing $\mathrm{E}_{d_{\pi_b}}[w(s)f(s)]$ in Eq.(20) with the RHS of this identity will recover Liu et al. (2018)'s original loss. Also, because of this reason, their original loss only applies when d_{π_b} is the discounted occupancy of the behavior policy, whereas this simplified loss (as well as the losses of MWL and MQL) applies even if d_{π_b} is not a valid occupancy but an arbitrary exploratory distribution.

Proof of Theorem 5. We prove the first statement. For fixed any q, we have

$$\begin{aligned}
|R_{\pi_e} - R_{\mathrm{q}}[q]| &= |(1-\gamma)\mathrm{E}_{d_0 \times \pi_e}[q(s,a)] - R_{\pi_e}| \\
&= |\mathrm{E}_{d_{\pi_b}}[-\gamma w_{\pi_e/\pi_b}(s,a) q(s', \pi_e) + w_{\pi_e/\pi_b}(s,a) q(s,a)] - \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a) r]| \\
&= |\mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a)\{r + \gamma q(s', \pi_e) - q(s,a)\}]| \\
&\leq \max_{g \in \mathrm{conv}(\mathcal{G})} |L_{\mathrm{q}}(q,g)| = \max_{g \in \mathcal{G}} |L_{\mathrm{q}}(q,g)|.
\end{aligned}$$

From the first line to the second line, we use Lemma 12 choosing $f(s,a)$ as $q(s,a)$. From second line to the third line, we use $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$.

Then, the second statement follows immediately based on the definition of $\hat{q}$. □

Remark 3. Similar to Remark 2 for Theorem 2, it is also straightforward to incorporate the approximation errors of $w_{\pi_e/\pi_b} \in \mathrm{conv}(\mathcal{G})$ into Theorem 5, and the more general bound is

$$|R_{\pi_e} - R_{\mathrm{q}}[\hat{q}]| \leq \min_{q \in \mathcal{Q}} \max_{g \in \mathcal{G}} |L_{\mathrm{q}}(q,g)| + \max_{q^\dagger \in \mathcal{Q}} \min_{w^\dagger \in \mathcal{G}} |L_{\mathrm{q}}(q^\dagger, w_{\pi_e/\pi_b} - w^\dagger)|.$$

The proof is omitted due to similarity to the MWL case.

Proof of Lemma 6. We have

$$\begin{aligned}
\max_{g \in \mathcal{G}} L_{\mathrm{q}}(q,g)^2 &= \max_{g \in \mathcal{G}} \mathrm{E}_{d_{\pi_b}}[g(s,a)(r + \gamma q(s', \pi_e) - q(s,a))]^2 \\
&= \max_{g \in \mathcal{G}} \mathrm{E}_{d_{\pi_b}}[\langle g, K((s,a),\cdot)\rangle_{\mathcal{H}_K}(r + \gamma q(s', \pi_e) - q(s,a))]^2 \\
&= \max_{g \in \mathcal{G}} \langle g, \mathrm{E}_{d_{\pi_b}}[K((s,a),\cdot)(r + \gamma q(s', \pi_e) - q(s,a))]\rangle_{\mathcal{H}_K}^2 \\
&= \max_{g \in \mathcal{G}} \langle g, g^*\rangle_{\mathcal{H}_K}^2 = \langle g^*, g^*\rangle_{\mathcal{H}_K}
\end{aligned}$$

where

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

From the first line to the second line, we use a reproducing property of RKHS; $g(s,a) = \langle g(\cdot), K((s,a),\cdot)\rangle_{\mathcal{H}_K}$. From the second line to the third line, we use a linear property of the inner product. From third line to the fourth line, we use a Cauchy–schwarz inequality since $\mathcal{G} = \{g; \langle g, g\rangle_{\mathcal{H}_K} \leq 1\}$.

Then, the last expression $\langle g^*, g^*\rangle_{\mathcal{H}_K}$ is equal to

$$\begin{aligned}
&\langle \mathrm{E}_{d_{\pi_b}}[K((s,a),\cdot)(r + \gamma q(s', \pi_e) - q(s,a))], \mathrm{E}_{d_{\pi_b}}[K((s,a),\cdot)(r + \gamma q(s', \pi_e) - q(s,a))]\rangle_{\mathcal{H}_K} \\
&= \langle \mathrm{E}_{d_{\pi_b}}[K((s,a),\cdot)(\mathrm{E}_{r \sim \mathcal{R}(s,a)}[r] + \mathrm{E}_{s' \sim P(s,a)}[\gamma q(s', \pi_e)] - q(s,a))], \\
&\mathrm{E}_{d_{\pi_b}}[K((\tilde{s},\tilde{a}),\cdot)(\mathrm{E}_{\tilde{r} \sim \mathcal{R}(\tilde{s},\tilde{a})}[\tilde{r}] + \mathrm{E}_{\tilde{s}' \sim P(\tilde{s},\tilde{a})}[\gamma v(\tilde{s}')] - q(\tilde{s},\tilde{a}))]\rangle \\
&= \mathrm{E}_{d_{\pi_b}}[\Delta^q(q; s,a,r,s')\Delta^q(q; \tilde{s},\tilde{a},\tilde{r},\tilde{s}')K((s,a),(\tilde{s},\tilde{a}))]
\end{aligned}$$

where $\Delta^q(q; s,a,r,s') = r + \gamma q(s', \pi_e) - q(s,a)$ and the expectation is taken with respect to the density $d_{\pi_b}(s,a,r,s') d_{\pi_b}(\tilde{s},\tilde{a},\tilde{r},\tilde{s}')$. Here, we have used a kernel property

$$\langle K((s,a),\cdot), K((\tilde{s},\tilde{a}),\cdot)\rangle_{\mathcal{H}_K} = K((s,a),(\tilde{s},\tilde{a})).$$

□

Theorem 18. *Assume Q^{π_e} is included in $\mathcal{Q}$ and $d_{\pi_b}(s,a) > 0 \forall (s,a)$. Then, if $\mathcal{G}$ is $L^2(\mathcal{X}, \nu)$, $\hat{q} = Q^{\pi_e}$. Also if $\mathcal{G}$ is a RKHS associated with an ISPD kernel, $\hat{q} = Q^{\pi_e}$*

Proof of Theorem 18. The first statement is obvious from Lemma 4. The second statement is proved similarly as Theorem 15. □

B.1. Proof for Example 3

Suppose $q_0(s,a) = C$. Then, for $g = w_{\pi_e/\pi_b}$, we have

$$\begin{aligned} L_{\mathrm{q}}(q,g) &= \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a)(r+\gamma C - C)] \\ &= \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a)r] - (1-\gamma)C = R_{\pi_e} - (1-\gamma)C. \end{aligned}$$

Therefore $C = R_{\pi_e}/(1-\gamma)$ satisfies $L_{\mathrm{q}}(q,g) = 0\ \forall g \in \mathcal{G}$. From Theorem 5 we further have $R_{\mathrm{q}}[\hat{q}] = R_{\pi_e}$.

