A Additional Lemmas

Lemma 4. *Let assumption 1 be given, and define*

$$V = \left(\frac{1}{n} \sum_{i=1}^{n} f(X_i) \right)^2 ,$$

where $\mathbb{E}[f(X)] = 0$*, and* $\|f(X)\|_\infty < \infty$*. Then* $\mathbb{E}[V] = O(1/n)$*.*

Proof. Given the assumption that $\mathbb{E}[f(X)] = 0$, it is clear that

$$\mathbb{E}[V] = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{V}[f(X_i)] + \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \text{Cov}[f(X_i), f(X_j)].$$

Now by assumption $\|f(X)\|_\infty < \infty$, thus $\mathbb{V}[f(X)]$ is finite so the first term in the above expression is in $O(1/n)$. Thus it remains to bound the second term. Next we note that the X values are independent between trajectories, thus we can partition this term according to

$$\frac{2}{n^2} \sum_{t=1}^{N} \sum_{i=1}^{T_t} \sum_{j=1}^{i-1} \text{Cov}[f(X_i^{(t)}), f(X_j^{(t)})],$$

where $X_i^{(t)}$ denotes the i'th observation of the t'th trajectory. Therefore if we can show that the t'th term in the outer sum is in $O(T_t)$ we are done, so without loss of generality we consider the case of a single trajectory of length n and show that the corresponding sum of covariances is in $O(n)$.

Now let $\alpha(k)$ denote the kth α-mixing coefficient. Since $X_{1:n}$ is a Markov chain we have that $\alpha(X_i, X_j) = \alpha(|i-j|)$. In addition, given any random variables U and W, it follows from Rio (2013, 1.12b) that $\text{Cov}[X, Y] \leq 2\alpha(U, W)\|U\|_\infty \|W\|_\infty$. Applying this result to our setting we obtain

$$\begin{aligned} \text{Cov}[f(X_i), f(X_j)] &\leq 2\alpha(|i-j|)\|f(X)\|_\infty^2 \\ &\leq 4\beta(|i-j|)\|f(X)\|_\infty^2, \end{aligned}$$

where the second inequality follows from the fact that β-mixing coefficients are larger than α-mixing coefficients (up to a factor of 2). Thus we can obtain the bound

$$\begin{aligned} \sum_{i=1}^{n} \sum_{j=1}^{i-1} \text{Cov}[f(X_i), f(X_j)] &\leq 4\|f(X)\|_\infty^2 \sum_{i=1}^{n} \sum_{j=0}^{i-1} \beta(j) \\ &\leq 4n\|f(X)\|_\infty^2 \sum_{j=1}^{\infty} \beta(j) \\ &\leq 4n\|f(X)\|_\infty^2 \sum_{j=1}^{\infty} j^{2/(p-2)} \beta(j) \\ &\leq O(n), \end{aligned}$$

where $2 < p \leq \infty$ is the constant referenced in assumption 1, and the final inequality follows from assumption 1.

Thus we have $\sum_{i=1}^{n} \sum_{j=1}^{i-1} \text{Cov}[f(X_i), f(X_j)] = O(n)$, which lets us conclude that $\mathbb{E}[V] = O(1/n)$. □

Lemma 5. *Assume that $\mathcal{G}$ is p-balancing-regular. Then for every constant $M \geq 0$ we have*

$$\inf_{W} \sup_{g \in \mathcal{G}} J_\lambda(W, g) \leq \sup_{g \in \mathcal{G}} \inf_{\|W\| \leq M} B(W, g)^2 + \frac{\lambda}{n^2} M^2.$$

Proof of lemma 5. By assumption $\mathcal{G}$ is compact, and $g \mapsto J_\lambda(W, g)$ is continuous for every W. This means that by the Extreme Value theorem we can replace the supremum over $\mathcal{G}$ with a maximum over $\mathcal{G}$ in the quantity we are bounding.

Given this, we will proceed by bounding $\min_W \max_{g\in\mathcal{G}} B(W,\mu)$ using von Neumann's minimax theorem to swap the minimum and the maximum, and then use this to establish the overall bound for $J_\lambda(W,\mu)$.

First, we can observe that $B(W,g)$ is linear, and therefore both convex and concave, in each of W and g. Next, by assumption $\mathcal{G}$ is convex and compact, and as argued already $g \mapsto B(W,g)$ is continuous for every W. In addition, $B(W,g)$ is also clearly continuous in W for fixed g, and the set $\{W : \|W\| \leq M\}$ is obviously compact and convex for any non-negative M. Thus by von Neumann's minimax theorem we have the following for every $M \geq 0$:

$$\min_{\|W\|\leq M} \max_{g\in\mathcal{G}} B(W,g) = \max_{\mu\in\mathcal{G}} \min_{\|W\|\leq M} B(W,g) \tag{3}$$

Given this, we can bound $\min_W \max_{\mu\in\mathcal{F}} J(W,\mu)$ as follows, which is valid for any M:

$$\begin{aligned}
\min_W \max_{g\in\mathcal{G}} J_\lambda(W,g) &= \min_W \max_{g\in\mathcal{G}} B(W,g)^2 + \frac{\lambda}{n^2}\|W\|^2 \\
&\leq \min_{\|W\|\leq M} \max_{g\in\mathcal{G}} B(W,g)^2 + \frac{\lambda}{n^2}\|W\|^2 \\
&\leq \min_{\|W\|\leq M} \max_{g\in\mathcal{G}} B(W,g)^2 + \frac{\lambda}{n^2}M^2 \\
&= (\min_{\|W\|\leq M} \max_{g\in\mathcal{G}} |B(W,g)|)^2 + \frac{\lambda}{n^2}M^2 \\
&= (\max_{g\in\mathcal{G}} \min_{\|W\|\leq M} B(W,g))^2 + \frac{\lambda}{n^2}M^2 \\
&\leq (\max_{g\in\mathcal{G}} \min_{\|W\|\leq M} |B(W,g)|)^2 + \frac{\lambda}{n^2}M^2 \\
&= \max_{g\in\mathcal{G}} \min_{\|W\|\leq M} B(W,g)^2 + \frac{\lambda}{n^2}M^2
\end{aligned}$$

In these inequalities we use the fact that $\min_W \max_g B(W,g) = \min_W \max_g |B(W,g)|$, which follows because $B(W,g) = -B(W,-g)$, and that $g \in \mathcal{G} \iff -g \in \mathcal{G}$. In addition we use the fact that $x \mapsto x^2$ is a monotonic function on $\mathbb{R}^+$.

□

Lemma 6. *Let some $g \in \mathcal{G}$ be given. Then as long as there exists $i \in [n]$ such that $h_g(Z_i) \neq 0$, there exists $W \in \mathbb{R}^n$ satisfying*

$$B(W,g) = 0$$

and

$$\|W\|^2 = \frac{(\sum_{i=1}^n k_g(Z_i)h_g(Z_i))^2}{4\sum_{i=1}^n h_g(Z_i)^2}$$

.

Proof of lemma 6. We will prove this non-constructively by considering the value of the solution to the constrained optimization problem

$$\begin{aligned}
&\min_W \sum_{i=1}^n W_i^2 \\
&\text{s.t. } \frac{1}{n}\sum_{i=1}^n h_g(Z_i)W_i - k_g(Z_i) = 0
\end{aligned}$$

The Lagrangian corresponding to this problem is

$$\mathcal{L}(W;\lambda) = \sum_{i=1}^n W_i^2 + \lambda(h_g(Z_i)W_i - k_g(Z_i))$$

It can easily be verified by taking derivatives that for fixed λ this is minimized by setting $W_i = -\frac{1}{2}\lambda h_g(Z_i)$. Plugging in this W, we obtain the dual problem

$$\begin{aligned} D &= \max_{\lambda \in \mathbb{R}} \sum_{i=1}^n -\frac{1}{4} h_g(Z_i)^2 \lambda^2 + \frac{k_g(Z_i) h_g(Z_i)}{2} \lambda \\ &= \max_{\lambda \in \mathbb{R}} -\frac{1}{4} \left(\sum_{i=1}^n h_g(Z_i)^2 \right) \lambda^2 + \frac{1}{2} \left(\sum_{i=1}^n k_g(Z_i) h_g(Z_i) \right) \lambda \end{aligned}$$

Again taking derivatives, it is clear that this objective is maximized by

$$\lambda = \frac{\sum_{i=1}^n k_g(Z_i) h_g(Z_i)}{\sum_{i=1}^n h_g(Z_i)^2}.$$

Plugging in this solution we have that the maximum dual value objective is given by

$$D^* = \frac{(\sum_{i=1}^n k_g(Z_i) h_g(Z_i))^2}{4 \sum_{i=1}^n h_g(Z_i)^2}$$

Finally we note that the original constrained optimization problem had only linear equality constraints, and under the assumption that $h_g(Z_i) \neq 0$ for some i we can construct a feasible solution, so Slater's condition applies. Thus we can conclude that the minimum euclidean norm of W satisfying $B(W, g) = 0$ is given by D^*, and therefore a W satisfying our conditions must exist.

□

Lemma 7. *Let assumption 1 be given, and assume that $\mathcal{G}$ is p-balancing-regular. Then we have*

$$\sup_{g \in \mathcal{G}^*} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i)^2 - \mathbb{E}[h(Z)^2] \right| = o_p(1),$$

where

$$h(Z) = \mathbb{E}[g_A(S, U) \mid Z].$$

Proof of lemma 7. Let $2 < p \leq \infty$ be the fixed value of p from assumption 1, and let $N(\epsilon) = \max_{a \in [n]} N_{[]}(\epsilon, \mathcal{G}_a^*, \mathcal{L}_p)$. It follows easily from our assumptions that $\int_0^\infty \sqrt{\log N(\epsilon)} d\epsilon < \infty$.

Now, let $\mathcal{F} = \{f : f(z) = \mathbb{E}[g_A(S, U) \mid Z = z]\}$. Given any $f \in \mathcal{F}$ indexed by some $g = (g_1, \ldots, g_m) \in \mathcal{G}^*$, we let $(l_1, r_1), \ldots, (l_m, r_m)$ be ϵ/m-brackets for $g_1, \ldots, g_m$ respectively in $\mathcal{L}_p$. Now clearly by linearity (f_l, f_r) = $(\mathbb{E}[l_A(S, U) \mid Z = z], \mathbb{E}[r_A(S, U) \mid Z = z])$ is a bracket for f, and we have

$$\begin{aligned} \mathbb{E}[|f_l(Z) - f_r(Z)|^p]^{1/p} &= \mathbb{E}[|\mathbb{E}[r_A(S, U) - l_A(S, U) \mid Z]|^p]^{1/p} \\ &\leq \mathbb{E}[\mathbb{E}[|r_A(S, U) - l_A(S, U)|^p \mid Z]]^{1/p} \\ &= \mathbb{E}[|r_A(S, U) - l_A(S, U)|^p]^{1/p} \\ &\leq \mathbb{E}\left[\sum_{a=1}^m |r_a(S, U) - l_a(S, U)|^p \right]^{1/p} \\ &\leq \sum_{a=1}^m \mathbb{E}\left[|r_a(S, U) - l_a(S, U)|^p \right]^{1/p} \\ &\leq \sum_{a=1}^m \frac{\epsilon}{m} \\ &= \epsilon. \end{aligned}$$

Thus the $\mathcal{L}_p$-bracketing number for $\mathcal{F}$ must be at most $N(\epsilon/m)^m$, since we can ensure that every $f \in \mathcal{F}$ is in an ϵ-bracket by constructing ϵ/m-brackets for each class $\mathcal{G}_a^*$, and then contstructing an ϵ-bracket for $\mathcal{F}$ from each possible combinatorial choice of selecting one $\mathcal{G}_a^*$ bracket for each $a \in [m]$ and combining these.

