A Notations, Terms, and Abbreviations

In this section, we summarize the notations used in this paper.

<table>
<thead>
<tr>
<th>Notations</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A, X, Y$</td>
<td>Action, covariate, reward</td>
</tr>
<tr>
<td>$E[\mu(X, A, Y)]$</td>
<td>Expected value of the action</td>
</tr>
<tr>
<td>$E[\mu(Z)]$</td>
<td>Expected value of the covariate</td>
</tr>
<tr>
<td>$\pi^b(a</td>
<td>x)$</td>
</tr>
<tr>
<td>$\pi^e(a</td>
<td>x)$</td>
</tr>
<tr>
<td>$R(\pi^e)$</td>
<td>$E[q(x)p(x) \pi(x</td>
</tr>
<tr>
<td>$r(x)$</td>
<td>$p(x)/q(x)$</td>
</tr>
<tr>
<td>$w(a, x)$</td>
<td>$\pi^e(a</td>
</tr>
<tr>
<td>$f(a, x)$</td>
<td>$E[q(x)p(y</td>
</tr>
<tr>
<td>$v(x)$</td>
<td>$E[\pi^e(x</td>
</tr>
<tr>
<td>$\text{Asmse}[^{\hat{R}}]$</td>
<td>$\lim_{n \to \infty} \mathbb{E}[(\hat{R} - R)^2]n$</td>
</tr>
<tr>
<td>$\Pi$</td>
<td>Policy class</td>
</tr>
<tr>
<td>$\otimes A$</td>
<td>$AA^T$</td>
</tr>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$k(\Pi)$</td>
<td>Entropy integral of $\Pi$ w.r.t $\epsilon$-Hamming distance</td>
</tr>
<tr>
<td>$n_{\text{hst}}$</td>
<td>Number of training data</td>
</tr>
<tr>
<td>$n_{\text{evl}}$</td>
<td>Number of evaluation data</td>
</tr>
<tr>
<td>$A \preceq B$</td>
<td>There exists an absolute constant $C$ s.t. $A \leq CB$</td>
</tr>
<tr>
<td>$C_1, C_2, R_{\text{max}}$</td>
<td>Upper bound of $r(X), w(A, X), Y$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$n_{\text{hst}}/(n_{\text{hst}} + n_{\text{evl}})$</td>
</tr>
<tr>
<td>$\mathcal{D}<em>{\text{hst}}, \mathcal{D}</em>{\text{evl}}$</td>
<td>Train data, evaluation data</td>
</tr>
<tr>
<td>$n_{\text{hst}}$, $n_{\text{evl}}^{\text{hst}}$</td>
<td>Split train data</td>
</tr>
<tr>
<td>$n_{\text{evl}}, n_{\text{evl}}^{\text{evl}}$</td>
<td>Split evaluation data</td>
</tr>
<tr>
<td>$\mathcal{D}_i$</td>
<td>Concatenation of $n_{\text{hst}}^{\text{hst}}$ and $n_{\text{evl}}^{\text{evl}}$ for $i = 1, 2$</td>
</tr>
<tr>
<td>$\mathcal{G}_{\text{hst}}$</td>
<td>$\sqrt{n_{\text{hst}}^{\text{hst}}}E[n_{\text{hst}}^{\text{hst}} - E]$ Empirical process based on train data</td>
</tr>
<tr>
<td>$\mathcal{G}_{\text{evl}}$</td>
<td>$\sqrt{n_{\text{evl}}^{\text{evl}}}E[n_{\text{evl}}^{\text{evl}} - E]$ Empirical process based on evaluation data</td>
</tr>
<tr>
<td>$\Upsilon(\pi^e)$</td>
<td>Semi-parametric lower bound of $R(\pi^e)$ under nonparametric model</td>
</tr>
<tr>
<td>$K_h(\cdot)$</td>
<td>Kernel with a bandwidth $h$</td>
</tr>
<tr>
<td>$n_{\text{hst}}^{\text{hst}}$, $n_{\text{evl}}^{\text{evl}}$</td>
<td>$k$-th train data</td>
</tr>
<tr>
<td>$n_{\text{evl}}^{\text{evl}}$</td>
<td>$k$-th evaluation data</td>
</tr>
</tbody>
</table>

B Identification under Potential Outcome Framework

We explain how to apply our results in the main draft under potential outcome framework, which is a common framework in causal inference literature (Rubin, 1987). In this section, our goal is justifying DM and IPWCS estimators under potential outcome framework.