B.2. Minimax Value Learning (MVL)

We can extend the loss function in Eq.(13) to the off-policy setting as follows: for state-value function $v : \mathcal{S} \to \mathbb{R}$, define the following two losses:

$$\max_{g \in \mathcal{G}^{\mathcal{S}}} \mathrm{E}_{\pi_b}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma v(s') - v(s)\}g(s)\right]^2, \tag{26}$$

$$\max_{g \in \mathcal{G}^{\mathcal{S}}} \mathrm{E}_{\pi_b}\left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma v(s')\} - v(s)\right)g(s)\right]^2. \tag{27}$$

Again, similar to the situation of MSWL as we discussed in Appendix A.5, these two losses are equal in expectation but exhibit different finite sample behaviors. 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., 2014; Bertsekas and Yu, 2009) and coincide with MSWL and its variant.

C. Proofs and Additional Results of Section 6 (DR and Sample Complexity)

Proof of Lemma 7. We have

$$\begin{aligned} & R[w,q] - R_{\pi_e} \\ =& R[w,q] - R[w_{\pi_e/\pi_b}(s,a), Q^{\pi_e}(s,a)] && (28)\\ =& \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{r - Q^{\pi_e}(s,a) + \gamma V^{\pi_e}(s')\}] + && (29)\\ & \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a)\{Q^{\pi_e}(s,a) - q(s,a) + \gamma q(s',\pi_e) - \gamma V^{\pi_e}(s')\}] + (1-\gamma)\mathrm{E}_{d_0}[q(s,\pi_e) - V^{\pi_e}(s)] + \\ & \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{Q^{\pi_e}(s,a) - q(s,a) + \gamma q(s',\pi_e) - \gamma V^{\pi_e}(s')\}] + \\ =& \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{Q^{\pi_e}(s,a) - q(s,a) + \gamma q(s',\pi_e) - \gamma V^{\pi_e}(s')\}]. && (30) \end{aligned}$$

From (28) to (29), this is just by algebra following the definition of $R[\cdot,\cdot]$. From (29) to (30), we use the following lemma. □

Lemma 19.

$$\begin{aligned} 0 &= \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}(s,a)\{Q^{\pi_e}(s,a) - q(s,a) + \gamma q(s',\pi_e) - \gamma V^{\pi_e}(s')\}] + (1-\gamma)\mathrm{E}_{d_0}[q(s,\pi_e) - V^{\pi_e}(s)], \\ 0 &= \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{r - Q^{\pi_e}(s,a) + \gamma V^{\pi_e}(s')\}]. \end{aligned}$$

Proof. The first equation comes form Lemma 12 with $f(s,a) = q(s,a) - Q^{\pi}(s,a)$. The second equation comes from Lemma 4 with $g(s,a) = w(s,a) - w_{\pi_e/\pi_b}(s,a)$. □

Proof of Theorem 8. We begin with the second statement, which is easier to prove from Lemma 7:

$$\begin{aligned} R[w',q] - R_{\pi_e} &= \mathrm{E}_{d_{\pi_b}}[\{w'(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{-\gamma V^{\pi_e}(s') - \gamma q(s',\pi_e) + \gamma q(s',\pi_e) + Q^{\pi_e}(s,a)\}] \\ &= L_{\mathrm{q}}(q, w' - w_{\pi_e/\pi_b}) - L_{\mathrm{q}}(Q^{\pi_e}, w' - w_{\pi_e/\pi_b}) \\ &= -L_{\mathrm{q}}(q, w_{\pi_e/\pi_b} - w') - 0. && \text{(Lemma 4)} \end{aligned}$$

Thus, if $(w_{\pi_e/\pi_b} - w') \in \mathrm{conv}(\mathcal{G})$

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

Next, we prove the first statement. From Lemma 7,

$$\begin{aligned} R[w, q'] - R_{\pi_e} &= \mathrm{E}_{d_{\pi_b}}[\{w(s,a) - w_{\pi_e/\pi_b}(s,a)\}\{-\gamma V^{\pi_e}(s') - \gamma q'(s', \pi_e) + \gamma q(s', \pi_e) + Q^{\pi_e}(s,a)\}] \\ &= L_{\mathrm{w}}(w, q' - Q^{\pi_e}) - L_{\mathrm{w}}(w_{\pi_e/\pi_b}, q' - Q^{\pi_e}) \\ &= -L_{\mathrm{w}}(w, Q^{\pi_e} - q') - 0. \qquad \text{(Lemma 12)} \end{aligned}$$

Then, if $(Q^{\pi_e} - q') \in \mathrm{conv}(\mathcal{F})$,

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

Finally, from the definition of $\hat{w}$ and $\hat{q}$, we also have

$$|R[\hat{w}, q'] - R_{\pi_e}| \leq \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)|, \; |R[w', \hat{q}] - R_{\pi_e}| \leq \min_{q \in \mathcal{Q}} \max_{g \in \mathcal{G}} |L_{\mathrm{q}}(q, g)|. \qquad \square$$

Proof of Theorem 9. We prove the first statement. The second statement is proved in the same way. We have

$$\begin{aligned} &|R[\hat{w}_n, q'] - R_{\pi_e}| \\ &\leq \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}_n, f)| \\ &= \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}_n, f)| - \max_{f \in \mathcal{F}} |L_{\mathrm{w},n}(\hat{w}_n, f)| + \max_{f \in \mathcal{F}} |L_{\mathrm{w},n}(\hat{w}_n, f)| - \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}, f)| + \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}, f)| \\ &\leq \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}_n, f)| - \max_{f \in \mathcal{F}} |L_{\mathrm{w},n}(\hat{w}_n, f) + \max_{f \in \mathcal{F}} |L_{\mathrm{w},n}(\hat{w}, f) - \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}, f)| + \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(\hat{w}, f)| \\ &\leq 2 \max_{f \in \mathcal{F}, w \in \mathcal{W}} ||L_{\mathrm{w},n}(w, f)| - |L_{\mathrm{w}}(w, f)|| + \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)|. \end{aligned} \tag{31}$$

The remaining task is to bound term $\max_{f \in \mathcal{F}, w \in \mathcal{W}} ||L_{\mathrm{w},n}(w, f)| - |L_{\mathrm{w}}(w, f)||$. This is bounded as follows;

$$\max_{f \in \mathcal{F}, w \in \mathcal{W}} ||L_{\mathrm{w},n}(w, f)| - |L_{\mathrm{w}}(w, f)|| \lesssim \mathfrak{R}'_n(\mathcal{F}, \mathcal{W}) + C_f C_w \sqrt{\log(1/\delta)/n}. \tag{32}$$

where $\mathfrak{R}'_n(\mathcal{F}, \mathcal{W})$ is the Rademacher complexity of the function class

$$\{(s, a, s') \mapsto |w(s,a)(\gamma f(s', \pi_e) - f(s,a))| : w \in \mathcal{W}, f \in \mathcal{F}\}.$$