Next, consider the function class $\mathcal{F}^2 = \{f : f(z) = \tilde{f}(z)^2, \tilde{f} \in \mathcal{F}\}$. Now, given a bracket (l, r) for $f \in \mathcal{F}$ we can construct a bracket (l_2, r_2) for the corresponding element f^2 of $\mathcal{F}^2$, where

$$\begin{aligned} l_2(z) &= \mathbb{1}\{\text{sign}(l(z)) = \text{sign}(r(z))\} \min(l(z)^2, r(z)^2) \\ r_2(z) &= \max(l(z)^2, r(z)^2). \end{aligned}$$

In the case that $\mathbb{1}\{\text{sign}(l(z)) = \text{sign}(r(z))\}$ we have $r_2(z) - l_2(z) = (r(z) - l(z))(r(z) + l(z)) \leq C(r(z) - l(z))$ for some constant C, which follows because uniformly bounded property of $\mathcal{G}$ implies that $\mathcal{F}$ must be uniformly bounded also. Also in the other case we have $r_2(z) - l_2(z) = r_2(z) \leq (r(z) - l(z))^2 \leq C(r(z) - l(z))$. Thus we have

$$\begin{aligned} \mathbb{E}[|r_2(Z) - l_2(Z)|^p]^{1/p} &\leq \mathbb{E}[C^p |r(Z) - l(Z)|^p]^{1/p} \\ &= C\mathbb{E}[|r(Z) - l(Z)|^p]^{1/p}. \end{aligned}$$

Thus any ϵ/C-bracketing of $\mathcal{F}$ gives a ϵ-bracketing of $\mathcal{F}^2$, so the $\mathcal{L}_p$-bracketing number of $\mathcal{F}^2$ must be at most $N(\epsilon/(mC))^m$. Therefore we have that the function class $\mathcal{F}^2$ satisfies

$$\begin{aligned} \int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}^2, \mathcal{L}_p)} d\epsilon &\leq \int_0^\infty \sqrt{m \log N(\epsilon/(mC))} d\epsilon \\ &= m^{3/2} C \int_0^\infty \sqrt{\log N(\alpha)} d\alpha \\ &< \infty. \end{aligned}$$

This finite uniform-entropy integral combined the β-mixing part of assumption 1 implies that the stochastic process over $\mathcal{F}^2$ defined by

$$G_n(f) = \sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n f(Z_i)^2 - \mathbb{E}[f(Z)^2] \right)$$

converges tightly to a limiting Gaussian process, by Kosorok (2007, Theorem 11.24). Thus the stochastic process $G_n/\sqrt{n}$ converges tightly to the zero random variable, meaning that $\sup_{f \in \mathcal{F}^2} G_n(f)/\sqrt{n} = o_p(1)$. Finally we can observe that by construction

$$\sup_{g \in \mathcal{G}^*} \left| \frac{1}{n} \sum_{i=1}^n h(Z_i)^2 - \mathbb{E}[h(Z)^2] \right| = \sup_{f \in \mathcal{F}^2} G_n(f)/\sqrt{n},$$

which gives us our final result.

□

B Omitted Proofs

Proof of theorem 1. We begin by providing a bound for the conditional MSE, $\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}]$. Define the sample average policy effect:

$$\text{SAPE}(\pi_e) = \frac{1}{n} \sum_{i=1}^n \sum_{a=1}^m d(S_i) \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i).$$

We note that following the derivation in section 4 we have $\mathbb{E}[\text{SAPE}(\pi_e)] = v(\pi_e)$. Given this and assumptions 1 to 3, it is clear that the conditions of lemma 4 apply to $\mathbb{E}_b[(\text{SAPE}(\pi_e) - v(\pi_e))^2]$, so this term must be $O(1/n)$. Thus by Markov's inequality and the law of total expectation we have $\mathbb{E}[(\text{SAPE}(\pi_e) - v(\pi_e))^2 \mid Z_{1:n}] = O_p(1/n)$. Then, using the fact that $(x + y)^2 \leq 2x^2 + 2y^2$, we have

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}] \leq 2\mathbb{E}[(\hat{\tau}_W - \text{SAPE}(\pi_e))^2 \mid Z_{1:n}] + O_p(1/n).$$

Next, we perform a bias variance decomposition of the RHS of this bound as follows:

$$\begin{aligned} \mathbb{E}[(\hat{\tau}_W - \text{SAPE}(\pi_e))^2 \mid Z_{1:n}] &= \mathbb{E}[\mathbb{E}[(\hat{\tau}_W - \text{SAPE}(\pi_e))^2 \mid Z_{1:n}, U_{1:n}] \mid Z_{1:n}] \\ &= \mathbb{E}[\mathbb{E}[\hat{\tau}_W - \text{SAPE}(\pi_e) \mid Z_{1:n}, U_{1:n}]^2 \mid Z_{1:n}] \\ &\quad + \mathbb{E}[\mathbb{V}[\hat{\tau}_W - \text{SAPE}(\pi_e) \mid Z_{1:n}, U_{1:n}] \mid Z_{1:n}] \\ &= \xi_1 + \xi_2, \end{aligned}$$

and we additionally define

$$\begin{aligned}\zeta_{ia} &= W_i \delta_{A_i a} R_i - d(S_i)\pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \\ \zeta_i &= \sum_{a=1}^m \zeta_{ia} = W_i R_i - d(S_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i).\end{aligned}$$

We note that our MDPUC structure implies that R_i and U_i are conditionally independent of all other states, actions, rewards, and confounders given Z_i, and therefore that $\mathbb{E}[\zeta_{ia} \mid Z_{1:n}] = \mathbb{E}[f_{ia}\mu_a(S_i, U_i) \mid Z_i]$. Given this, the first term of the above bias variance decomposition can be broken down as:

$$\begin{aligned}\xi_1 &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia}\right)^2 \;\middle|\; Z_{1:n}\right] \\ &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia} \;\middle|\; Z_{1:n}\right]^2 + \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia} \;\middle|\; Z_{1:n}\right] \\ &= \left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \mathbb{E}[f_{ia}\mu_a(S_i, U_i) \mid Z_i]\right)^2 + \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \zeta_i \mid Z_{1:n}\right] \\ &\leq \left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \mathbb{E}[f_{ia}\mu_a(S_i, U_i) \mid Z_i]\right)^2 + \frac{2\sigma^2}{n^2}\sum_{i=1}^n W_i^2 \\ &\quad + 2\mathbb{V}\left[\frac{1}{n}\sum_{1=1}^n d(S_i)\sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \mid Z_{1:n}\right] \\ &= B(W, \mu)^2 + \frac{2\sigma^2}{n^2}\|W\|^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{1=1}^n d(S_i)\sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \mid Z_{1:n}\right],\end{aligned}$$

where the inequality step follows from assumption 3 and the identity $(x+y)^2 \leq 2x^2 + 2y^2$. Similarly, we bound the the second error term ξ_2 as:

$$\begin{aligned}\xi_2 &= \mathbb{E}\left[\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \zeta_i \;\middle|\; Z_{1:n}, U_{1:n}\right] \;\middle|\; Z_{1:n}\right] \\ &\leq \mathbb{E}\left[\frac{2\sigma^2}{n^2}\sum_{i=1}^n W_i^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i)\sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \;\middle|\; Z_{1:n}, U_{1:n}\right] \;\middle|\; Z_{1:n}\right] \\ &\leq \frac{2\sigma^2}{n^2}\|W\|^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{1=1}^n d(S_i)\sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \;\middle|\; Z_{1:n}\right],\end{aligned}$$

where in the first inequality step follows again from assumption 3 and the identity $(x+y)^2 \leq 2x^2 + 2y^2$, and the second inequality step follows from the law of total variance.

Next, by assumptions 1 to 3, it follows from lemma 4 that

$$\mathbb{V}\left[\frac{1}{n}\sum_{1=1}^n d(S_i)\sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i)\right] = O(1/n),$$

and therefore it follows from Markov's inequality that the conditional variance version is $O_p(1/n)$.

Next, putting the above bounds together we get

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n})] \leq 2\left(B(W, \mu)^2 + \frac{4\sigma^2}{n^2}\|W\|^2\right) + O_p(1/n).$$

It follows from this that if $\lambda \geq 4\sigma^2$ and $J_\lambda(W, \mu) = O_p(r_n)$, then $\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}] = O_p(\max(1/n, r_n))$. Finally, it follows from Kallus (2016, Lemma 31) that $(\hat{\tau}_W - v(\pi_e))^2 = O_p(\max(1/n, r_n))$, and thus $\hat{\tau}_W = v(\pi_e) + O_p(\max(n^{-1/2}, r_n^{1/2}))$.

□

Proof of theorem 2. We first note that by lemma 5 we have for every $M \geq 0$:

$$\inf_{W \in \mathbb{R}^n} \sup_{g \in \mathcal{G}} J_\lambda(W, g) \leq \sup_{g \in \mathcal{G}} \inf_{\|W\| \leq M} B(W, g)^2 + \frac{\lambda}{n^2} M^2.$$

Therefore it is sufficient to ensure that, for each $g \in \mathcal{G}$, that we can find $W(g)$ in response such that $B(W(g), g) = 0$ and $\sup_{g \in \mathcal{G}} \|W(g)\|^2 = O_p(n)$. In the case that $\|g\| = 0$ we trivially have $B(0, g) = 0$, so we can restrict our attention to non-zero g.

Next, define

$$\begin{aligned} h_g(z) &= \mathbb{E}[g_A(S, U) \mid Z = z] \\ k_g(z) &= \mathbb{E}[d(S) \sum_a \pi_e(a \mid S, U) g_a(S, U) \mid Z = z] \,. \end{aligned}$$

Given the decomposition $B(W, g) = \frac{1}{n} \sum_{i=1}^n W_i h_g(Z_i) - k_g(Z_i)$, as long as $h(Z_i) \neq 0$ for some $i \in [n]$ it follows from lemma 6 that we can find $W(g)$ satisfying

$$\begin{aligned} B(W(g), g) &= 0 \\ \|W(g)\|^2 &= \frac{(\sum_{i=1}^n h(Z_i) k(Z_i))^2}{4 \sum_{i=1}^n h(Z_i)^2} \\ &= n \frac{(\frac{1}{n} \sum_{i=1}^n h(Z_i) k(Z_i))^2}{4 \frac{1}{n} \sum_{i=1}^n h(Z_i)^2} \\ &= n \frac{(\frac{1}{n} \sum_{i=1}^n h(Z_i) k(Z_i))^2}{4(\mathbb{E}[h(Z)^2] + (\frac{1}{n} \sum_{i=1}^n h(Z_i)^2 - \mathbb{E}[h(Z)^2]))}. \end{aligned}$$

We note that this equation clearly satisfies $\|W(g)\|^2 = \|W(\lambda g)\|^2$ for any $\|g\| \neq 0$ and $\lambda > 0$. Thus it follows that $\sup_{g \in \mathcal{G}} \|W(g)\|^2 = \sup_{g \in \mathcal{G}^*} \|W(g)\|$. Furthermore, by assumption 4 we have that $P(h_g(Z) > 0)$ for every $g \in \mathcal{G}^*$, and thus $\mathbb{E}[h_g(Z)^2] > 0$. Now let $\alpha = \inf_{g \in \mathcal{G}^*} \mathbb{E}[h(Z)^2]$. Given the compactness of continuity properties of $\mathcal{G}$ the extreme value theorem applies and we have $\alpha > 0$. Next, it follows easily from the uniform boundedness of $\mathcal{G}$ that $(\frac{1}{n} \sum_{i=1}^n h(Z_i) k(Z_i))^2 \leq \beta(\frac{1}{n} \sum_{i=1}^n d(S_i))$ for some $0 < \beta < \infty$. Furthermore assumption 1 gives us that $d(S_i)$ is stationary, it follows from the Markov chain law of large numbers that $\beta(\frac{1}{n} \sum_{i=1}^n d(S_i)) = O_p(1)$. Thus for any $g \in \mathcal{G}^*$ we have

$$\|W(g)\|^2 \leq \left(\frac{n}{4}\right) \frac{O_p(1)}{\alpha + \epsilon(g) + \frac{1}{n} \sum_{i=1}^n h(Z_i)^2 - \mathbb{E}[h(Z)^2]},$$

where $\alpha > 0$, and $\epsilon(g) \geq 0$. Next, lemma 7 implies that the stochastic equicontinuity term $(1/n) \sum_{i=1}^n h(Z_i)^2 - \mathbb{E}[h(Z)^2]$ converges in probability to 0 uniformly over $\mathcal{G}$. Thus by the continuous mapping theorem we have that the RHS of the previous bound converges in probability to $O_p(n)/(\alpha + \epsilon(g)) \leq O_p(n)$ uniformly over $g \in \mathcal{G}^*$.