Let us denote counterfactual variables based on stochastic policies (interventions) as $Y(B)$, where $B$ is a random variable from the conditional density $\pi^e(b | Z)$ and $Z$ is a random variable following the evaluation density $q(z)$. Here, note what we can observe is data:

$$\{X_i, A_i, Y_i\}_{i=1}^{n_{\text{hst}}} \sim p(x)\pi^e(a | x)p(y | a, x), \{Z_j\}_{j=1}^{n_{\text{evl}}} \sim q(z).$$

A detailed review of the stochastic intervention is shown in Muñoz & Van Der Laan (2012); Young et al. (2014).

Then, let us put the following assumptions:

- Consistency: $Y = Y(a)$ if $A = a$ for $\forall a \in \mathcal{A}$.

The reason why we use $B$ is to distinguish it from the random variable $A$.
• Unconfoundedness: \( A \) and \( Y(a) \) are conditionally independent given \( X \) for any \( a \in \mathcal{A}, G \) and \( Y(a) \) are conditionally independent given \( Z \) for any \( a \in \mathcal{A}, c \in \mathcal{X} \).

• Transportability: \( E[Y(a) \mid Z = c] = E[Y(a) \mid X = c] \) for any \( a \in \mathcal{A}, c \in \mathcal{X} \).

Note that transportability is a weaker assumption compared with the assumption in the main draft:

\[
p_{\text{train}}(Y(a) \mid c) = p_{\text{test}}(Y(a) \mid c),
\]

where \( p_{\text{train}}(\cdot \mid \cdot) \) is a condition density of \( Y(a) \) given \( Z \), \( p_{\text{test}}(\cdot \mid \cdot) \) is a condition density of \( Y(a) \) given \( X \). Following Lemma 1 (Kennedy, 2019), we can prove the following lemma.

**Lemma 2 (G-formula).** \( E[Y(B)] = \int E[Y \mid A = a, X = x] \pi^e(a \mid x) q(x) d(a, x) \).

**Proof.**

\[
E[Y(B)] = \int E[Y(b) \mid B = b, Z = z] \pi^e(b \mid z) q(z) d(b, z)
\]

\[
= \int E[Y(g) \mid Z = z] \pi^e(b \mid z) q(z) d(b, z)
\]

\[
= \int E[Y(g) \mid X = z] \pi^e(b \mid z) q(z) d(b, z)
\]

\[
= \int E[Y(g) \mid A = g, X = z] \pi^e(b \mid z) q(z) d(b, z)
\]

\[
= \int E[Y(a) \mid A = a, X = x] \pi^e(a \mid x) q(x) d(a, x)
\]

\[
= \int E[Y \mid A = a, X = x] \pi^e(a \mid x) q(x) d(a, x).
\]

From the first line to the second line, we use a unconfoundedness assumption. From the second line to the third line, we use a transportability assumption. From the third line to the fourth line, we use a unconfoundedness assumption. From the fourth line to the fifth line, the random variables \( a, x \) are replaced with \( b, z \). From the fifth line to the sixth line, we use a consistency assumption.

\[\square\]

From this lemma, the DM method can be naturally introduced. Note this is equivalent to a transport formula [Pearl & Bareinboim, 2014 (3.1)] when the evaluation policy is atomic. The G-formula described here is its extension when the evaluation policy is stochastic.

**Theorem 8 (IPWCS).** \( E[Y(B)] = E[r(X) w(A, X) Y] \)

**Proof.**

\[
E[r(X) w(A, X) Y] = E[r(X) w(A, X) E[Y \mid A, X]]
\]

\[
= \int E[Y \mid A = a, X = x] r(x) w(a, x) \pi^b(a \mid x) p(x) d(a, x)
\]

\[
= \int E[Y \mid A = a, X = x] \pi^e(a \mid x) q(x) d(a, x)
\]

\[
= E[Y(B)].
\]

From the third line to the fourth line, we use a Lemma\[\square\]

\[\square\]

**C Density Ratio Estimation**

Here, we introduce the formulation of LSIF. In LSIF, we estimate the density ratio \( r(x) = \frac{q(x)}{p(x)} \) directly. Let \( \mathcal{S} \) be the class of non-negative measurable functions \( s : \mathcal{X} \rightarrow \mathbb{R}^+ \). We consider minimizing the following squared error between \( s \) and \( r \):

\[
E_{p(x)}[(s(x) - r(x))^2] = E_{p(x)}[(r(x))^2] - 2E_{q(x)}[s(z)] + E_{p(x)}[(s(x))^2]. \tag{7}
\]
D.1 Cramér-Rao lower bound