Here, we just used an uniform law of large number based on the Rademacher complexity noting $|\,(w(s,a)(\gamma f(s', \pi_e) - f(s,a)))\,|$ is uniformly bounded by $C_f C_w$ up to some constant (Bartlett and Mendelson, 2003, Theorem 8). From the contraction property of the Rademacher complexity (Bartlett and Mendelson, 2003, Theorem 12),

$$\mathfrak{R}'_n(\mathcal{F}, \mathcal{W}) \leq 2\mathfrak{R}_n(\mathcal{F}, \mathcal{W})$$

where $\mathfrak{R}_n(\mathcal{F}, \mathcal{W})$ is the Rademacher complexity of the function class

$$\{(s, a, s') \mapsto w(s,a)(\gamma f(s', \pi_e) - f(s,a)) : w \in \mathcal{W}, f \in \mathcal{F}\}. \tag{33}$$

Finally, Combining (31), (32) and (33), the proof is concluded. $\square$

C.1. Relaxing the i.i.d. data assumption

Although the sample complexity results in Section 6 are established under i.i.d. data, we show that under standard assumptions we can also handle dependent data and obtain almost the same results. For simplicity, we only include the result for $\hat{w}_n$.

In particular, we consider the setting mentioned in Section 2, that our data is a single long trajectory generated by policy π_b:

$$s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T.$$

We assume that the Markov chain induced by π_b is ergodic, and s_0 is sampled from its stationary distribution so that the chain is stationary. In this case, d_{π_b} corresponds to such a stationary distribution, which is also the marginal distribution of any s_t. We convert this trajectory into a set of transition tuples $\{(s_i, a_i, r_i, s_i')\}_{i=0}^{n-1}$ with $n = T$ and $s_i' = s_{i+1}$, and then apply our estimator on this data. Under the standard β-mixing condition[16] (see e.g., Antos et al., 2008), we can prove a similar sample complexity result:

Corollary 20. *Assume $\{s_i, a_i, r_i, s_i'\}_{i=1}^n$ follows a stationary β–mixing distribution with β–mixing coefficient $\beta(k)$ for $k = 0, 1, \cdots$. For any $a_1, a_2 > 0$ with $2a_1a_2 = n$ and $\delta > 4(a_1 - 1)\beta(a_2)$, with probability at least $1 - \delta$, we have (all other assumptions are the same as in Theorem 9(1))*

$$|R[\hat{w}_n, q] - R_{\pi_e}| \lesssim \min_{w \in \mathcal{W}} \max_{f \in \mathcal{F}} |L_{\mathrm{w}}(w, f)| + \hat{\mathfrak{R}}_{a_1}(\mathcal{F}, \mathcal{W}) + C_f C_w \sqrt{\frac{\log(1/\delta')}{a_1}}$$

where $\hat{\mathfrak{R}}_{a_1}(\mathcal{F}, \mathcal{W})$ is the empirical Rademacher complexity of the function class $\{(s, a, s') \mapsto \{w(s, a)(\gamma f(s', \pi_e) - f(s, a)) : w \in \mathcal{W}, f \in \mathcal{F}\}$ based on a selected subsample of size a_1 from the original data (see Mohri and Rostamizadeh (2009, Section 3.1) for details), and $\delta' = \delta - 4(a_1 - 1)\beta(a_2)$.

Proof Sketch of Corollary 20. We can prove in the same way as for Theorem 9. The only difference is we use Theorem 2 (Mohri and Rostamizadeh, 2009) to bound the term $\sup_{f \in \mathcal{F}, w \in \mathcal{W}} ||L_{\mathrm{w},n}(w, f)| - |L_{\mathrm{w}}(w, f)||$. □

D. Statistical Efficiency in the Tabular Setting

D.1. Statistical efficiency of MWL and MQL

As we already have seen in Example 7, when $\mathcal{W}, \mathcal{F}, \mathcal{Q}, \mathcal{G}$ are the same linear class, MWL, MQL, and LSTDQ give the same OPE estimator. These methods are also equivalent in the tabular setting—as tabular is a special case of linear representation (with indicator features)—which also coincides with the model-based (or certainty-equivalent) solution. Below we prove that this tabular estimator can achieve the semiparametric lower bound for infinite horizon OPE (Kallus and Uehara, 2019a). [17] Though there are many estimators for OPE, many of the existing OPE methods do not satisfy this property.

Here, we have the following theorem; the proof is deferred to Appendix D.4.

Theorem 21 (Restatement of Theorem 10)**.** *Assume the whole data set $\{(s, a, r, s')\}$ is geometrically Ergodic [18]. 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).

Two remarks are in order:

1. Theorem 21 could be also extended to the continuous sample space case in a nonparametric manner, i.e., replacing $\phi(s, a)$ with some basis functions for L^2–space and assuming that its dimension grows with some rate related to n and the data–generating process has some smoothness condition (Newey and Mcfadden, 1994). The proof is not obvious and we leave it to future work.

2. In the contextual bandit setting, it is widely known that the importance sampling estimator with plug-in weight from the empirical distribution and the model-based approach can achieve the semiparametric lower bound (Hahn, 1998; Hirano et al., 2003). Our findings are consistent with this fact and is novel in the MDP setting to the best of our knowledge.

[16] Refer to (Meyn and Tweedie, 2009) regarding the definition.

[17] Semiparametric lower bound is the non–parametric extension of Cramer–Rao lower bound (Bickel et al., 1998). It is the lower bound of asymptotic MSE among regular estimators (van der Vaart, 1998).

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

Table 3. Summary of the connections between several OPE methods and LSTD, and their optimality in the tabular setting.

	MWL	MQL	MSWL	MVL
Definition	Sec 4	Sec 5	Appendix A.5	Appendix B.2
Linear case	LSTDQ		Off-policy LSTD	
Optimality in tabular	Yes		No	

D.2. Statistical inefficiency of MSWL and MVL for OPE

Here, we compare the statistical efficiency of MWL, MQL with MSWL, MVL in the tabular setting. First, we show that MSWL, MVL positing the linear class is the same as the off-policy LSTD (Bertsekas and Yu, 2009; Dann et al., 2014). Then, we calculate the asymptotic MSE of these estimators in the tabular case and show that this is larger than the ones of MWL and MQL.