Now, recall that this bound was valid in the event that at least one $h(Z_i)$ is non-zero, which must occur with probability $1 - \delta_n(g)$, where $\delta_n(g) = O_p(p(g)^{-n})$, and $p(g) = P(h(Z) = 0)$. Furthermore, given assumption 4 and applying the extreme value theorem as above, we have $\sup_{g \in \mathcal{G}^*} \delta_n(g) = O_p(p^{-n})$, for some $p < 1$. In the event that for some g every $h(Z_i)$ is zero, we can instead choose $W(g) = 0$, giving a bound of $J_\lambda(W(g), g) \leq (\sum_{i=1}^n k(Z_i))^2 = O_p(1)$ uniformly over $g \in \mathcal{G}^*$, since $\frac{1}{n} \sum_{i=1}^n k(Z_i)$ can be bounded uniformly over $\mathcal{G}^*$ by applying assumption 2 and the uniform boundedness of $\mathcal{G}$.

Therefore, we can conclude by putting the above bounds together, which gives us

$$\inf_{W \in \mathbb{R}^n} \sup_{g \in \mathcal{G}} J_\lambda(W, g) \leq (1 - O_p(p^{-n})) O_p(1/n) + O_p(p^{-n}) O_p(1) = O_p(1/n).$$

□

Proof of lemma 2. Recall that for this lemma we have made the assumption that π_e is measurable with respect to S only. That is, $\pi_e(a \mid s, u) = \pi_e(a \mid s)$. Let b be some constant such that $g_a(s,u) \leq b$ for every $g \in \mathcal{G}$, $a \in [m]$, $s \in \mathcal{S}$, and $u \in \mathcal{U}$, and let c be some constant such that $d(s) \leq c$ for every $s \in \mathcal{S}$. We note that both these constants must exist given assumption 2 and the uniform boundedness of $\mathcal{G}$. In addition we define the estimated versions of the quantities in our analysis as follows.

$$\begin{aligned}
\hat{f}_{ia} &= W_i \delta A_i a - \hat{d}(S_i)\pi_e(a \mid S_i) \\
\hat{B}(W,g) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} \hat{f}_{ia}\mathbb{E}_{\hat{\varphi}}[g_a(S_i,U_i) \mid Z_i] \\
\hat{J}_\lambda(W,g) &= \hat{B}(W,g)^2 + \frac{\lambda}{n^2}\|W\|^2.
\end{aligned}$$

Given this, for any W measurable in $Z_{1:n}$, we can obtain the bound

$$\begin{aligned}
&|\sup_{g\in\mathcal{G}} B(W,g) - \sup_{g\in\mathcal{G}} \hat{B}(W,g)| \\
&\quad \leq \sup_{g\in\mathcal{G}} |B(W,g) - \hat{B}(W,g)| \\
&\quad \leq \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} f_{ia}(\mathbb{E} - \hat{\mathbb{E}})[g_a(S_i,U_i) \mid Z_i] \right| \\
&\qquad + \sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} (d(S_i) - \hat{d}(S_i))\hat{\mathbb{E}}\left[\sum_{a=1}^{m} \pi_e(a \mid S_i) g_a(S_i,U_i) \;\middle|\; Z_i \right] \right| \\
&\quad \leq \frac{F}{n}\sum_{i=1}^{n}\sum_{a=1}^{m} |f_{ia}| D_{\mathcal{F}}(\varphi(Z_i),\hat{\varphi}(Z_i)) + \frac{b}{n}\sum_{i=1}^{n} |d(S_i) - \hat{d}(S_i)| \\
&\quad \leq \frac{F}{n}\sum_{i=1}^{n} (|W_i| + c) D_{\mathcal{F}}(\varphi(Z_i),\hat{\varphi}(Z_i)) + \frac{b}{n}\sum_{i=1}^{n} |d(S_i) - \hat{d}(S_i)| \\
&\quad \leq \frac{cF}{n}\sum_{i=1}^{n} D_{\mathcal{F}}(\varphi(Z_i),\hat{\varphi}(Z_i)) + \frac{F\|W\|}{\sqrt{n}} \left(\frac{1}{n}\sum_{i=1}^{n} D_{\mathcal{F}}(\varphi(Z_i),\hat{\varphi}(Z_i))^2 \right)^{1/2} \\
&\qquad + \frac{b}{n}\sum_{i=1}^{n} |d(S_i) - \hat{d}(S_i)| \\
&\quad \leq O_p(r_n) + \frac{\|W\|}{\sqrt{n}} O_p(r_n) + O_p(r_n),
\end{aligned}$$

where in the second last inequality we apply Cauchy Schwartz, and in the final inequality we apply the assumptions that $D_{\mathcal{F}}(\varphi(Z_i),\hat{\varphi}(Z_i)) = O_p(r_n)$ and $|d(S_i) - \hat{d}(S_i)| = O_p(r_n)$ for every $i \in [n]$. Now, let $\tilde{W} = \arg\min_W \sup_{g\in\mathcal{G}} J_\lambda(W,g)$. It easily follows from theorem 2 that $\|\tilde{W}\| = O_p(\sqrt{n})$, so from the above we have $\sup_{g\in\mathcal{G}} \hat{B}(\tilde{W},g) = \sup_{g\in\mathcal{G}} B(\tilde{W},g) + O_p(r_n)$. In addition it also follows from theorem 2 that $\sup_{g\in\mathcal{G}} B(\tilde{W},g) = O_p(1/\sqrt{n})$. Putting all of the above together we get $\sup_{g\in\mathcal{G}} \hat{J}_\lambda(\tilde{W},g) = O_p(\max(1/n, r_n^2))$, and therefore $\sup_{g\in\mathcal{G}} \hat{J}_\lambda(W^*,g) = O_p(\max(1/n, r_n^2))$. Given this, it follows that $\|W^*\| = O_p(\max(\sqrt{n}, n r_n))$, and therefore applying the bound above again we get

$$\begin{aligned}
&\sup_{g\in\mathcal{G}} B(W^*,g) = O_p(\max(\sqrt{n} r_n^2, r_n)) \\
&\implies \sup_{g\in\mathcal{G}} J_\lambda(W^*,g) = O_p(\max(1/n, r_n^2, n r_n^4)) = O_p(\max(1/n, n r_n^4)),
\end{aligned}$$

where the final equality follows since it is always the case that either $1 \leq n r_n^2 \leq n^2 r_n^4$, or $1 \geq n r_n^2 \geq n^2 r_n^4$ (depending on whether $n r_n^2 \geq 1$ or not). It immediately follows that $J_\lambda(W^*, \mu) = O_p(\max(1/n, n r_n^4))$. Therefore plugging in $\lambda = 4\sigma^2$, the required result immediately follows by applying theorem 1.

□

Proof of theorem 4. First, following exactly the same argument as in the proof of lemma 2, we can obtain the bound

$$\sup_{g \in \mathcal{G}} J_\lambda(W^*, g) = O_p(\max(1/n, nr_n^4)).$$

We note that none of the arguments or theorems used in the derivation of the above bound, or theorem 2 which is used in the argument, depend on the assumption that the confounders are independent, and therefore this bound still holds in the case that $U_{1:n}$ are distributed according to a Markov chain.

Next, we define the following terms similar to those in our core theory, recalling that for this lemma we have assumed that pi_e is measurable with respect to the observed state only (that is, $\pi_e(s, u) = \pi_e(s)$).

$$f_{ia}^* = W_i^* \delta_{A_i a} - d(S_i, U_i) \pi_e(a \mid S_i)$$
$$B^*(W^*, g) = \frac{1}{n} \sum_{i=1}^n \sum_{a=1}^m f_{ia}^* \mathbb{E}[g_a(S_i, U_i) \mid Z_{1:n}].$$

In addition, we define the error term

$$\epsilon(W^*, \mu) = |B^*(W^*, \mu)^2 - B(W^*, \mu)^2|^{1/2}.$$

Then it follows from lemma 8 (described and proved in appendix C) that

$$\mathbb{E}[(\hat{\tau}_{W^*} - v(\pi_e))^2 \mid Z_{1:n}] \leq 2 J_{4\sigma^2}(W^*, \mu) + \epsilon(W^*, \mu)^2.$$

Given this, the assumption that $\lambda \geq 4\sigma^2$, and the bound $\sup_{g \in \mathcal{G}} J_\lambda(W^*, g) = O_p(\max(1/n, nr_n^4))$, it follows from Kallus (2016, Lemma 31) that

$$(\hat{\tau}_{W^*} - v(\pi_e))^2 = \epsilon(W^*, \mu)^2 + O_p(\max(1/n, nr_n^4))$$

That is, by bounding $\epsilon(W^*, \mu)$ we can bound the irreducible MSE from our balanced policy evaluation in the non-iid setting.

Next, we let b be a constant such that $|\mu_a(s, u)| \leq b \; \forall a, s, u$ (which must exist given assumption 3). Given our assumption that π_e is measurable with respect to S, it follows from the conditions of this lemma that the assumptions of lemma 9 are satisfied (described and proved in appendix C). Then, applying the fact that $J_\lambda(W^*, g) = O_p(\max(1/n, nr_n^4))$ implies that $|B(W^*, \mu)| = O_p(\max(n^{-1/2}, n^{1/2} r_n^2))$, as well as Cauchy Schwartz and the inequality $(x + y)^2 \leq 2x^2 + 2y^2$, this lemma gives us

$$\epsilon(W^*, \mu) \leq Fc \left(\frac{1}{n} \sum_{i=1}^n D_{\mathcal{F}}(\varphi_{1:n}, \varphi_i)^2 \right)^{1/2} + b \| d(S, U) - d(S) \|_2 + O_p(\max(n^{-1/2}, n^{1/2} r_n^2)),$$

where $c = \sqrt{2}(\|W^*\|^2/n + 1)^{1/2}$, which gives us our final result.