In this section, we discuss the efficiency bound. Depending on the bracketing entropy of $H$, the term $\frac{1}{2}E_{p(x)}[(s(x))^{2}] - E_{q(z)}[s(z)]$ does not affect the result of minimization and we can ignore the term, i.e., the density ratio is estimated through the following minimization problem:

$$s^{*} = \arg \min_{s \in S} \left[ \frac{1}{2}E_{p(x)}[(s(x))^{2}] - E_{q(z)}[s(z)] \right],$$

where $S$ is a hypothesis class of the density ratio. As mentioned above, to minimize the empirical version of (7), we use uLSIF [Sugiyama et al. 2012]. Given a hypothesis class $\mathcal{H}$, we obtain $\hat{r}$ by

$$\hat{r} = \arg \min_{s \in \mathcal{H}} \left[ \frac{1}{2}E_{n_{\text{eval}}}(s(X))^{2} - E_{n_{\text{eval}}}[s(Z)] + \mathcal{R}(s) \right],$$

where $\mathcal{R}$ is a regularization term. For a model of uLSIF, Kanamori et al. [2012] proposed using kernel based hypothesis to estimate the density ratio nonparametrically. Kanamori et al. [2012] called uLSIF with kernel based hypothesis as KuLSIF. Kanamori et al. [2012] showed that, under some assumptions, the convergence rate of KuLSIF is $O(\min (n_{\text{eval}}, n_{\text{target}})^{-\frac{1}{2\gamma}})$, where $0 < \gamma < 2$ is a constant depending on the bracketing entropy of $\mathcal{H}$.

D. Efficiency bound for the stratified sampling mechanism

In this section, we discuss the efficiency bound.

D.1 Cramér-Rao lower bound

First, we show the Cramér-Rao lower bound when the DGP is a stratified sampling with the historical data $\{\alpha_{i}\}_{i=1}^{n_{\text{hist}}}$ and evaluation data $\{\beta_{i}\}_{i=1}^{n_{\text{eval}}}$, where $\alpha_{i}$ and $\beta_{i}$ are random variables. Let $H_{n_{\text{hist}}}$ and $G_{n_{\text{eval}}}$ be the distributions of $\{\alpha_{i}\}_{i=1}^{n_{\text{hist}}}$ and $\{\beta_{i}\}_{i=1}^{n_{\text{eval}}}$. Let us define a set of distributions as $\mathcal{M}_{n} = \{H_{n_{\text{hist}}}, G_{n_{\text{eval}}}\}$. A model $\mathcal{M}_{n_{\text{para}}}$ is called a regular parametric submodel if the model can be written as $\mathcal{M}_{n_{\text{para}}} = \{H_{\theta_{1}, n_{\text{hist}}}, G_{\theta_{2}, n_{\text{eval}}}\}$, where $\theta_{1} \in \Theta_{1}$, $\theta_{2} \in \Theta_{2}$ and it matches the true distribution at $\theta_{1}^{*}$ and $\theta_{2}^{*}$, and it has a density

$$h_{\theta_{1}, n_{\text{hist}}} (\{\alpha_{i}\}) = \prod_{i=1}^{n_{\text{hist}}} h(\alpha_{i}; \theta_{1}), g_{\theta_{2}, n_{\text{eval}}} (\{\beta_{i}\}) = \prod_{i=1}^{n_{\text{eval}}} g(\beta_{i}; \theta_{2}).$$

Let $R(H, G) \rightarrow \mathbb{R}$ be a target functional. Then, the Cramér-Rao lower bound of the functional $R$ under the parametric submodel $\mathcal{M}_{n_{\text{para}}}$ is

$$\text{CR}(\mathcal{M}_{n_{\text{para}}}, R) = \nabla_{\theta_{1}} R(H_{\theta_{1}}, G_{\theta_{2}}) \mathbb{E} [\nabla_{\theta_{1}} \log h_{\theta_{1}, n_{\text{hist}}} ]^{-1} \nabla_{\theta_{1}} R(H_{\theta_{1}}, G_{\theta_{2}}) + \nabla_{\theta_{2}} R(H_{\theta_{1}}, G_{\theta_{2}}) \mathbb{E} [\nabla_{\theta_{2}} \log g_{\theta_{2}, n_{\text{eval}}} ]^{-1} \nabla_{\theta_{2}} R(H_{\theta_{1}}, G_{\theta_{2}}).$$

Before that, we calculate the Cramér-Rao lower bound in a tabular setting, where the state, action and reward spaces are finite.