Equivalence of MSWL, MVL with linear models and off-policy LSTD By slightly modifying Liu et al. (2018, Theorem 4), MSWL is introduced based on the following relation;

$$\mathrm{E}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\left(\gamma\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}f(s')-\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}f(s)\right)\right]+(1-\gamma)\mathrm{E}_{d_0}[f(s)]=0\,\forall f\in L^2(\mathcal{S},\nu).$$

Then, the estimator for $\frac{d_{\pi_e}(s)}{d_{\pi_b}(s)}$ is given as

$$\min_{w^{\mathcal{S}}}\max_{f\in\mathcal{F}^{\mathcal{S}}}\left\{\mathrm{E}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\left(\gamma w(s)f(s')-w(s)f(s)\right)\right]+(1-\gamma)\mathrm{E}_{d_0}[f(s)]\right\}^2. \tag{34}$$

As in Example 2, in the linear model case, let $z(s)=\phi(s)^\top\alpha$ where $\phi(s)\in\mathbb{R}^d$ is some basis function and α is the parameters. Then, the resulting estimator for $d_{\pi_e,\gamma}(s)/d_{\pi_b}(s)$ is

$$\hat{\alpha}=(1-\gamma)\mathrm{E}_n\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\left\{-\gamma\phi(s')+\phi(s)\right\}\phi^\top(s)\right]^{-1}\mathrm{E}_{d_0}[\phi(s)]$$

Then, the final estimator for R_{π_e} is

$$(D_{v_1})^\top D_{v_2}^{-1}D_{v_3},$$

where

$$\begin{aligned}
D_{v_1}&=(1-\gamma)\mathrm{E}_{d_0}[\phi(s)],\\
D_{v_2}&=\mathrm{E}_n\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\phi(s)\left\{-\gamma\phi^\top(s')+\phi^\top(s)\right\}\right],\\
D_{v_3}&=\mathrm{E}_n\left[r\frac{\pi_e(a|s)}{\pi_b(a|s)}\phi(s)\right].
\end{aligned}$$

In MVL, the estimator for $V^{\pi_e}(s)$ is constructed based on the relation;

$$\mathrm{E}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\left\{r+\gamma V^{\pi_e}(s')-V^{\pi_e}(s)\right\}g(s)\right]=0\,\forall g\in L^2(\mathcal{S},\nu).$$

Then, the estimator for $V^{\pi_e}(s)$ is given by

$$\min_{v\in\mathcal{V}}\max_{g\in\mathcal{G}^{\mathcal{S}}}\mathrm{E}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r+\gamma v(s')-v(s)\}g(s)\right]^2.$$

As in Example 7, in the linear model case, let $v(s) = \phi(s)^\top \beta$ where $\phi(s) \in \mathbb{R}^d$ is some basis function and β is the parameters. Then, the resulting estimator for $V^{\pi_e}(s)$ is

$$\hat{\beta} = \mathrm{E}_n\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\phi(s)\left\{-\gamma\phi^\top(s') + \phi^\top(s)\right\}\right]^{-1}\mathrm{E}_n\left[r\frac{\pi_e(a|s)}{\pi_b(a|s)}\phi(s)\right].$$

Then, the final estimator for R_{π_e} is still $(D_{v_1})^\top D_{v_2}^{-1} D_{v_3}$. This is exactly the same as the estimator obtained by off-policy LSTD (Bertsekas and Yu, 2009).

Another formulation of MSWL and MVL According to Liu et al. (2018, Theorem 4), we have

$$\mathrm{E}_{d_{\pi_b}}\left[\left(\gamma\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\frac{\pi_e(a|s)}{\pi_b(a|s)} - \frac{d_{\pi_e,\gamma}(s')}{d_{\pi_b}(s')}\right)f(s')\right] + (1-\gamma)\mathrm{E}_{d_0}[f(s)] = 0\,\forall f \in L^2(\mathcal{S},\nu).$$

They construct an estimator for $\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}$ as;

$$\min_{w\in\mathcal{W}^{\mathcal{S}}}\max_{f\in\mathcal{F}^{\mathcal{S}}}\left\{\mathrm{E}_{d_{\pi_b}}\left[\left(\gamma w(s)f(s')\frac{\pi_e(a|s)}{\pi_b(a|s)} - w(s)f(s)\right)\right] + (1-\gamma)\mathrm{E}_{d_0}[f(s)]\right\}^2.$$

Note that compared with the previous case (34), the position of the importance weight π_e/π_b is different. In the same way, MVL is constructed base on the relation;

$$\mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{\pi_e(a|s)}{\pi_b(a|s)}(r + \gamma V^{\pi_e}(s')) - V^{\pi_e}(s)\right\}g(s)\right] = 0\,\forall g \in L^2(\mathcal{S},\nu).$$

The estimator for $V^{\pi_e}(s)$ is given by

$$\min_{v\in\mathcal{V}}\max_{g\in\mathcal{G}^{\mathcal{S}}}\mathrm{E}_{\pi_b}\left[\left(\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma v(s')\} - v(s)\right)g(s)\right]^2.$$

When positing linear models, in both cases, the final estimator for R_{π_e} is

$$D_{v_1}^\top\{D_{v_4}\}^{-1}D_{v_3},$$

where

$$D_{v_4} = \mathrm{E}_n\left[\phi(s)\left\{-\gamma\frac{\pi_e(a|s)}{\pi_b(a|s)}\phi^\top(s') + \phi^\top(s)\right\}\right].$$

This is exactly the same as the another type of off-policy LSTD (Dann et al., 2014).

Statistical inefficiency of MSWL, MVL and off-policy LSTD Next, we calculate the asymptotic variance of $D_{v1}D_{v2}^{-1}D_{v3}$ and $D_{v1}D_{v4}^{-1}D_{v3}$ in the tabular setting. It is shown that these methods cannot achieve the semiparametric lower bound (Kallus and Uehara, 2019a). These results show that these methods are statistically inefficient. Note that this implication is also brought to the general continuous sample space case since the the asymptotic MSE is generally the same even in the continuous sample space case with some smoothness conditions.

Theorem 22. *Assume the whole data is geometrically Ergodic. In the tabular setting,* $\sqrt{n}(D_{v1}D_{v2}^{-1}D_{v3} - R_{\pi_e})$ *weakly converges to the normal distribution with mean* 0 *and variance;*

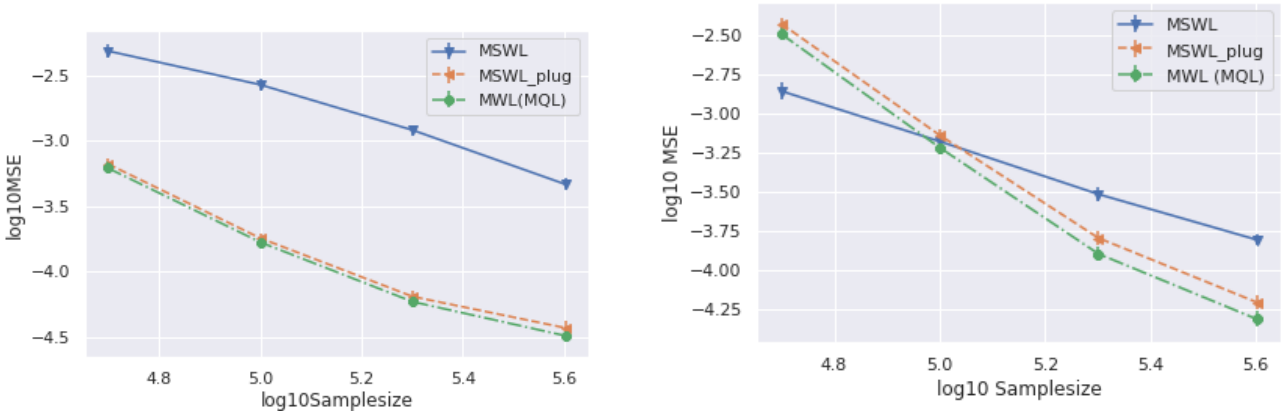
$$\mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2\mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + V^{\pi_e}(s') - V^{\pi_e}(s)\}|s\right]\right].$$

This is larger than the semiparametric lower bound.