□

Proof of Lemma 3. First, note that for any $\epsilon' > 0$, then for sufficiently large i we have $J_\lambda(W, \mu) \leq \sup_{g \in \mathcal{G}_i} J_\lambda(W, g) + \epsilon' \|W\|_1^2/n^2$ for all W, which follows from the definition of universal approximation and the fact that by assumption d and g are universally bounded. This implies that $J_\lambda(W^*, \mu) \leq \inf_W \sup_{g \in \mathcal{G}_i} J_\lambda(W) + \epsilon' \|W^*\|_1^2/n^2$. In addition, under our assumptions we have $\min_W \sup_{g \in \mathcal{G}_i} J_\lambda(W, g) \to 0$ in probability for each class $\mathcal{G}_i$.

The above observations immediately suggest that we can obtain our required result by being careful of choosing the sequence i_n. Specifically, let some arbitrary sequence $\{\epsilon_i\}$ be given, such that $\epsilon_i \to 0$. Then first, we choose some non-decreasing sequence j_i such that the error term $\epsilon' \|W^*\|_1^2/n^2$ above is bounded by ϵ_i when using $\mathcal{G}_{j_i}$. This can be ensured given that $\|W^*\|_1^2/n^2$ is well behaved. Next, choose a second non-decreasing sequence n_i such that $\inf_W \sup_{g \in \mathcal{G}_{j_i}} J_\lambda^{(n_i)}(W, g) \to 0$ in probability as $i \to \infty$, where $J_\lambda^{(n_i)}$ is the adversarial objective for $n = n_i$. Note that this can be ensured given that $\inf_W \sup_{g \in \mathcal{G}_i} J_\lambda(W, g) \to 0$ for each $\mathcal{G}_i$. Finally, we can choose i_n to be any non-decreasing sequence such that $i_{n_i} = j_{n_i}$ for each i, from which it immediately follows that $J_\lambda(W^*, \mu) \to 0$ in probability, where $W^* = \arg\min_W \sup_{g \in \mathcal{G}_{i_n}} J_\lambda(W, g)$.

□

Proof of theorem 5. We first note that $\mathbb{E}[d(S)] = 1$ follows trivially for any stationary density ratio, by the definition of d and the fact that all probability measures have total measure 1.

Next, let U' denote the successor of U (analogously to S'), and let f_b and f_e refer to measures (or conditional measures) with respect to the stationary distributions of π_b and π_e. Then we have

$$\begin{aligned}
\mathbb{E}_b[d(S,U,A) \mid S', U'] &= \int d(s,u,a) f_b(s,u,a \mid S', U') ds\ du\ da \\
&= \int \frac{f_e(s,u,a)}{f_b(s,u,a)} f_b(s,u,a \mid S', U') ds\ du\ da \\
&= \int \frac{f_e(s,u,a)}{f_b(s,u,a)} \frac{f_b(S',U' \mid s,u,a) f_b(s,u,a)}{f_b(S',U')} ds\ du\ da \\
&= \int \frac{f_e(s,u,a) f_e(S',U' \mid s,u,a)}{f_b(S',U')} ds\ du\ da \\
&= d(S',U').
\end{aligned}$$

In the second last step we appeal to the fact that the conditional density of S', U' given S, U, A is the same under both π_b and π_e given our MDPUC assumptions. Note also that the fractions in the above derivation should be interpreted as Radon–Nikodym derivatives where appropriate, in the case that the random variables are not continuous.

Next, we note that $d(S,U,A) = d(S,U)\pi_e(A \mid S,U)/\pi_b(A \mid S,U)$, and that by our MDPUC assumtions we have that $d(S,U) = d(S)$. Therefore we have

$$\mathbb{E}_b\left[d(S) \frac{\pi_e(A \mid S,U)}{\pi_b(A \mid S,U)} - d(S') \;\middle|\; S', U' \right] = 0.$$

Next we note that from our MDPUC indepdence assumptions (S, A, U) are indepdendent of U' given S', so we can marginalize over U' and obtain

$$\mathbb{E}_b\left[d(S) \frac{\pi_e(A \mid S,U)}{\pi_b(A \mid S,U)} - d(S') \;\middle|\; S' \right] = 0.$$

Finally we can iterate expectations on Z to obtain

$$\mathbb{E}_b\left[d(S)\beta(Z) - d(S') \;\middle|\; S' \right] = 0.$$

Now we have established that the true stationary density ratio must satisfy the regular and conditional moment conditions described in theorem 5. For the reverse result, we note first that assumption 1 implies that the stationary distribution of our Markov chain is unique. Now as argued in Liu et al. (2018), it is clear given ergodicity that any d satisfying this conditional moment restriction must correspond to a scalar multiple of the true stationary density ratio, since the construction of the conditional moment restriction is exactly identical to that of Liu et al. (2018) if we consider (S, U) to be the state. Thus the additional restriction that $\mathbb{E}[d(S)] = 1$ ensures that any d satisfying both moment conditions must the true stationary density ratio.

□

Proof of lemma 1. First we observe that by construction $\mathcal{G}_K = \mathcal{G}_K^*$, so we will only discuss the former. Define

$$B_s = \sup_{s \in \mathcal{S}, u \in \mathcal{U}} \sqrt{K((s,u),(s,u))}.$$

By our bounded kernel assumption we have that $0 < B_s < \infty$. Now, for any $g \in \mathcal{G}_a^*$ we have $g((s,u)) = \langle g, K_{s,u} \rangle \leq B_s \|g\|$, where $K_{s,u}$ denotes the reproducing element for evaluation at s, u. Thus $\|g\|_\infty \leq B_2 \|g\|$, which gives us that $\mathcal{G}$ is uniformly bounded.

Next, from Cucker & Smale (2002, Theorem D) we have that the covering number under the $\mathcal{L}_\infty$-norm of an RKHS ball of unit radius with bounded, continuous kernel is given by

$$\sqrt{\log N(\epsilon, \mathcal{G}_a^*, \mathcal{L}_\infty)} \leq (C_b/\epsilon)^b,$$

for some constant $C_b > 0$ depending only on b and any $0 < b < 1$. Thus it is easy to argue by constructing separate finite covering sets for each $a \in [m]$ that we satisfy the compactness condition.

In order to deal with the bracketing entropy condition, we note that an $\mathcal{L}_\infty$ covering number bound gives a corresponding $\mathcal{L}_\infty$ bracketing number bound, given our uniformly bounded condition. Concretely, given any $g \in \mathcal{G}_a^*$, we let g' be a function such that $\|g - g'\|_\infty < \epsilon$. This implies that the bracket $(g' - \epsilon, g' + \epsilon)$ is a valid bracket for g. Therefore $N_{[]}(\epsilon, \mathcal{G}_a^*, \mathcal{L}_\infty) \leq N(\epsilon, \mathcal{G}_a^*, \mathcal{L}_\infty)$. Thus we have that:

$$\sqrt{\log N_{[]}(\epsilon, \mathcal{G}_a^*, \mathcal{L}_\infty)} \leq (C_b/\epsilon)^b,$$

which is sufficient to ensure the bracketing entropy condition, since $\int_0^C (1/\epsilon)^b d\epsilon < \infty$ for any $0 < C < \infty$ when $0 < b < 1$, and from uniform boundedness we have that $\sqrt{\log N_{[]}(\epsilon, \mathcal{G}_a^*, \mathcal{L}_\infty)} = 0$ when $\epsilon \geq B_s$.

Finally, we note that the symmetry and convexity properties are trivial from the definition of $\mathcal{G}_K$, as is the continuity condition since RKHSs are continuous with respect to function application.

□

Proof of Theorem 6. First, by the representer theorem we note that there must exist solutions for the interior and exterior optimization problems respectively, given by

$$\begin{aligned} h(s) &= \sum_{i=1}^{n_s} a_i K_H(s_i, s) \\ d(s) &= \sum_{i=1}^{n_s} b_i K_D(s_i, s)\,. \end{aligned}$$

Therefore, we can re-frame the optimization problem as min-max problem over the vectors a and b, where we define $a_{n_s+1} = c$. Consider the interior optimization problem first. Plugging in the above equation for h and c, we have

$$\begin{aligned} U_n(d, \tilde{d}, h, c) &= \sum_{i=1}^{n_s} a_i \frac{1}{n} \sum_{j=1}^{n} \big(K_H(s_i, S_j') d(S_j) \beta(Z_j) + a_{n_s+1}(d((S_j) - 1) \big) \\ &\quad - \frac{1}{4n} \sum_{j=1}^{n} \left(\sum_{i=1}^{n_s} a_i K_H(s_i, S_j') \tilde{d}(S_j) \beta(Z_j) + a_{n_s+1}(\tilde{d}(S_j) - 1) \right)^2 \\ &= g^T a - \frac{1}{4} a^T G a\,, \end{aligned}$$

where

$$\begin{aligned} g_i &= \frac{1}{n} \sum_{j=1}^{n} K_H(s_i, S_j') d(S_j) \beta(Z_j) \quad \forall i \in [n_s] \\ g_{n_s+1} &= \frac{1}{n} \sum_{j=1}^{n} d(S_j) - 1 \\ G &= \frac{1}{n} \sum_{j=1}^{n} q_j q_j^T\,, \end{aligned}$$

and q_j is defined as in the theorem statement for each $j \in [n]$. Now, assuming we enforce the norm constraints on h and c by Lagrangian regularization using hyperparameters λ_h and λ_c, the interior optimization problem is given by

$$\sup_{h,c} U_n(d, \tilde{d}, h, c) = \sup_a g^T a - \frac{1}{4} a^T Q a\,,$$

where Q is defined as in the theorem statement. Now, it easily follows by taking derivatives that this objective is maximized by $a = 2Q^{-1}g$, so it follows that

$$\sup_{h,c} U_n(d, \tilde{d}, h, c) = g^T Q^{-1} g\,.$$

Next, we consider the exterior optimization problem in b. We note that Q is constant in b, so therefore we only need to consider the effect of b on g. Given the definition $d(s) = \sum_{i=1}^{n_s} b_i K_D(s_i, s)$, we can obtain

$$g_i = \sum_{k=1}^{n_s} b_k \phi_{i,k} + \phi_{i,0} \quad \forall i \in [n_s + 1] \,,$$

where

$$\begin{aligned} \phi_{i,k} &= \frac{1}{n} \sum_{j=1}^{n} K_H(s_i, S'_j) K_D(s_k, S_j) \beta(Z_j) \quad \forall i \in [n_s] \\ \phi_{n_s+1,k} &= \frac{1}{n} \sum_{j=1}^{n} K_D(s_k, S_j) \\ \phi_{i,0} &= 0 \quad \forall i \in [n_s] \\ \phi_{n_s+1,0} &= -1 \,. \end{aligned}$$

Plugging in these to the above, we have

$$\sup_{h,c} U_n(d, \tilde{d}, h, c) = b^T \Omega^T Q^{-1} \Omega b - 2(\Omega^T Q^{-1} \omega)^T b + \omega^T Q^{-1} \omega \,,$$

where $\Omega_{i,j} = \phi_{i,k} \; \forall \; i \in [n_s + 1], j \in [n_s]$, and $\omega_i = -\phi_{i,0} \; \forall \; i \in [n_s + 1]$. We note that these definitions of Ω and ω exactly match those in the theorem statement. Assuming that we enforce the constraint on d by Lagrangian regularization using hyperparameter λ_d, the exterior optimization problem then is given by

$$\inf_d \sup_{h,c} U_n(d, \tilde{d}, h, c) = \inf_b b^T (\Omega^T Q^{-1} \Omega + \lambda_d k_D) b - 2(\Omega^T Q^{-1} \omega)^T b + \omega^T Q^{-1} \omega \,.$$

Finally, by taking derivatives again, it follows that the b minimizing the RHS above is given by

$$b = (\Omega^T Q^{-1} \Omega + \lambda_d k_D)^{-1} \Omega^T Q^{-1} \omega \,,$$

as required.