Theorem 9. In a tabular case, $n\text{CR}(\mathcal{M}_{n_{\text{para}}}, R)$ is

$$\rho^{-1} \mathbb{E}[\nabla^{2} (X)w^{2}(A, X) \text{var}[Y | A, X]] + (1 - \rho)^{-1} \text{var}[v(Z)].$$

Proof of Theorem 9: In our setting, we have $\{X_{i}, A_{i}, Y_{i}\}_{i=1}^{n_{\text{hist}}}$ and $\{Z_{j}\}_{j=1}^{n_{\text{eval}}}$. The target functional, i.e., the value of the evaluation policy $\pi^{e}$ defined in (1) is

$$R(\pi^{e}) = \int q(y(x); \pi^{e}(a | x))p(y | a, x)d\mu(a, x, y),$$

where $\mu$ is a baseline measure such as Lebesgue or counting measure. The scaled Cramér-Rao lower bound for regular parametric models under $\mathcal{M}_{n_{\text{para}}}$ is given by

$$n\text{CR}(\mathcal{M}_{n_{\text{para}}}, R) = \rho^{-1} A_{1} B_{1}^{-1} A_{1}^{\top} + (1 - \rho)^{-1} A_{2} B_{2}^{-1} A_{2}^{\top},$$

$$A_{1} = \mathbb{E}_{x \sim q(x), a \sim \pi(\alpha | x), y \sim p(y | a, x)} [\nabla_{\theta_{1}} \log p(y | a, x; \theta_{1})],$$

$$A_{2} = \mathbb{E}_{x \sim q(x), a \sim \pi(\alpha | x), y \sim p(y | a, x)} [\nabla_{\theta_{2}} \log q(x; \theta_{2})],$$

$$B_{1} = \mathbb{E}_{x \sim p(x), a \sim \pi(\alpha | x), y \sim p(y | a, x)} [\nabla_{\theta_{1}} \log p(y | a, x; \theta_{1})],$$

$$B_{2} = \mathbb{E}_{x \sim q(x)} [\nabla_{\theta_{2}} \log q(x; \theta_{2})].$$
Then, from the Cauchy Schwartz inequality [Tripathi, 1999], we have the following inequality:
\[
\mathbb{E}[A(Z)B^\top(Z)]\mathbb{E}[B(Z)B^\top(Z)]^{-1}\mathbb{E}[A(Z)B^\top(Z)]^\top \leq \mathbb{E}[A^2(Z)],
\]
where \(\mathbb{E}[A(Z)] = 0, \mathbb{E}[B(Z)] = 0\). Then, we obtain the following upper bound:
\[
A_1B_1^{-1}A_1^\top = \mathbb{E}[r(X)w(A, X)Y\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]\mathbb{E}[\nabla\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]^{-1} \\
\times \mathbb{E}[r(X)w(A, X)Y\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]
\]
\[
= \mathbb{E}[r(x)w(A, X)\{Y - \mathbb{E}[Y \mid A, X]\}\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]\mathbb{E}[\nabla\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]^{-1} \\
\times \mathbb{E}[r(x)w(A, X)\{Y - \mathbb{E}[Y \mid A, X]\}\nabla_{\theta_1} \log p(Y \mid A, X; \theta_1)]
\]
\[
\leq \mathbb{E}[r^2(X)w^2(A, X)\{Y - \mathbb{E}[Y \mid A, X]\}^2] = \mathbb{E}[r^2(X)w^2(A, X)\text{var}[Y \mid A, X]].
\]

In the same way,
\[
A_2B_2^{-1}A_2^\top = \mathbb{E}[v(Z)\nabla_{\theta_2} \log q(Z; \theta_2)]\mathbb{E}[\nabla\nabla_{\theta_2} \log q(Z; \theta_2)]^{-1} \mathbb{E}[v(Z)\nabla_{\theta_2} \log q(z; \theta_2)]
\]
\[
= \mathbb{E}[\{v(Z) - \mathbb{E}[v(Z)]\}\nabla_{\theta_2} \log q(Z; \theta_2)]\mathbb{E}[\nabla\nabla_{\theta_2} \log q(Z; \theta_2)]^{-1} \mathbb{E}[\{v(Z) - \mathbb{E}[v(Z)]\}\nabla_{\theta_2} \log q(Z; \theta_2)]
\]
\[
\leq \mathbb{E}[\{v(Z) - \mathbb{E}[v(Z)]\}^2] = \text{var}[v(Z)].
\]