Theorem 23. *Assume the whole data is geometrically Ergodic. In the tabular setting,* $\sqrt{n}(D_{v1}D_{v4}^{-1}D_{v3} - R_{\pi_e})$ *weakly converges to the normal distribution with mean* 0 *and variance;*

$$\mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2\mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + V^{\pi_e}(s')\}|s\right]\right].$$

This is larger than the semiparametric lower bound.



(a) $\alpha = 0.2$ (b) $\alpha = 0.4$

Figure 2. MSE as a function of sample size. α controls the difference between π_b and π_e; see Appendix D.3 for details.

D.3. Experiments

We show some empirical results that back up the theoretical discussions in this section. We conduct experiments in the Taxi environment (Dietterich, 2000), which has 20000 states and 6 actions; see Liu et al. (2018, Section 5) for more details. We compare three methods, all using the tabular representation: MSWL with exact π_e, MSWL with estimated π_e ("plug-in"), and MWL (same as MQL). As we have mentioned earlier, this comparison is essentially among off-policy LSTD, plug-in off-policy LSTD, and LSTDQ.

We choose the target policy π_e to be the one obtained after running Q-learning for 1000 iterations, and choose another policy π_+ after 150 iterations. The behavior policy is $\pi_b = \alpha\pi_e + (1-\alpha)\pi_+$. We report the results for $\alpha \in \{0.2,\ 0.4\}$. The discount factor is $\gamma = 0.98$.

We use a single trajectory and vary the truncation size T as $[5, 10, 20, 40] \times 10^4$. For each case, by making 200 replications, we report the Monte Carlo MSE of each estimator with their 95% interval in Figure 2.

It is observed that MWL is significantly better than MSWL and MWL is slightly better than plug-in MSWL. This is because MWL is statisitcally efficient and MSWL is statistically inefficient as we have shown earlier in this section. The reason why plug-in MSWL is superior to the original MSWL is that the plug-in based on MLE with a well specified model can be viewed as a form of control variates (Henmi and Eguchi, 2004; Hanna et al., 2019). Whether the plug-in MSWL can achieve the semiparametric lower bound remains as future work.

D.4. Proofs of Theorems 21, 22, and 23

Proof of Theorem 21. Recall that the estimator is written as $D_{q1}^\top D_{q2}^{-1} D_{q3}$, where

$$
\begin{aligned}
D_{q1} &= (1-\gamma)\mathrm{E}_{d_0 \times \pi_e}[\phi(s,a)] \\
D_{q2} &= \mathrm{E}_n[-\gamma\phi(s,a)\phi^\top(s',\pi_e) + \phi(s,a)\phi(s,a)^\top] \\
D_{q3} &= \mathrm{E}_n[r\phi(s,a)].
\end{aligned}
$$

Recall that $D_{q2}^{-1} D_{q3} = \hat{\beta}$ is seen as Z–estimator with a parametric model $q(s,a;\beta) = \beta^\top\phi(s,a)$. More specifically, the estimator $\hat{\beta}$ is given as a solution to

$$
\mathrm{E}_n[\{r + \gamma q(s', \pi_e; \beta) - q(s,a;\beta)\}\phi(s,a)] = 0.
$$

Following the standard theory of Z–estimator (van der Vaart, 1998), the asymptotic MSE of β is calculated as a sandwich estimator;

$$
\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, D_1^{-1} D_2 D_1^{-1^\top})
$$

where $\beta_0^\top \phi(s,a) = Q^{\pi_e}(s,a)$ and

$$
\begin{aligned}
D_1 &= \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\{-\gamma\phi(s',\pi_e)+\phi(s,a)\}^\top]|_{\beta_0}, \\
D_2 &= \mathrm{var}_{d_{\pi_b}}[\{r+\gamma q(s',\pi_e;\beta)-q(s,a;\beta)\}\phi(s,a)]|_{\beta_0} && (35)\\
&= \mathrm{E}_{d_{\pi_b}}[\mathrm{var}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s')-Q^{\pi_e}(s,a)|s,a]\phi(s,a)\phi^\top(s,a)] && (36)\\
&+ \mathrm{var}_{d_{\pi_b}}[\mathrm{E}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s')-Q^{\pi_e}(s,a)|s,a]\phi(s,a)\phi^\top(s,a)] \\
&= \mathrm{E}_{d_{\pi_b}}[\mathrm{var}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s')-Q^{\pi_e}(s,a)|s,a]\phi(s,a)\phi^\top(s,a)]. && (37)
\end{aligned}
$$

Here, we use a variance decomposition to simplify D_2 from (35) to (36). We use a relation $\mathrm{E}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)|s,a] = 0$ from (36) to (37). Then, by delta method,

$$
\sqrt{n}(D_{q1}^\top D_{q2}^{-1} D_{q3} - R_{\pi_e}) \xrightarrow{d} \mathcal{N}(0, D_{q1}^\top D_1^{-1} D_2 {D_1^{-1}}^\top D_{q1}).
$$

From now on, we simplify the expression $D_{q1}^\top D_1^{-1} D_2 {D_1^{-1}}^\top D_{q1}$. First, we observe

$$
[D_{q1}]_{|S|i_1+i_2} = (1-\gamma) d_0(S_{i_1}) \pi_e(A_{i_2}|S_{i_1}),
$$

where $[D_{q1}]_{|S|i_1+i_2}$ is a element corresponding (S_{i_1}, A_{i_2}) of D_{q1}. In addition,

$$
\begin{aligned}
D_1^{-1} &= \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\{-\gamma\phi(s',\pi_e)+\phi(s,a)\}^\top]^{-1} \\
&= \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\phi^\top(s,a)(-\gamma P^{\pi_e}+I)^\top]^{-1} \\
&= \{(-\gamma P^{\pi_e}+I)^{-1}\}^\top \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\phi^\top(s,a)]^{-1}.
\end{aligned}
$$

where P^{π_e} is a transition matrix between (s,a) and (s',a'), and I is an identity matrix.