□

Proof of theorem 7. First we will find a closed form expression for $\sup_{g \in \mathcal{G}_K} (\frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} \mathbb{E}[f_{ia} g_(S_i, U_i) \mid Z_i])^2$, In this derivation we will use the shorthand φ_i for the conditional density of U_i given Z_i, and T_K for the kernel intergral operator defined according to $T_K f = \int_{\mathcal{Z}} K(\cdot, z) f(z) dz$. In this derivation we will make use of the fact that $\langle f, g \rangle_2 = \langle f, g \rangle_K$ for any square integrable f and g, where these inner products refer to $\mathcal{L}_2$ and the RKHS $\mathcal{H}_K$ respectively. Note that in this derivation we calculate $\mathcal{L}_2$ inner products with respect to the Borel measure $\mathbb{R}$, rather than the measure from the stationary distribution of π_b, which allows us to write conditional expectations as explicit inner products. Given all this we can obtain:

$$
\begin{aligned}
&\sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} \mathbb{E}\left[f_{ia} g_a(S_i, U_i) \;\middle|\; Z_i \right] \right)^2 \\
&= \sum_{a=1}^{m} \sup_{g \in \mathcal{G}} \left(\frac{1}{n} \sum_{i=1}^{n} \langle g_a, \varphi_i f_{ia} \rangle_2^2 \right)^2 \\
&= \sum_{a=1}^{m} \sup_{g \in \mathcal{G}} \left(\langle g_a, T_K \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia} \rangle_K^2 \right)^2 \\
&= \sum_{a=1}^{m} \frac{\langle T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia}, T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia} \rangle_K^2}{\| T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia} \|_K} \\
&= \sum_{a=1}^{m} \langle T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia}, T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia} \rangle_K \\
&= \sum_{a=1}^{m} \langle \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia}, T_k \frac{1}{n} \sum_{i=1}^{n} \varphi_i f_{ia} \rangle_2 \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{a=1}^{m} \int \varphi_i(u) f_{ia} \left(\int K((S_i, u), (S_j, u')) \varphi_j(u') f_{ja} \, du' \right) du \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{a=1}^{m} \int \int \varphi_i(u) f_{ia} \varphi_j(u') f_{ja} K((S_i, u), (S_j, u')) \, du \, du' \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{a=1}^{m} \mathbb{E}[f_{ia} \tilde{f}_{ja} K(((S_i, U_i), (S_j, \tilde{U}_j)) \mid Z_i, Z_j]
\end{aligned}
$$

Next, we convert this into a quadratic objective in W. Recall that $f_{ia} = W_i \delta_{A_i a} - d(S_i) \pi_e(a \mid S_i, U_i)$, and $k_{ij} = K((S_i, U_i), (S_j, \tilde{U}_j))$. Then given this immediately follows from basic matrix algebra that

$$
\begin{aligned}
\sup_{g \in \mathcal{G}_K} J_\lambda(W, g) &= \sup_{g \in \mathcal{G}_K} B(W, g) + \frac{\lambda}{n^2} \sum_{i=1}^{n} W_i^2 \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} \sum_{a=1}^{m} \mathbb{E}[f_{ia} \tilde{f}_{ja} k_{ij} \mid Z_i, Z_j] + \frac{\lambda}{n^2} \sum_{i=1}^{n} W_i^2 \\
&= \frac{1}{n^2} \sum_{i,j=1}^{n} W_i W_j \left(\delta_{A_i A_j} \mathbb{E}[k_{ij} \mid Z_i, Z_j] + \lambda \delta_{ij} \right) \\
&\quad - 2 \frac{1}{n^2} \sum_{i,j=1}^{n} W_i d(S_j) \mathbb{E}[\pi_e(A_i \mid S_j, U_j) k_{ij} \mid Z_i, Z_j] \\
&\quad + \frac{1}{n^2} \sum_{i,j=1}^{n} d(S_i) d(S_j) \mathbb{E}[\sum_{a=1}^{m} (\pi_e(a \mid S_i, U_i) \pi_e(a \mid S_j, U_j)) k_{ij} \mid Z_i, Z_j] \,.
\end{aligned}
$$

Next, for each $i \in [n_z]$, define $W'_i = (\sum_{j=1}^{n} \mathbb{1}\{Z_j = z_i\} W_i) / (\sum_{j=1}^{n} \mathbb{1}\{Z_j = z_i\})$. That is, W'_i measures the average W value over all indices where $Z = z_i$, and let N be defined as in the problem statement. Then given the above, we clearly have $\sup_{g \in \mathcal{G}_K} J_\lambda(W, g) = (W')^T G W' - 2 g^T W' + C$, where G and g are defined as in the proof statement, and

$$
C = \frac{1}{n^2} \sum_{i,j=1}^{n} d(S_i) d(S_j) \mathbb{E}[\sum_{a=1}^{m} (\pi_e(a \mid S_i, U_i) \pi_e(a \mid S_j, U_j)) k_{ij} \mid Z_i, Z_j].
$$

Therefore, we can conclude that the objective only depends on the average W value at all indices where $Z = z_i$ for each $i \in [n_z]$, so it is sufficient to optimize over W' and just set $W_i = W'_{\nu(i)} \; \forall i \in [n]$. Thus, given that C is constant in W, the required result immediately follows from the above.

□

C Sensitivity Theory

In this appendix we present some details on the sensitivity of our theory under minor violations of the iid confounders assumption. We consider a generalization of the MDPUC model depicted in fig. 2, where we allow the unobserved confounder values to be correlated rather than assuming them to be iid. For this analysis we define the following terms similar to those in our core theory:

$$\begin{aligned} f^*_{ia} &= W_i \delta_{A_i a} - d(S_i, U_i)\pi_e(a \mid S_i, U_i) \\ B^*(W, g) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m}\mathbb{E}[f^*_{ia} g_a(S_i, U_i) \mid Z_{1:n}] \\ J^*_\lambda(W, g) &= B^*(W, g)^2 + \frac{\lambda}{n^2}\|W\|^2. \end{aligned}$$

We note that these only differ from the original terms in two respects: (1) conditioning on all observed triplets $Z_{1:n}$ rather than the single observed triplet Z_i in the i'th term; and (2) use of density ratio $d(S, U)$ rather than $d(S)$. Given this, we can first obtain the following lemma under a mild modification of our overlap assumption.

Assumption 5. *$\|d(S, U)\|_q < \infty$, where $2 < q \leq \infty$ is the same value referred to in assumption 1.*

Lemma 8. *Let assumptions 1, 3 and 5 be given. Then we have*

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}] \leq 2J^*_{4\sigma^2}(W, \mu) + O_p(1/n).$$

This proof of this lemma is almost identical to that of theorem 1, and is detailed in appendix C.1. Next, we define the error term

$$\epsilon(W, \mu) = |B^*(W, \mu)^2 - B(W, \mu)^2|^{1/2}.$$

Then it follows that

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}] \leq 2J_{4\sigma^2}(W, \mu) + \epsilon(W, \mu)^2,$$

and therefore if we choose W such that $J_{4\sigma^2}(W, \mu) = O_p(r_n)$, then applying Kallus (2016, Lemma 31) gives us

$$(\hat{\tau}_W - v(\pi_e))^2 = \epsilon(W, \mu)^2 + O_p(\max(1/n, r_n)).$$

That is, by bounding $\epsilon(W, \mu)$ we can bound the irreducible bias from our balanced policy evaluation in the non-iid setting. Note that our theory for providing conditions where $J_{4\sigma^2}(W, \mu) = O_p(1/n)$ (from theorem 2, assuming no nuisance error), or $J_{4\sigma^2}(W, \mu) = O_p(\max(1/n, nr_n^4))$ (assuming $O_p(r_n)$ nuisance error) does not depend on the assumption that U_i values are iid, and therefore still applies here.

Next, we let b be a constant such that $|\mu_a(s, u)| \leq b \; \forall a, s, u$ (which must exist given assumption 3), and we let $D_{\mathcal{F}}$ be defined as in section 4.2, and we let φ_i and φ^*_i be defined as in theorem 4. Given these definitions, we provide the following result on the residual bias $\epsilon(W, \mu)$:

Lemma 9. *Suppose F is some constant such that for every $S \in \mathcal{S}, A \in [m]$ we have $\|\mu_A(S, \cdot)\|_{\mathcal{F}} \leq F$ and $\|\sum_{a=1}^{m}\pi_e(a \mid S, \cdot)\mu_a(S, \cdot)\|_{\mathcal{F}} \leq F$. Then given assumptions 1, 3 and 5, we have*

$$\epsilon(W, \mu) \leq F\left(\frac{1}{n}\sum_{i=1}^{n}(|W_i| + 1)D_{\mathcal{F}}(q_{1:n}, q_i)\right) + b\|d(S, U) - d(S)\|_2 + 2|B(W, \mu)| + O_p(1/n).$$

Then given lemma 8, it follows that Lemma 9 gives a bound on the irreducible squared bias of $\hat{\tau}_W$ as $n \to \infty$.

We note that this bound is an explicit function of the difference between $P(U_i \mid Z_i)$ and $P(U_i \mid Z_{1:n})$ for each i, and the difference between $d(S)$ and $d(S, U)$. Furthermore in the iid confounder case this bound on the squared bias vanishes to zero as $n \to \infty$, as long as $J_{4\sigma^2}(W, \mu) = o_p(1)$, as is ensured by our balancing theory under the assumptions in section 4. This provides some concrete justification for our intuition that in “near-iid” settings our estimator should be close to consistent.

Next, we observe that if one uses the supremum norm for $D_{\mathcal{F}}$ then the corresponding IPM is total variation distance, and we easily satisfy the theorem requirements with $F = b$ given assumption 3. However the general form of the theorem allows for alternate tighter bounds in terms of weaker IPMs under assumptions on the norm of μ. In particular if we assume μ is contained in an RKHS, as in the case of our kernel-based algorithm, another natural choice for $D_{\mathcal{F}}$ would be the corresponding maximum mean discrepancy (MMD).

In addition we note that given some assumptions on μ, all terms in the bound can be estimated in practice for a given weighted estimator. This means practitioners can estimate the bound under different non-iid model assumptions and assumptions on μ, in order to perform sensitivity analysis. Furthermore given the dependence of the first term in this estimator on $\|W\|_\infty$, this may motivate additional regularization on W in non-iid settings. However we leave further exploration of this idea to future work.

Finally, we provide the cautionary note that in the non-iid setting the identification assumptions for $d(S)$ are invalid, and therefore our proposed algorithm for learning the state density ratio may be inconsistent. Therefore the $d(S_i)$ terms in the above theorem should be interpreted as coming from the possibly biased d function used by the optimal balancing algorithm. The theorem then provides an explicit bound on the incurred bias due to this. Note however that in the case that $d(S) \approx d(S, U)$, the estimating equations in section 5.1 are approximately correct, so we do not expect this to be a major issue in practice. This is further justified by the strong positive results of our sensitivity experiments.