Therefore,
\[
\rho^{-1}A_1B_1^{-1}A_1^\top + (1 - \rho)^{-1}A_2B_2^{-1}A_2^\top \leq \rho^{-1}\mathbb{E}[r^2(X)w^2(A, X)\text{var}[Y \mid A, X]] + (1 - \rho)^{-1}\text{var}[v(Z)].
\]

Finally, we have to show this inequality is equality. This is obvious because our setting is tabular.

\[\square\]

D.2 Reduction to i.i.d setting

The Cramér-Rao lower bound we have seen so far can be extended when the model is semiparametric. However, because our DGP is not i.i.d, we cannot directly apply standard semiparametric theory here. To circumvent this problem, we regard the whole \(n\) data at hand as one observation and consider the case where we observe \(m\) observations. Then, as \(m\) goes to infinity, the total data size \(n' := nm\) goes to infinity. Because each one observation (\(n\) data) is i.i.d, we can apply standard semiparametric theory. Only in this section, we regard \(n'\) as the total hypothetical sample size when discussing asymptotics. The value \(n\) is the sample size of the actual data at hand, which is fixed.

We explain the definition of the efficient influence function (EIF). This is a function for one observation
\[
o = (x_1, \cdots, x_{n\text{hat}}, a_1, \cdots, a_{n\text{hat}}, y_1, \cdots, y_{n\text{hat}}, z_1, \cdots, z_{n\text{eval}}).
\]
This is defined given the target functional and the model. In our context, the EIF has the following property.

**Theorem 10.** [van der Vaart, 1998] Theorem 25.20) The EIF \(\phi(o)\) is the gradient of \(R(\pi^e)\) \(w.r.t\) the model \(M_n\), which has the smallest \(L_2\)-norm. It satisfies that for any regular estimator \(\hat{R}\) of \(R(\pi^e)\) \(w.r.t\) the model \(M_n\), \(\text{AMSE}[^{\hat{R}}] \geq \text{var}[\phi(o)]\), where \(\text{AMSE}[^{\hat{R}}]\) is the second moment of the limiting distribution of \(\sqrt{n}(\hat{R} - R(\pi^e))\).

This states that \(\text{var}[\phi(o)]\) is the lower bound in estimating \(R(\pi^e)\). We call \(n\text{var}[\phi(o)]\) the efficiency bound because what we want to consider is the lower bound of \(\sqrt{n}(\hat{R} - R(\pi^e))\). Note that \(n\) is fixed here.

For the current case, the EIF and efficiency bound are explicitly calculated as follows.

**Theorem 11.** The EIF of \(R(\pi^e)\) \(w.r.t\) the model \(M_n\) is
\[
\phi(o) = \frac{1}{n_{\text{hat}}} \sum_{i=1}^{n_{\text{hat}}} r(x_i)w(a_i, x_i)\{y_i - q(x_i, a_i)\} + \frac{1}{n_{\text{eval}}} \sum_{j=1}^{n_{\text{eval}}} v(z_j) - R(\pi)
\]

The (scaled) efficiency bound \(n\text{var}[\phi(o)]\) is
\[
\rho^{-1}\mathbb{E}[r^2(X)w^2(A, X)\text{var}[Y \mid A, X]] + (1 - \rho)^{-1}\text{var}[v(Z)].
\]
When assuming the model $\mathcal{M}_n^{\text{fix}}$ where $\pi^b(a|x)$ and $p(x)$ are fixed at true values, we can also show that the EIF and the efficiency bound are the same.

**Proof of Theorem** We follow the following steps.

1. Calculate some gradient (a candidate of EIF) of the target functional $R(\pi^*)$ w.r.t $\mathcal{M}_n$.
2. Calculate the tangent space w.r.t $\mathcal{M}_n$.
3. Show that the candidate of EIF in Step 1 lies in the tangent space. Then, this concludes that a candidate of EIF in Step 1 is actually the EIF.