Therefore, by defining $g(s,a) = \mathrm{var}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)|s,a]$ and $(I-\gamma P^{\pi_e})^{-1} D_{q1} = D_3$ the asymptotic variance is

$$
\begin{aligned}
&D_3^\top \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\phi^\top(s,a)]^{-1} \mathrm{E}_{d_{\pi_b}}[g(s,a)\phi(s,a)\phi(s,a)^\top] \mathrm{E}_{d_{\pi_b}}[\phi(s,a)\phi^\top(s,a)]^{-1} D_3 \\
&= \sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \{d_{\pi_b}(\tilde{s},\tilde{a})\}^{-1} g(\tilde{s},\tilde{a}) \{D_3^\top I_{\tilde{s},\tilde{a}}\}^2
\end{aligned}
$$

where $I_{\tilde{s},\tilde{a}}$ is a $|S||A|$–dimensional vector, which the element corresponding $(\tilde{s},\tilde{a})$ is 1 and other elements are 0. Noting $D_3^\top I_{\tilde{s},\tilde{a}} = d_{\pi_e,\gamma}(\tilde{s},\tilde{a})$, the asymptotic variance is

$$
\begin{aligned}
&\sum_{\tilde{s}\in\mathcal{S},\tilde{a}\in\mathcal{A}} \{d_{\pi_b}(\tilde{s},\tilde{a})\}^{-1} g(\tilde{s},\tilde{a}) d_{\pi_e,\gamma}^2(\tilde{s},\tilde{a}) \\
&= \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}^2(s,a) \mathrm{var}_{d_{\pi_b}}[r+\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)|s,a]] \\
&= \mathrm{E}_{d_{\pi_b}}[w_{\pi_e/\pi_b}^2(s,a)(r+\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a))^2].
\end{aligned}
$$

This concludes the proof. □

Proof of Theorem 22. Recall that $D_{v2}^{-1} D_{v3} = \hat{\beta}$ is seen as Z–estimator with a parametric model $v(s;\beta) = \beta^\top \phi(s)$. More specifically, the estimator $\hat{\beta}$ is given as a solution to

$$
\mathrm{E}_n\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r+v(s';\beta)-v(s;\beta)\}\phi(s)\right] = 0.
$$

Following the standard theory of Z–estimator (van der Vaart, 1998), the asymptotic variance of β is calculated as a sandwich estimator;

$$
\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} \mathcal{N}(0, D_1^{-1} D_2 (D_1^{-1})^\top),
$$

where $\beta_0^\top \phi(s) = V^{\pi_e}(s)$ and

$$
\begin{aligned}
D_1 &= \mathrm{E}_{d_{\pi_b}}[\phi(s)\{-\gamma\phi(s') + \phi(s)\}^\top] \\
D_2 &= \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + v(s';\beta) - v(s;\beta)\}\phi(s)\right]|_{\beta_0} \\
&= \mathrm{E}_{d_{\pi_b}}\left[\mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma V^{\pi_e}(s';\beta) - V^{\pi_e}(s)\}|s\right]\phi(s)\phi^\top(s)\right].
\end{aligned}
$$

Then, by delta method,

$$
\sqrt{n}(D_{v1}^\top D_{v2}^{-1} D_{v3} - R_{\pi_e}) \xrightarrow{d} \mathcal{N}(0, D_{v1}^\top D_1^{-1} D_2 (D_1^{-1})^\top D_{v1}).
$$

From now on, we simplify the expression $D_{v1}^\top S_1^{-1} S_2 (S_1^{-1})^\top D_{v1}$. First, we observe

$$
[D_{v1}]_i = (1-\gamma) d_0(S_i),
$$

where $[D_{v1}]_i$ is i–th element. In addition,

$$
\begin{aligned}
D_1^{-1} &= \mathrm{E}_{d_{\pi_b}}[\phi(s)\{-\gamma\phi(s') + \phi(s)\}^\top]^{-1} \\
&= \mathrm{E}_{d_{\pi_b}}[\phi(s)\phi^\top(s)\{-\gamma P^{\pi_e} + I\}^\top]^{-1} \\
&= (\{-\gamma P^{\pi_e} + I\}^\top)^{-1} \mathrm{E}_{d_{\pi_b}}[\phi(s)\phi^\top(s)]^{-1},
\end{aligned}
$$

where P^{π_e} is a transition matrix from the current state to the next state.

Therefore, by defining $g(s) = \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + V^{\pi_e}(s') - V^{\pi_e}(s)\}|s\right]$ and $\{-\gamma P^{\pi_e} + I\}^{-1} D_{v1} = D_3$, the asymptotic variance is

$$
\begin{aligned}
&D_3^\top \mathrm{E}_{d_{\pi_b}(s)}[\phi(s)\phi^\top(s)]^{-1} \mathrm{E}_{d_{\pi_b}(s)}[g(s)\phi(s)\phi^\top(s)] \mathrm{E}_{d_{\pi_b}(s)}[\phi(s)\phi^\top(s)]^{-1} D_3 \\
&= \sum_{\tilde{s}\in\mathcal{S}} d_{\pi_b(\tilde{s})}^{-1} g(\tilde{s}) \{D_3^\top I_{\tilde{s}}\}^2,
\end{aligned}
$$

where $I_{\tilde{s}}$ is $|S|$–dimensional vector, which the element corresponding $\tilde{s}$ is 1 and other elements are 0. Noting $D_3^\top I_{\tilde{s}} = d_{\pi_e,\gamma}(\tilde{s})$, the asymptotic variance is

$$
\sum_{\tilde{s}\in\mathcal{S}} d_{\pi_b(\tilde{s})}^{-1} g(\tilde{s}) d_{\pi_e}^2(\tilde{s}) = \mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma V^{\pi_e}(s') - V^{\pi_e}(s)\}|s\right]\right].
$$

Finally, we show this is larger than the semiparametric lower bound. This is seen as

$$
\begin{aligned}
&\mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma V^{\pi_e}(s') - V^{\pi_e}(s)\}|s\right]\right] \\
&\geq \mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{E}_{d_{\pi_b}}\left[\mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma V^{\pi_e}(s';\beta) - V^{\pi_e}(s)\}|s,a\right]\right]\right] \\
&= \mathrm{E}_{d_{\pi_b}}\left[w_{\pi_e/\pi_b}^2(s,a)\mathrm{var}[r + \gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)|s,a]\right].
\end{aligned}
$$

Here, from the first line to the second line, we use a general inequality $\mathrm{var}[x] = \mathrm{var}[\mathrm{E}[x|y]] + \mathrm{E}[\mathrm{var}[x|y]] \geq \mathrm{E}[\mathrm{var}[x|y]]$. □

Proof of Theorem 23. By refining $g(s) = \mathrm{var}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + V^{\pi_e}(s)\}|s\right]$ in the proof of Theorem 22, the asymptotic variance is

$$
\sum_{\tilde{s}\in\mathcal{S}} d_{\pi_b(\tilde{s})}^{-1} g(\tilde{s}) d_{\pi_e,\gamma}^2(\tilde{s}) = \mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r + \gamma V^{\pi_e}(s')\}|s\right]\right].
$$