C.1 Omitted Proofs for Sensitivity Theory

Proof of lemma 8. First we define sample average policy effect slightly differently for the non-iid setting, as:

$$\text{SAPE}^*(\pi_e) = \frac{1}{n} \sum_{i=1}^{n} \sum_{a=1}^{m} d(S_i, U_i) \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i).$$

Again, following the derivation in section 4 we have $\mathbb{E}[\text{SAPE}(\pi_e)] = v(\pi_e)$. Given this and assumptions 1, 3 and 5, it is clear that the conditions of lemma 4 apply to $\mathbb{E}_b[(\text{SAPE}^*(\pi_e) - v(\pi_e))^2]$, so this term must be $O(1/n)$. Thus by Markov's inequality and the law of total expectation we have $\mathbb{E}[(\text{SAPE}^*(\pi_e) - v(\pi_e))^2 \mid Z_{1:n}] = O_p(1/n)$. Then, using the fact that $(x+y)^2 \leq 2x^2 + 2y^2$, we have

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n}] \leq 2\mathbb{E}[(\hat{\tau}_W - \text{SAPE}^*(\pi_e))^2 \mid Z_{1:n}] + O_p(1/n).$$

Next, we perform a bias variance decomposition of the RHS of this bound as follows:

$$\begin{aligned} \mathbb{E}[(\hat{\tau}_W - \text{SAPE}^*(\pi_e))^2 \mid Z_{1:n}] &= \mathbb{E}[\mathbb{E}[(\hat{\tau}_W - \text{SAPE}^*(\pi_e))^2 \mid Z_{1:n}, U_{1:n}] \mid Z_{1:n}] \\ &= \mathbb{E}[\mathbb{E}[\hat{\tau}_W - \text{SAPE}^*(\pi_e) \mid Z_{1:n}, U_{1:n}]^2 \mid Z_{1:n}] \\ &\quad + \mathbb{E}[\mathbb{V}[\hat{\tau}_W - \text{SAPE}^*(\pi_e) \mid Z_{1:n}, U_{1:n}] \mid Z_{1:n}] \\ &= \xi_1^* + \xi_2^*, \end{aligned}$$

and we additionally define

$$\begin{aligned} \zeta_{ia}^* &= W_i \delta_{A_i a} R_i - d(S_i, U_i) \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i) \\ \zeta_i^* &= \sum_{a=1}^{m} \zeta_{ia}^* = W_i R_i - d(S_i, U_i) \sum_{a=1}^{m} \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i). \end{aligned}$$

Given this, the first term of the above bias variance decomposition can be broken down as:

$$\begin{aligned}
\xi_1^* &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia}^*\right)^2 \,\middle|\, Z_{1:n}\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia}^* \,\middle|\, Z_{1:n}\right]^2 + \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \zeta_{ia}^* \,\middle|\, Z_{1:n}\right] \\
&= \left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \mathbb{E}[f_{ia}^* \mu_a(S_i, U_i) \mid Z_{1:n}]\right)^2 + \frac{1}{n^2}\mathbb{V}[\sum_{i=1}^n \zeta_i^* \mid Z_{1:n}] \\
&\leq \left(\frac{1}{n}\sum_{i=1}^n \sum_{a=1}^m \mathbb{E}[f_{ia}^* \mu_a(S_i, U_i) \mid Z_{1:n}]\right)^2 + \frac{2\sigma^2}{n^2}\sum_{i=1}^n W_i^2 \\
&\quad + 2\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i, U_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \,\middle|\, Z_{1:n}\right] \\
&= B^*(W, \mu)^2 + \frac{2\sigma^2}{n^2}\|W\|^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i, U_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \,\middle|\, Z_{1:n}\right].
\end{aligned}$$

Similarly, we bound the the second error term ξ_2^2 as:

$$\begin{aligned}
\xi_2^* &= \mathbb{E}\left[\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n \zeta_i^* \,\middle|\, Z_{1:n}, U_{1:n}\right] \,\middle|\, Z_{1:n}\right] \\
&\leq \mathbb{E}\left[\frac{2\sigma^2}{n^2}\sum_{i=1}^n W_i^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i, U_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \,\middle|\, Z_{1:n}, U_{1:n}\right] \,\middle|\, Z_{1:n}\right] \\
&\leq \frac{2\sigma^2}{n^2}\|W\|^2 + 2\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i, U_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i) \,\middle|\, Z_{1:n}\right],
\end{aligned}$$

where the final inequality follows from the law of total variance. Next, applying assumptions 1, 3 and 5, it clearly follows from lemma 4 that

$$\mathbb{V}\left[\frac{1}{n}\sum_{i=1}^n d(S_i, U_i) \sum_{a=1}^m \pi_e(a \mid S_i, U_i)\mu_a(S_i, U_i)\right] = O(1/n),$$

and therefore by Markov's inequality the corresponding conditional variance is $O_p(1/n)$.

Putting the above bounds together we get

$$\mathbb{E}[(\hat{\tau}_W - v(\pi_e))^2 \mid Z_{1:n})] \leq 2\left(B^*(W, \mu)^2 + \frac{4\sigma^2}{n^2}\|W\|^2\right) + O_p(1/n),$$

which gives us our required result immediately.

□

Proof of lemma 9. First we can obtain the bound

$$\begin{aligned}
\epsilon(W, \mu)^2 &= |B^*(W, \mu)^2 - B(W, \mu)^2| \\
&= |B^*(W, \mu) - B(W, \mu)||B^*(W, \mu) + B(W, \mu)| \\
&= |B^*(W, \mu) - B(W, \mu)||2B(W, \mu) + (B^*(W, \mu) - B(W, \mu))| \\
&\leq |B^*(W, \mu) - B(W, \mu)|(2|B(W, \mu)| + |B^*(W, \mu) - B(W, \mu)|) \\
&\leq (2|B(W, \mu)| + |B^*(W, \mu) - B(W, \mu)|)^2.
\end{aligned}$$

Next, let b be a constant such that $|\mu_a(s,u)| \leq b \; \forall a, s, u$, which by assumption 3 must exist, and define the notation shorthand

$$\begin{aligned} e_i(\cdot) &= \mathbb{E}[\cdot \mid Z_i] \\ e_{1:n}(\cdot) &= \mathbb{E}[\cdot \mid Z_{1:n}]. \end{aligned}$$

Given this we can obtain the bound

$$\begin{aligned} |B^*(W,\mu) - B(W,\mu)| &\leq \left| \frac{1}{n} \sum_{i=1}^{n} (e_{1:n} - e_i) \left(\sum_{a=1}^{m} f_{ia}^* \mu_a(S_i, U_i) \right) \right| \\ &\quad + \left| \frac{1}{n} \sum_{i=1}^{n} e_i \left(\sum_{a=1}^{m} (f_{ia}^* - f_{ia}) \mu_a(S_i, U_i) \right) \right| \\ &\leq \frac{1}{n} \sum_{i=1}^{n} |W_i| \left| (e_{1:n} - e_i) \sum_{a=1}^{m} \delta_{A_i a} \mu_a(S_i, U_i) \right| \\ &\quad + \frac{1}{n} \sum_{i=1}^{n} \left| (e_{1:n} - e_i) \sum_{a=1}^{m} \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i) \right| \\ &\quad + \left| \mathbb{E}\left[(d(S_i, U_i) - d(S_i)) \sum_{a=1}^{m} \pi_e(a \mid S_i, U_i) \mu_a(S_i, U_i) \right] \right| + O_p(1/n) \\ &\leq \frac{F}{n} \sum_{i=1}^{n} (|W_i| + 1) D_{\mathcal{F}}(\varphi_i, \varphi_i^*) + b \mathbb{E}[(d(S,U) - d(S))^2]^{1/2} + O_p(1/n) \\ &\leq \frac{F}{n} \sum_{i=1}^{n} (|W_i| + 1) D_{\mathcal{F}}(\varphi_i, \varphi_i^*) + b \|d(S,U) - d(S)\|_2 + O_p(1/n), \end{aligned}$$

where in the second inequality we apply the Markov chain law of large numbers, in the third and final inequalities we apply Cauchy Schwartz and our $\|\cdot\|_{\mathcal{F}}$ bound assumptions. Putting the above together, we obtain the final bound:

$$\epsilon(W,\mu) \leq F \left(\frac{1}{n} \sum_{i=1}^{n} (|W_i| + 1) D_{\mathcal{F}}(\varphi_i, \varphi_i^*) \right) + b(\|d(S,U) - d(S)\|_2 + 2\|B(W,\mu)\| + O_p(1/n).$$

$\square$

D Discussion of Nuisance Estimation

We discuss here some of the existing theory regarding the estimation of the posterior distributions φ, and the state density ratio d, including the assumptions neeed for identification and for the rates of convergence required by our theory.

D.1 Estimation of Confounder Posterior Distribution

We provide some discussion here for convergence rates of $D_{\mathcal{F}}(\hat{\varphi}(Z), \varphi(Z))$ in the case where $\|f\|_{\mathcal{F}} = \|f\|_\infty$, which corresponds to total variation distance, since this metric dominates most other integral probability metrics (IPMs) of interest.

First, for any given z we can obtain the bound

$$\begin{aligned} D_{\mathcal{F}}(\varphi(z), \hat{\varphi}(z)) &= \sup_{\|f\|_\infty} \left| \int f d\varphi(z) - \int f d\hat{\varphi}(z) \right| \\ &\leq \left| \int \varphi(z)(u) - \hat{\varphi}(z)(u) du \right| \\ &\leq \sup_{u \in \mathcal{U}} |\varphi(z)(u) - \hat{\varphi}(z)(u)| \int du. \end{aligned}$$

Now, under the assumption that $\mathcal{U}$ is compact, we have $\int du < \infty$, so it is sufficient to consider the convergence rate of $\sup_{u \in \mathcal{U}} |\varphi(z)(u) - \hat{\varphi}(z)(u)|$. We analze this convergence for multiple cases below.

D.1.1 Discrete States and Confounders

The simplest case to consider here is the case where both S and U are discrete, as in our experiments. Under this assumption, the above bound translates to requiring that $|\varphi(z)(u) - \hat{\varphi}(z)(u)|$ converges sufficiently fast for each U and Z level. Fortunately, in this case the probabilities $P(U \mid Z)$ are given by parameters in some parametric latent variable model, which can be fit using approaches such as expectation maximization (EM) (Dempster et al., 1977), Bayesian estimators (Lehmann & Casella, 2006), or spectral methods (Hsu et al., 2009; Shaban et al., 2015). In particular, maximum likelihood-based approaches such as the EM algorithm, are known to be efficient and achieve the $O_p(n^{-1/2})$-convergence required for $O_p(n^{-1/2})$ OPE consistency (Van der Vaart, 2000). Note that in the case of EM this depends on solving the difficult non-convex optimization problem, however this challenge may be mitigated by initializing EM with some non-local optimization method (Shaban et al., 2015). This analysis depends on the assumption that the confounder model is well-specified (i.e. confounders are actually discrete, and we do not underestimate the number of confounder levels). In addition it depends on standard identifiability conditions needed for latent variable models in general (Dempster et al., 1977).