**Calculation of the gradient** As mentioned, the model $\mathcal{M}_n^{\text{para}}$ for a nonparametric model $\mathcal{M}_n$ is

$$p(o; \theta) = \prod_{i=1}^{n_{\text{hat}}} p(x_i; \theta_x) \pi^b(a_i | x_i; \theta_a) p(y_i | x_i, a_i; \theta_y) \prod_{j=1}^{n_{\text{evl}}} q(z_j; \theta_z),$$

$$\theta = (\theta_x^T, \theta_a^T, \theta_y^T, \theta_z^T)^T, o = \{x_i, a_i, y_i, z_j\}_{i=1}^{n_{\text{hat}}, j=1}^{n_{\text{evl}}}. $$

We define the corresponding gradients:

$$g_x = \nabla_{\theta_x} \log p(x; \theta_x), \quad g_a|x = \nabla_{\theta_a} \log \pi^b(a|x; \theta_a), \quad g_y|a,x = \nabla_{\theta_y} \log p(y|a, x; \theta_y), \quad q_z = \nabla_{\theta_z} \log q(z; \theta_z).$$

To derive some gradient of the target functional $R(\pi^*)$ w.r.t $\mathcal{M}_n$, what we need is finding a function $f(o)$ satisfying

$$\nabla R(\theta) = \mathbb{E}[f(D) \nabla \log p(D; \theta)]$$

$$= \mathbb{E} \left[ f(D) \left\{ \frac{1}{n_{\text{hat}}} \sum_{i=1}^{n_{\text{hat}}} \{g_x(X_i) + g_a|x(X_i, A_i) + g_y|a,x(X_i, A_i, Y_i)\} + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} g_z(Z_j) \right\} \right].$$

We take the derivative as follows:

$$\nabla R(\theta) = \mathbb{E}_{p(x|z)p(y|a,x)} \left[ y \left\{ g_z(x) + g_y|a,x(y|a, x) \right\} \right].$$

By some algebra, this is equal to

$$\mathbb{E} \left[ \left\{ \frac{1}{n_{\text{hat}}} \sum_{i=1}^{n_{\text{hat}}} r(X_i) w(A_i, X_i) \{Y_i - q(X_i, A_i)\} + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} v(Z_j) - R(\pi) \right\} \nabla \log p(D; \theta) \right].$$

Thus, the following function

$$\phi(o) = \frac{1}{n_{\text{hat}}} \sum_{i=1}^{n_{\text{hat}}} r(x_i) w(a_i, x_i) \{y_i - q(x_i, a_i)\} + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} v(z_j) - R(\pi)$$

is a derivative.

**Calculation of the tangent space** Following a standard derivation way (Tsiatis, 2006; van der Vaart, 1998), the tangent space of the model $\mathcal{M}_n$ is

$$\left\{ \frac{1}{n_{\text{hat}}} \sum_{i=1}^{n_{\text{hat}}} \{t_x(x_i) + t_a|x(x_i, a_i) + t_y|a,x(x_i, a_i, y_i)\} + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} t_z(z_j) \in L_2(o) \right\}.$$

where $L_2(o)$ is an $L_2$ space at the true density,

$$\mathbb{E}[t_x(X)] = 0, \mathbb{E}[t_a|x(X, A)|X] = 0, \mathbb{E}[t_y|a,x(X, A, Y)|X, A] = 0, \mathbb{E}[t_z(Z)] = 0.$$
Last Part We can easily check that \( \phi(o) \) lies in the tangent space by taking
\[
t_x = 0, t_{a|x} = 0, t_{y|a,x} = r(x)w(a, x)\{y - q(a, x)\}, t_z(z) = v(z) - R(\pi).
\]
Thus, \( \phi(o) \) is the EIF.

Remark 6. We can easily see that the EIF is \( \phi(o) \) when assuming the model \( \mathcal{M}^\text{fix}_n \) where \( p(x) \) and \( \pi^b(a|x) \) are fixed at true values. This model is represented as
\[
\left\{ \prod_{i=1}^{n_{\text{fix}}} p_*(x_i) \pi^b_{\text{fix}}(a_i | x_i) p(y_i | x_i, a_i; \theta_y) \prod_{j=1}^{n_{\text{evl}}} q(z_j; \theta_z) \right\}
\]
where \( * \) emphasizes that these are fixed at true densities.