Then, we show this is larger than the semiparamatric lower bound. This is seen as

$$\begin{aligned}\sum_{\tilde{s}\in\mathcal{S}} d^{-1}_{\pi_b(\tilde{s})} g(\tilde{s}) d^2_{\pi_e,\gamma}(\tilde{s}) &= \mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r+\gamma V^{\pi_e}(s')\}|s\right]\right] \\ &\geq \mathrm{E}_{d_{\pi_b}}\left[\left\{\frac{d_{\pi_e,\gamma}(s)}{d_{\pi_b}(s)}\right\}^2 \mathrm{E}_{\pi_b}\left[\mathrm{var}_{d_{\pi_b}}\left[\frac{\pi_e(a|s)}{\pi_b(a|s)}\{r+\gamma V^{\pi_e}(s')\}|s,a\right]\right]\right] \\ &= \mathrm{E}_{d_{\pi_b}}\left[w^2_{\pi_e/\pi_b}(s,a)\mathrm{var}[r+\gamma V^{\pi_e}(s') - Q^{\pi_e}(s,a)|s,a]\right]. \end{aligned}$$

□

E. Experiments in the Function Approximation Setting

In this section, we empirically evaluate our new algorithms MWL and MQL, and make comparison with MSWL (Liu et al., 2018) and DualDICE (Nachum et al., 2019b) in the function approximation setting.

E.1. Setup

We consider infinite-horizon discounted setting with $\gamma = 0.999$, and test all the algorithms on CartPole, a control task with continuous state space and discrete action space. Based on the implementation of OpenAI Gym (Brockman et al., 2016), we define a new state-action-dependent reward function and add small Gaussian noise with zero mean on the transition dynamics.

To obtain the behavior and the target policies, we first use the open source code[19] of DQN to get a near-optimal Q function, and then apply softmax on the Q value divided by an adjustable temperature τ:

$$\pi(a|s) \propto \exp(\frac{Q(s,a)}{\tau}) \tag{38}$$

We choose $\tau = 1.0$ as the behavior policy and $\tau = 0.25, 0.5, 1.5, 2.0$ as the target policies. The training datasets are generated by collecting trajectories of the behavior policy with fixed horizon length 1000. If the agent visits the terminal states within 1000 steps, the trajectory will be completed by repeating the last state and continuing to sample actions.

E.2. Algorithms Implementation

We use the loss functions and the OPE estimators derived in this paper and previous literature (Liu et al., 2018; Nachum et al., 2019b), except for MWL, in which we find another equivalent loss function can work better in practice:

$$L = \mathbb{E}_{d_{\pi_0}}[\gamma\Big(w(s,a)f(s',\pi) - w(s',a')f(s',a')\Big)] - (1-\gamma)\mathbb{E}_{d_0\times\pi_0}[w(s,a)f(s,a)] + (1-\gamma)\mathbb{E}_{d_0\times\pi}[f(s,a)] \tag{39}$$

In all algorithms, we keep the structure of the neural networks the same, which have two hidden layers with 32 units in each and ReLU as activation function. Besides, the observation is normalized to zero mean and unit variance, and the batch size is fixed to 500. In MSWL, the normalized states are the only input, while in the others, the states and the actions are concatenated and fed together into neural networks.

As for DualDICE, we conduct evaluation based on the open source implementation[20]. The learning rate of $\nu-$network and $\zeta-$network are changed to be $\eta_\nu = 0.0005$ and $\eta_\zeta = 0.005$, respectively, after a grid search in $\eta_\nu \times \eta_\zeta \in \{0.0001, 0.0005, 0.001, 0.0015, 0.002\} \times \{0.001, 0.005, 0.01, 0.015, 0.02\}$.

In MSWL, we use estimated policy distribution instead of the true value to compute the policy ratio. To do this, we train a 64x64 MLP with cross-entropy loss until convergence to approximate the distribution of the behavior policy. The learning rate is set to be 0.0005.

Besides, we implement MQL, MWL and MSWL with $\mathcal{F}$ corresponding to a RKHS associated with kernel $K(\cdot,\cdot)$. The new loss function of MWL (39) can be written as:

$$L = \gamma^2 \mathbb{E}_{s,a,s',\tilde{s},\tilde{a},\tilde{s}'\sim d_{\pi_0}}[w(s,a)w(\tilde{s},\tilde{a})\mathbb{E}_{a',\tilde{a}'\sim\pi}[K((s',a'),(\tilde{s}',\tilde{a}'))]]$$

[19] https://github.com/openai/baselines

[20] https://github.com/google-research/google-research/tree/master/dual_dice

$$
\begin{aligned}
&+\gamma^2\mathbb{E}_{s',a',\tilde{s}',\tilde{a}'\sim d_{\pi_0}}[w(s',a')w(\tilde{s}',\tilde{a}')K((s',a'),(\tilde{s}',\tilde{a}'))]\\
&+(1-\gamma)^2\mathbb{E}_{s,a,\tilde{s},\tilde{a}\sim d_0\times\pi_0}[w(s,a)w(\tilde{s},\tilde{a})K((s,a),(\tilde{s},\tilde{a}))]\\
&+(1-\gamma)^2\mathbb{E}_{s,a,\tilde{s},\tilde{a}\sim d_0\times\pi}[K((s,a),(\tilde{s},\tilde{a}))]\\
&-2\gamma^2\mathbb{E}_{s,a,s',\tilde{s}',\tilde{a}'\sim d_{\pi_0}}[w(s,a)w(\tilde{s}',\tilde{a}')\mathbb{E}_{a'\sim\pi}[K((s',a'),(\tilde{s}',\tilde{a}'))]]\\
&-2\gamma(1-\gamma)\mathbb{E}_{s,a,s'\sim d_{\pi_0},\tilde{s},\tilde{a}\sim d_0\times\pi_0}[w(s,a)w(\tilde{s},\tilde{a})\mathbb{E}_{a'\sim\pi}[K((s',a'),(\tilde{s},\tilde{a}))]]\\
&+2\gamma(1-\gamma)\mathbb{E}_{s,a,s'\sim d_{\pi_0},\tilde{s},\tilde{a}\sim d_0\times\pi}[w(s,a)\mathbb{E}_{a'\sim\pi}[K((s',a'),(\tilde{s},\tilde{a}))]]\\
&+2\gamma(1-\gamma)\mathbb{E}_{s',a'\sim d_{\pi_0},\tilde{s},\tilde{a}\sim d_0\times\pi_0}[w(s',a')w(\tilde{s},\tilde{a})K((s',a'),(\tilde{s},\tilde{a}))]\\
&-2\gamma(1-\gamma)\mathbb{E}_{s',a'\sim d_{\pi_0},\tilde{s},\tilde{a}\sim d_0\times\pi}[w(s',a')K((s',a'),(\tilde{s},\tilde{a}))]\\
&-2(1-\gamma)^2\mathbb{E}_{s,a\sim d_0\times\pi_0,\tilde{s},\tilde{a}\sim d_0\times\pi}[w(s,a)K((s,a),(\tilde{s},\tilde{a}))]. \qquad (40)
\end{aligned}
$$

We choose the RBF kernel, defined as

$$K(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{2\sigma^2}) \qquad (41)$$

where $\mathbf{x}_i,\mathbf{x}_j$ corresponds to state vectors in MSWL, and corresponds to the vectors concatenated by state and action in MWL and MQL. Denote h as the median of the pairwise distance between $\mathbf{x}_i$, we set σ equal to h, $\frac{h}{3}$ and $\frac{h}{15}$ in MSWL, MWL and MQL, respectively. The learning rates are fixed to 0.005 in these three methods.