D.1.2 Continuous States and Discrete Confounders

In this next case U is still assumed to be discrete, so again it is sufficient to ensure that for any given z, we have that $|\varphi(z)(u) - \hat{\varphi}(z)(u)|$ converges sufficiently fast for each $u \in \mathcal{U}$. If we assume a parametric model such that $\varphi(z) = \varphi_{\theta_0}(z)$ for some finite-dimensional parameter space Θ and some $\theta_0 \in \Theta$, then θ_0 can be estimated using the kinds of approaches described in the previous section. Under standard correct-specification and identifiability assumptions it easily follows that we can obtain $O_p(n^{-1/2})$ consistency for estimating θ_0. Then under some smoothness assumptions of $\varphi_\theta(z)$ (e.g. locally Lipschitz at θ_0), it follows that $|\varphi(z)(u) - \hat{\varphi}(z)(u)| = O_p(n^{-1/2})$, and therefore we can obtain the same parametric rate for our policy value estimate. Alternatively, if we assume some kind of semi- or non-parametric model for $\varphi(z)$, then we may still be able to estimate $\varphi(z)(u)$ at some rate in between $O_p(n^{-1/4})$ and $O_p(n^{-1/2})$ using machine learning methods, under some smoothness assumptions, as is standard for flexible nuisance estimation in causal inference (see for example discussion in Chernozhukov et al. (2016)).

D.1.3 Continuous States and Confounders

In this final most general case, we can again consider estimating $\varphi(z)$ either by assuming a parametric model, or using flexible machine learning methods that exploit smoothness. Again this can result in estimates of $\varphi(z)(u)$ that are either $O_p(n^{-1/2})$-consistent under parametric assumptions, or consistent at some slower rate under more general smoothness assumptions. This allows us to guarantee convergence for any fixed $u \in \mathcal{U}$, however in this case we have the additional complexity that the space $\mathcal{U}$ is not finite, and therefore we need to establish the convergence of $\sup_{u \in \mathcal{U}} |\varphi(z)(u) - \hat{\varphi}(z)(u)|$. Let $Q_n(u) = (\varphi(z) - \hat{\varphi}(z))/r_n$. Then if we assume that Q_n is uniformly sub-Gaussian in $\mathcal{U}$ for every $n \in \mathbb{N}$ (that is there exists some semi-metric d on $\mathcal{U}$ such that $P(|Q_n(u) - Q_n(u')| > x) \leq 2\exp(-\frac{1}{2}x^2/d(u,u')^2)$ for every $n \in \mathbb{N}$, $u, u' \in \mathcal{U}$), it follows easily from standard chaining arguments (Kosorok, 2007, Corollary 8.5 and Theorem 2.1) that $\sup_{u \in \mathcal{U}} |\varphi(z)(u) - \hat{\varphi}(z)(u)| = O_p(r_n)$. Note that following standard empirical process theory arguments, this required sub-Gaussian assumption may be justified based on compactness of $\mathcal{U}$ and Lipschitz continuity assumptions.

D.2 Estimation of State Density Ratio

Here we discuss the rate of convergence of the state density ratio d. First, in the case that $\mathcal{S}$ is discrete, as in our experiments, the variational GMM algorithm we proposed reduces to a standard efficient GMM algorithm for a finite number of parameters (in the case that $\mathcal{S} = [n_s]$, these parameters are $d(1), \ldots, d(n_s)$) as discussed in appendix E. These algorithms are known to be semi-parametrically efficient, with $O_p(n^{-1/2})$ consistency (Hansen, 1982), as required for $O_p(n^{-1/2})$-consistent estimation of $v(\pi_e)$.

In the more general case, where $\mathcal{S}$ is continuous, the theory on the rate of convergence of $\hat{d}$ is less clear. If we replaced the RKHS class for $\mathcal{D}$ used in our algorithm with a parametric class, then under an identifiability assumption on the class $\mathcal{H}$ (that it is sufficiently rich to identify d), and the assumption that $\mathcal{H}$ has a finite basis (such as in the case of a polynomial kernel), then again this corresponds to a standard efficient GMM estimate and $O_p(n^{-1/2})$-consistency would follow from standard GMM theory (Hansen, 1982). On the other hand in the more general case we consider in section 5.1, where $\mathcal{D}$ and $\mathcal{H}$ are both flexible potentially non-parametric function classes, consistency of $\hat{d}$ could be established using a proof almost identical to that in Bennett et al. (2019). However the rate of convergence in general settings where $\mathcal{D}$ and $\mathcal{H}$ can both be arbitrary RKHSs is unclear, and we leave this problem to future work.

E Derivation of Algorithm for State Density Ratio Estimation

We discuss here the theoretical derivation of the variational GMM algorithm presented in section 5.1 for state density ratio estimation.

First, we observe that it follows easily from a generalization of Bennett et al. (2019, Lemma 1) (replacing the instrumental variable regression conditional moment restrictions there with the state density ratio conditional moment restrictions) that if $\mathcal{H}$ is the vector space spanned by functions $\{h_1, \ldots, h_k\}$, and $\mathcal{D}$ is given by some parametric class, then the estimator

$$\hat{d} = \operatorname*{arg\,min}_{d \in \mathcal{D}} \sup_{h \in \mathcal{H}, c \in \mathbb{R}} U_n(d, \tilde{d}, h, c)$$

is exactly the same as the standard optimally-weighted GMM estimator (Hansen, 1982) given by the $k+1$ standard moment restrictions

$$\begin{gathered}\mathbb{E}[h_i(S')(d(S)\beta(Z) - d(S'))] = 0 \; \forall i \in [k] \\ \mathbb{E}[d(S) - 1] = 0.\end{gathered}$$

Given standard regularity assumptions, that $d \in \mathcal{D}$, the $k+1$ moment restrictions are sufficient to uniquely identify d, and that the parametric class for $\mathcal{D}$ is sufficiently smooth, then it follows from standard theory that this estimator is root-n consistent and asymptotically normal, and if the prior estimate $\tilde{d}$ is consistent then the estimator is statistically efficient relative to all other estimators based on these $k+1$ moment conditions. Note that given the above, efficiency is easily ensured by running the adversarial optimization at least twice, starting with an initial arbitrary guess for $\tilde{d}$ and then each time using the previous iterate estimate $\hat{d}$ for $\tilde{d}$, as proposed in section 5.1.

Given this, it is natural to consider extending this standard GMM estimator by replacing $\mathcal{D}$ and $\mathcal{H}$ by sufficiently regularized flexible function classes, such as neural networks or RKHSs. This is motivated by wanting to avoid the known curse of dimensionality issues of seive estimators using increasingly large numbers of standard moment conditions. Previously Bennett et al. (2019) proposed to use such an estimator for the instrumental variable regression problem using neural networks for both function classes. On the other hand we propose to use RKHSs, which has the nice benefit that the optimization can be performed analytically by appealing to the representor theorem (as given by theorem 6).

F Additional Experiment Details

F.1 Baseline Descriptions

Direct Method: This method works by using the approximate confounder model to directly fit an outcome model. Specifically, first we use the confounder-imputed dataset to fit a model $\hat{\mu}$ for μ via regressing R on $(S, \hat{U})$ for each $a \in [m]$. Given that our experiments work with discrete states and confounders, this is done simply by averaging the observed reward for each possible (s, u) pair. Then we use the estimated outcome model, stationary density ratio, and confounder model to directly estimate $v(\pi_e)$, according to

$$\begin{aligned}\hat{\tau}_{\mathrm{DM}}^{(i)} &= \sum_{u,a} \hat{\varphi}(u \mid Z_i) \pi_e(a \mid S_i, u) \hat{\mu}_a(S_i, u) \\ \hat{\tau}_{\mathrm{DM}} &= \frac{1}{n} \sum_{i=1}^{n} \hat{d}(S_i) \hat{\tau}_{\mathrm{DM}}^{(i)}. \end{aligned} \tag{4}$$

Doubly Robust This method combines the Direct Method and our weighted estimator approach. Specifically given weights $W_{1:n}$ and an outcome model $\hat{\mu}$ fit as above, we calculate

$$\begin{aligned}\hat{\tau}_{\mathrm{DR}}^{(i)} &= \sum_{u} \hat{\varphi}(u \mid Z_i) \hat{\mu}_{A_i}(S_i, u) \\ \hat{\tau}_{\mathrm{DR}} &= \hat{\tau}_{\mathrm{DM}} + \frac{1}{n} \sum_{i=1}^{n} W_i (R_i - \hat{\tau}_{\mathrm{DR}}^{(i)}). \end{aligned} \tag{5}$$

Inverse Propensity Score (IPS) This is a recently proposed effective approach to infinite-horizon OPE (Liu et al., 2018), under the naive assumption of no hidden confounding. This method works by fitting both inverse propensity scores and the state density ratio, using similar conditional moment conditions as in section 5.1.

Black-Box This is a state-of-the-art approach to OPE Mousavi et al. (2020), which is similar in nature to *IPS* but works under more general assumptions and tends to be more robust to behavior data sampling distributions than IPS . It also naively assumes no hidden confounding.

F.2 Hyperparameter Details

Estimating State Density Ratio. We follow the algorithm described by theorem 6. In both cases, for each of $\mathcal{H}$ and $\mathcal{D}$ identity kernels ($K(s, s') = \mathbb{1}\{s = s'\}$). Furthermore, we set the hyperparameters λ_h, λ_c, and λ_d all equal to 10^{-8} in both cases. Finally, we initialize $\tilde{d}$ to be a vector of all ones, and we iterate the min-max calculation of $\hat{d}$ five times, each time using the previous iterate solution as $\tilde{d}$.

Calculating Optimal Balancing Weights. We use the following kernel for our RKHS for $\mathcal{G}_K$: $k((s, u), (s', u')) = 0.5\mathbb{1}\{s = s'\} + 0.5\mathbb{1}\{u = u'\}$, which takes into account the tuple structure of the input of μ. In addition we use $\lambda = 10^{-3}$ in all experiments, as we found this gave consistently good performance (as in Bennett & Kallus (2019), we find that small values of λ perform well).

IPS and Black-Box. In general, both of these approaches use neural networks as parametric models to learn the weights of the estimator. However, both environments that we have studied in this paper (i.e., confounded Modelwin and GridWorld) have finite and discrete state space. Therefore, as suggested in Section 5 of Liu et al. (2018) (and similarly in Mousavi et al. (2020)) we can optimize the weights of the estimator in the space of all possible functions. This corresponds to using a delta kernel in terms of the RKHS used for defining the maximum mean discrepancy in both methods. Accordingly, minimizing loss functions in both baselines (i.e., eq. (12) in Liu et al. (2018) and eq. (11) in Mousavi et al. (2020)) reduce to quadratic optimization problems, which we solve using constrained optimization by linear approximation (COBYLA).

F.3 Environment Details

C-Modelwin. C-Modelwin has 3 states (denoted s_0, s_1, and s_2) and 2 actions (denoted a_0 and a_1). The agent always begins in s_0. At time t, the agent chooses between the actions a_0 and a_1 with probabilities $1 - \pi - U_t$ and $\pi + U_t$ respectively regardless of the current state, where π is a scalar policy parameter. In our experiments, we use a behavior policy with $\pi = 0.7$, and an evaluation policy with $\pi = 0.1$. In addition, $U_{i:n}$ are iid variables taking value 0.1 or 0.2 with probabilities 0.3 and 0.7 respectively.

Transitions and rewards occur as follows. If the agent is in state s_0 at time t and takes action a_0, it transitions to s_1 or s_2 with probabilities $0.7 + U_t$ and $0.3 - U_t$ respectively. Alternatively, if it takes action a_1 in state s_0 then it transitions to s_1 or s_2 with probabilities $0.3 + U_i$ and $0.7 - U_i$ respectively. In either case it receives zero reward transitioning from s_0. If the agent is in state s_1 or s_2 it transitions to s_0, regardless of the action taken. Furthermore, when it transitions from s_1 to s_0 it receives a reward of $10 + 20U_i$, and when it transitions from s_2 to s_0 it receives a reward of $-10 - 20U_i$. In both cases the reward doesn't depend on the action taken.