The function \( \phi(o) \) in the proof is still a gradient of \( R(\pi^o) \) w.r.t \( \mathcal{M}^\text{fix}_n \) because the model \( \mathcal{M}^\text{fix}_n \) is smaller than the model \( \mathcal{M}_n \). Besides, \( \phi(o) \) belongs to the tangent space induced by the model \( \mathcal{M}^\text{fix}_n \) because the tangent space induced by \( \mathcal{M}^\text{fix}_n \) is
\[
\left\{ \frac{1}{n_{\text{fix}}} \sum_{i=1}^{n_{\text{fix}}} \{t_{y|a,x}(x_i, a_i, y_i)\} + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} t_z(z_j) \in L_2(o) \right\}
\]
where
\[
\mathbb{E}[t_{y|a,x}(X, A, Y)|X, A] = 0, \mathbb{E}[t_z(Z)] = 0.
\]
We can easily check that \( \phi(o) \) lies in the above tangent space by taking
\[
t_{y|a,x} = r(x)w(a, x)\{y - q(a, x)\}, t_z(z) = v(z) - R(\pi).
\]
Thus, \( \phi(o) \) is the EIF regarding \( \mathcal{M}^\text{fix}_n \).

E Proofs

In this section, we show the proofs of theorems. In the proofs of Theorem [2][14] we prove the case where we use a two-fold cross-fitting. The extension of two fold cross-fitting to the general \( K \)-fold cross-fold is straightforward.

E.1 Required conditions

In order to show Theorems [6][13], we use the following Theorem [12] which shows the convergence rate of kernel regression. Here, we have data \( \{B_i, C_i\}_{i=1}^n \), which are i.i.d. from \( p(b, c) = p(c|b)p(b) \), and \( B_i \) takes a value in \( \mathcal{B} \). Then, let us consider a kernel estimation:
\[
n^{-1} \sum_{i=1}^n K_h(B_i - b) C_i,
\]
where \( K_h(b) = h^{-d}K(b/h^d) \), where \( d \) is a dimension of \( b \). Then, we have the following theorem following [Newey & McFadden](1994).

Theorem 12. Assume
- the space \( \mathcal{B} \) is compact and \( p(b) > 0 \) on \( \mathcal{B} \).
- the kernel \( K(u) \) has the bounded derivative of order \( k \), satisfies \( \int K(u)du = 1 \), and has zero moments of order \( \leq m - 1 \) and a nonzero \( m \)-th order moment,
- \( \mathbb{E}[C | B = b] \) is continuously differentiable to order \( k \) with bounded derivatives on the opening set in \( \mathcal{B} \).
- there is \( v \geq 4 \) such that \( \mathbb{E}||C|^v| B = b|p(b) \) is bounded.

Then, when \( h = h(n) \) and \( h(n) \to 0 \),
\[
n^{-1} \sum_{i=1}^n K_h(B_i - b) C_i - p(b)\mathbb{E}[C|b]| = O_p \left( \frac{\log n}{nh^{d+2k})^{1/2} + h^m} \right).
\]

Then, under \( n^{-1/2}\sqrt{h/d}\log n \to \infty, \sqrt{nh^{d+2k}} \to \infty, \sqrt{nh^{2m}} \to 0 \), the above \( l_\infty \) risk is \( o_p(n^{-1/2}) \) [Newey & McFadden](1994).
Additional assumptions: regarding Theorem [12] we assume the following assumptions when we prove Theorems [13–6]:

**Theorem [13]:** condition when replacing $B$ with $X$, $C$ with $w(A, X)Y$, condition when replacing $B$ with $Z$, $C$ with 1.

**Theorem [5]:** condition when replacing $B$ with $X$, $C$ with $w(A, X)Y$, condition when replacing $B$ with $X$, $C$ with 1, condition when replacing $B$ with $Z$, $C$ with 1.

**Theorem [6]:** condition when replacing $B$ with $(X, A)$, $C$ with $Y$, condition when replacing $B$ with $(X, A)$, $C$ with $1$.

### E.2 Warming up

As a warm up, first, we prove the asymptotic property of some simple estimator. When $p(x)$ and $\pi^b(a \mid x)$ are known, let us define an IPW estimator:

$$
\hat{R}_{\text{IPW}1}(\pi^e) = \mathbb{E}_{\text{bhat}} \left[ \frac{\hat{q}(X) \pi^e(A \mid X)Y}{p(X) \pi^b(A \mid X)} \right].
$$

**Theorem 13.** When $\hat{q}(x) = \hat{q}_h(x)$, the asymptotic MSE of $\hat{R}_{\text{IPW}1}$ is

$$
\rho^{-1} \text{var}[r(X)w(A, X)Y] + (1 - \rho)^{-1} \text{var}[v(Z)].
$$

**Proof of Theorem [13]** We follow the proof of [Newey & McFadden 1994]. For the ease of notation, assume $\rho = k_1/(k_1 + k_2)$. In this case, $n_{\text{bhat}} = k_1 N_o$ and $n_{\text{evl}} = k_2 N_o$, where $N_o = n/(k_1 + k_2)$.