In MSWL and MWL, to ensure that the predicted density ratio is non-negative, we apply $\log(1+\exp(\cdot))$ as the activation function in the last layer of the neural networks. Moreover, we normalize the ratio to have unit mean value in each batch, which works better.

E.3. Results

We generate N datasets with different random seeds. For each dataset, we run all these four algorithms from the beginning until convergence, and consider it as one trial. Every 100 training iterations, the estimation of value function is logged, and the average over the last five logged estimations will be recorded as the result in this trial. For each algorithm, we report the normalized MSE, defined by

$$\frac{1}{N}\sum_{i=1}^{N}\frac{(\hat{R}_{\pi_e}^{(i)}-R_{\pi_e})^2}{(R_{\pi_b}-R_{\pi_e})^2} \qquad (42)$$

where $\hat{R}_{\pi_e}^{(i)}$ is the estimated return in the i-th trial; R_{π_e} and R_{π_b} are the true expected returns of the target and the behavior policies, respectively, estimated by 500 on-policy Monte-Carlo trajectories (truncated at $H=10000$ steps to make sure γ^H is sufficiently small). This normalization is very informative, as a naïve baseline that treats $R_{\pi_e}\approx R_{\pi_b}$ will get 0.0 (after taking logarithm), so any method that beats this simple baseline should get a negative score. We conduct $N=25$ trials and plot the results in Figure 3.

E.4. Error bars

We denote $x_i=\frac{(\hat{R}_{\pi_e}^{(i)}-R_{\pi_e})^2}{(R_{\pi_b}-R_{\pi_e})^2}$. As we can see, the result we plot (Eq.(42)) is just the average of N i.i.d. random variables $\{x_i\}_{i=1}^n$. We plot twice the standard error of the estimation—which corresponds to 95% confidence intervals—under logarithmic transformation. That is, the upper bound of the error bar is

$$\log(\frac{1}{N}\sum_{i=1}^{N}x_i+\frac{2\sigma}{\sqrt{N}})$$

and the lower bound of the error bar is

$$\log(\frac{1}{N}\sum_{i=1}^{N}x_i-\frac{2\sigma}{\sqrt{N}})$$

where σ is the sample standard deviation of x_i.

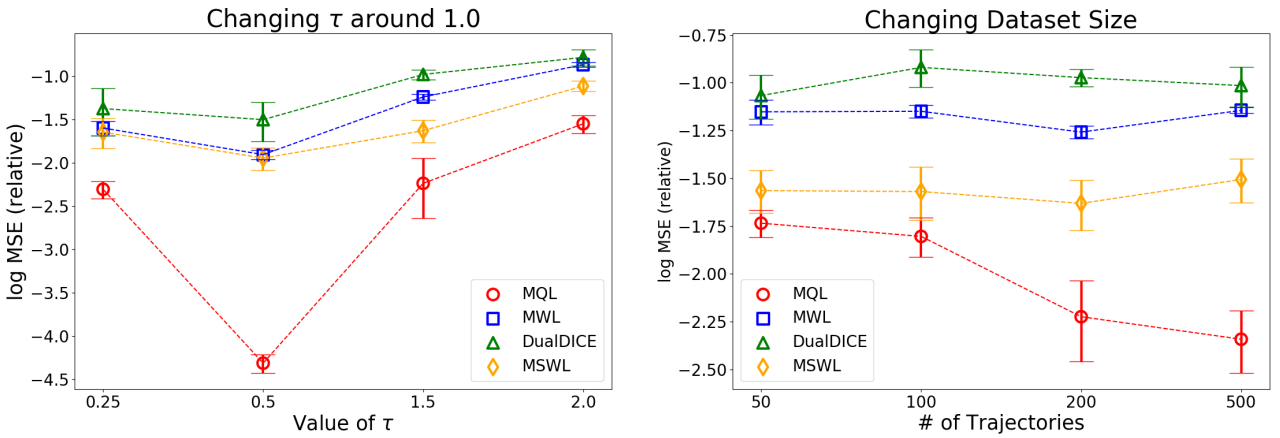


Figure 3. A Comparison of four algorithms: MQL (red circle), MWL (blue square), DualDICE (green triangle) and MSWL (yellow diamond). **Left**: We fix the number of trajectories to 200 and change the target policies. **Right**: We keep the target policy as $\tau = 1.5$, and vary the number of samples.

F. Step-wise IS as a Special Case of MWL

We show that step-wise IS (Precup et al., 2000) in discounted episodic problems can be viewed as a special case of MWL, and sketch the proof as follows. In addition to the setup in Section 2, we also assume that the MDP always goes to the absorbing state in H steps from any starting state drawn from d_0. The data are trajectories generated by π_b. We first convert the MDP into an equivalent *history-based* MDP, i.e., a new MDP where the state is the history of the original MDP (absorbing states are still treated specially). We use h_t to denote a history of length t, i.e., $h_t = (s_0, a_0, r_0, s_1, \dots, s_t)$. Since the history-based MDP still fits our framework, we can apply MWL as-is to the history-based MDP. In this case, each data trajectory will be converted into H tuples in the form of (h_t, a_t, r_t, h_{t+1}).

We choose the following $\mathcal{F}$ class for MWL, which is the space of *all* functions over histories (of various lengths up to H). Assuming all histories have non-zero density under π_e (this assumption can be removed), from Lemma 1 we know that the only w that satisfies $\forall f \in \mathcal{F},\ L_{\mathrm{w}}(w, f) = 0$ is

$$\frac{d_{\pi_e,\gamma}(h_t, a_t)}{d_{\pi_b}(h_t, a_t)} = \frac{(1-\gamma)\gamma^t}{1/H} \prod_{t'=0}^{t} \frac{\pi_e(a_{t'}|s_{t'})}{\pi_b(a_{t'}|s_{t'})}.$$ [21]

Note that $R_{\mathrm{w}}[w]$ with such an w is precisely the step-wise IS estimator in discounted episodic problems. Furthermore, the true marginalized importance weight in the original MDP $\frac{d_{\pi_e,\gamma}(s,a)}{d_{\pi_b}(s,a)}$ is not feasible under this "overly rich" history-dependent discriminator class (see also related discussions in Jiang (2019)).

[21] Here the term $\frac{(1-\gamma)\gamma^t}{1/H}$ appears because a state at time step t is discounted in the evaluation objective but its empirical frequency in the data is not. Other than that, the proof of this equation is precisely how one derives sequential IS, i.e., density ratio between histories is equal to the cumulative product of importance weights on actions.