GridWorld. The environment consists of a 10×10 grid, and each state corresponds to the agent's location in the grid (meaning that there are 100 different states). The agent starts from the bottom-left of the grid, and its goal is to reach the top-right of the grid. There are four possible actions: moving *up* (a_0), *right* (a_1), *down* (a_2), and *left* (a_3). We consider a class of hierarchical policies that first decide whether to move towards the top-right or towards the bottom-left, and then consider whether to move up or right (in case of moving towards top-right), or whether to move down or left (in case of moving towards bottom-left). Specifically, we consider policies that are parameterized by a single scalar parameter π. At time t, the agent first decides to move towards the bottom-left with probability $\pi + U_t$, or the top-right with probability $1 - \pi - U_t$. In the case of moving towards the bottom-left, the agent moves down with probability $0.5\pi + U_t$, or left with probability $1 - 0.5\pi - U_t$. Converseley, in the case of moving towards the top-right, the action taken depends on whether the agent is above or below the diagonal from the bottom-left to top-right: if the agent is below this diagonal they move up with probability $\pi + U_t$ or right with probability $1 - \pi - U_t$; if they are above this diagonal they move up with probability $1 - \pi - U_t$ or right with probability $\pi + U_t$; and if they are on the diagonal they move up with probability $0.5\pi + 0.5U_t$ or right with probability $1 - 0.5\pi - 0.5U_t$. As in C-ModelWin, the confounders $U_{1:n}$ are iid variables taking value 0.1 or 0.2 with probabilities 0.3 and 0.7 respectively, and we use $\pi = 0.7$ for the behavior policy, and $\pi = 0.1$ for the evaluation policy.

State transitions are mostly simple and deterministic; unless the agent is at the goal position of the top-right corner of the grid, it moves one space in the direction indicated by the action (up, right, down, or left). In the case that the agent cannot move in that direction because they are at the edge of the grid (for example if it is at the very right and takes the right action) they simply do not move. On the other hand if the agent is at the top-right corner before taking the action, they transition to the bottom-left corner regardless of the action taken.

Rewards are also simple and deterministic. At time t, if the agent is at the goal position of the top-right corner it receives a reward of $100 + 100U_t$, regardless of the action taken. Otherwise, it receives a deterministic reward based on the action taken regardless of the state: $1 + 20U_t$ for up, $1 + 30U_t$ for right, $-1 - 30U_t$ for down, and $-1 - 40U_t$ for left. Note that the agent still receives this reward if it is at the edge of the grid and therefore cannot move.

F.4 Model Misspecification Details

As discussed in the section 6, in our sensitivity to model misspecification experiments we assume confounders are distributed according to $\alpha\mathcal{P}_{\text{iid}} + (1-\alpha)\mathcal{P}_{\text{alt}}$ where $\mathcal{P}_{\text{iid}}$ denotes the original distribution in which confounder values all independent, $\mathcal{P}_{\text{alt}}$ denotes a distribution in which the confounder value at time t depends on the confounder value at time $t-1$, and α is a model hyperparameter.

Next, as described in appendix F.3, in both environments the original model $\mathcal{P}_{\text{iid}}$ is given by a simple categorical distribution, where each confounder takes the value 0.1 or 0.2 with probabilities 0.3 and 0.7 respectively. On the other hand, in the alternative model $\mathcal{P}_{\text{alt}}$ the confounder still takes the value 0.1 or 0.2, with probabilities that depend on the previous confounder value. Specifically, for the initial time step the respective probabilities are 0.3 and 0.7, as in $\mathcal{P}_{\text{iid}}$, and for future time steps the respective probabilities are 0.08 and 0.92 if the previous confounder value was 0.1, or 0.82 and 0.18 if the previous confounder value was 0.2.

F.5 Posterior Noise Injection Details

We describe here both how we inject noise in the posterior distributions $\varphi(z)$, and how we measure this noise. Recall that $\varphi(z)$ is shorthand for the posterior distribution of U given $Z = z$, that is $\varphi(z)(u) = P(U = u \mid Z = z)$. In our experiments all U and Z values are discrete, so we have a finite number of posterior distributions $\varphi(z)$, each represented by a finite-length vector. Let logits(p) denote the vector of log-odds corresponding to the vector of probabilities p. Then for each possible value z, we independently injected noise in $\varphi(z)$ by adding a random Gaussian vector to logits($\varphi(z)$), and then converting the perturbed logits back to probabilities (by taking the expits of the vector entries and re-normalizing). This was done for a wide variety of different variances of the random Gaussian vectors (all with spherical covariances).

It is difficult to interpret the scale of posterior error caused by a given variance for the Gaussian vector we added to the posterior logits, so we came up with the more interpretable metric average standard deviation (ASD). In this metric the average is taken over the distribution of Z values and levels of U, and the standard deviation is taken over the distribution of random noise vectors. Formally, let n_s be a number of Z values to sample from the stationary distribution of π_b, let n_e be a number of random Gaussian vectors to sample for each sampled Z value, and let n_u be the number of levels of U. In practice in our experiments we use $n_e = 50$ and $n_s = 5$. In addition, let Z_i be the i'th sampled Z value, let $\epsilon_{i,j}$ be the j'th sampled Gaussian vector for the i'th sampled Z value. In addition let $\psi(Z_i, \epsilon_{i,j})$ denote the vector of probabilities given by perturbing logits($\varphi(Z_i)$) by $\epsilon_{i,j}$, as described above. Then the ASD metric is given by

$$ASD = \frac{1}{n_s n_u} \sum_{i=1}^{n_s} \sum_{u=1}^{n_u} \left(\frac{1}{n_e - 1} \sum_{j=1}^{n_e} \left(\psi(Z_i, \epsilon_{i,j})_u - \frac{1}{n_e} \sum_{j'=1}^{n_e} \psi(Z_i, \epsilon_{i,j'})_u \right)^2 \right)^{1/2}$$

F.6 Additional Plots

In this section we present sensitivity of the direct method and doubly robust estimator to model misspecification and noise in the oracle for the posterior distribution of confounders. For the sake of visualization and clarity, we have repeated plots of off-policy estimates and RMSEs of different methods.

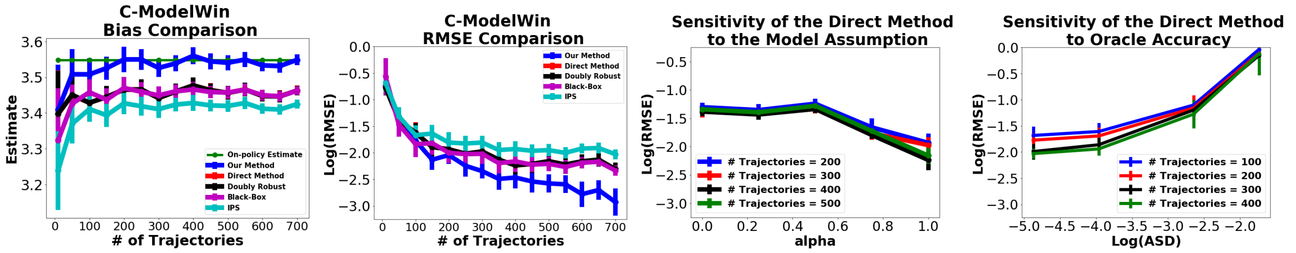

Figure 5: C-ModelWin Results. From left to right: The off-policy estimate, The $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of the direct method to model misspecification, and to noise in the confounders posterior distribution.

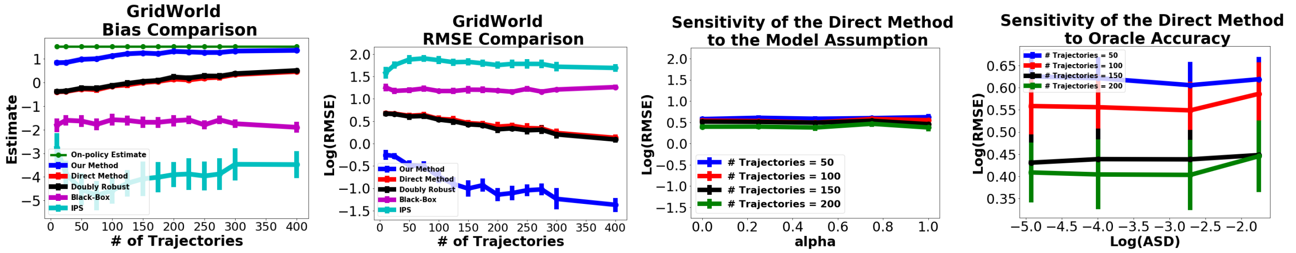

Figure 6: Confounded GridWorld Results. From left to right: The off-policy estimate, The $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of the direct method to model misspecification, and to noise in the confounders posterior distribution.

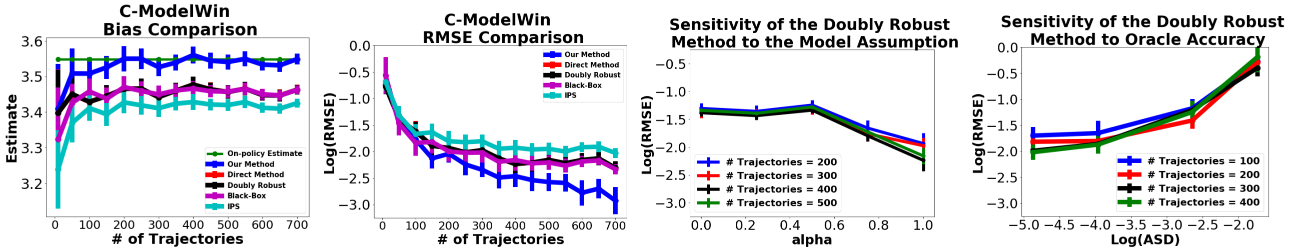

Figure 7: C-ModelWin Results. From left to right: The off-policy estimate, The $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of the doubly robust estimator to model misspecification, and to noise in the confounders posterior distribution.

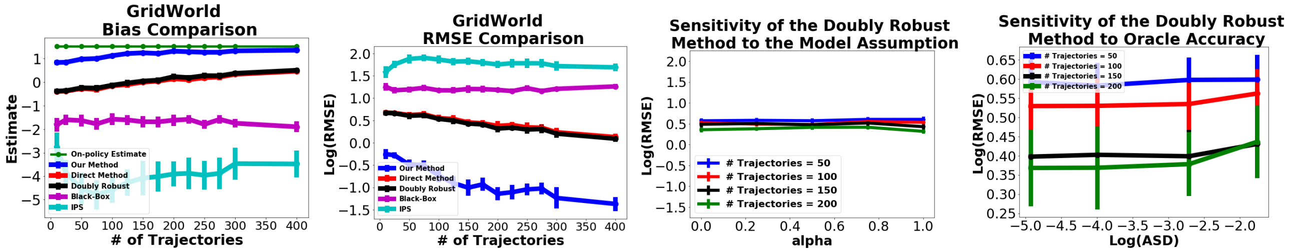

Figure 8: Confounded GridWorld Results. From left to right: The off-policy estimate, The $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of the doubly robust estimator to model misspecification, and to noise in the confounders posterior distribution.