Note that in this asymptotic regime, $N_o \to \infty$. Therefore, we reindex the sample set as

$$
\{X_i\}_{i=1}^{n_{\text{bhat}}} = \{X_{b,i}\} \ (1 \leq b \leq k_1, \ 1 \leq i \leq N_o),
$$

$$
\{Z_j\}_{j=1}^{n_{\text{evl}}} = \{Z_{c,j}\} \ (1 \leq c \leq k_2, \ 1 \leq j \leq N_o).
$$

Here, we only consider the estimator $\hat{R}_{\text{IPW}1}(\pi^e)$ based on based on $\{X_{b,i}\}_{i=1}^{N_o}$ and $\{Z_{c,j}\}_{j=1}^{N_o}$, and denote it as $\hat{R}_{b,c}$. Then, the final estimator $\hat{R}_{\text{IPW}1}(\pi^e)$ using all set of samples is equal to

$$
\frac{1}{k_1 k_2} \sum_{b=1}^{k_1} \sum_{c=1}^{k_2} \hat{R}_{b,c},
$$

because the kernel estimator has a linear property. More specifically, we have

$$
\hat{R}_{\text{IPW}1}(\pi^e) = \frac{1}{n_{\text{bhat}}} \sum_{i=1}^{n_{\text{bhat}}} \left\{ \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} K_h(Z_j - X_i) \right\} \frac{\pi^e(A_i \mid X_i)Y_i}{\pi^b(A_i \mid X_i)p(X_i)}
$$

$$
= \frac{1}{n_{\text{bhat}} n_{\text{evl}}} \sum_{i=1}^{n_{\text{bhat}}} \sum_{j=1}^{n_{\text{evl}}} K_h(Z_j - X_i) \frac{\pi^e(A_i \mid X_i)Y_i}{\pi^b(A_i \mid X_i)p(X_i)}
$$

$$
= \frac{1}{k_1 k_2} \sum_{b=1}^{k_1} \sum_{c=1}^{k_2} \left\{ \frac{1}{n_{\text{bhat}} n_{\text{evl}}} \sum_{i=1}^{n_{\text{bhat}}} \sum_{j=1}^{n_{\text{evl}}} K_h(Z_{c,j} - X_{b,i}) \frac{\pi^e(A_{b,i} \mid X_{b,i})Y_{b,i}}{\pi^b(A_{b,i} \mid X_{b,i})p(X_{b,i})} \right\}
$$

$$
= \frac{1}{k_1 k_2} \sum_{b=1}^{k_1} \sum_{c=1}^{k_2} \hat{R}_{b,c}.
$$

First, we analyze $\hat{R}_{1,1}$. 

20
Step 1 We prove the following in this step:

\[
\hat{R}_{b,c} = \frac{1}{2d} \sum_{i=1}^{N_i} r(X_{b,i})w(X_{b,i}, A_{b,i})Y_{b,i} + \frac{1}{2d} \sum_{i=1}^{N_i} v(Z_{c,i}) + o_p(n^{-1/2}).
\]

Especially, we prove the statement for \(\hat{R}_{1,1}\) when \(k_1 = 1, k_2 = 1, n^{\text{hst}} = n^{\text{evl}} = n/2\). We have

\[
\hat{R}_{1,1} = \mathbb{E}_{n^{\text{hst}}} \left[ \frac{q_h(X)}{p(X)} \pi^e(A \mid X)^Y \pi^b(A \mid X) \right]
= \frac{1}{n^{\text{hst}}} \sum_{i=1}^{n^{\text{hst}}} \frac{q_h(X)}{p(X)} \pi^b(A_i \mid X_i) \left\{ \frac{1}{n^{\text{evl}}} \sum_{j=1}^{n^{\text{evl}}} \mathbb{E}_{h}(Z_j - X_i) \right\}
= \mathbb{E}_{n^{\text{hst}}} \left[ \frac{q(X)}{p(X)} \pi^e(A \mid X)^Y \pi^b(A \mid X) \right] + \frac{1}{n^{\text{hst}}n^{\text{evl}}} \sum_{i=1}^{n^{\text{hst}}} \sum_{j=1}^{n^{\text{evl}}} a_{i,j}
= \mathbb{E}_{n^{\text{hst}}} \left[ \frac{q(X)}{p(X)} \pi^e(A \mid X)^Y \pi^b(A \mid X) \right] + \frac{2}{n^{\text{hst}}n^{\text{evl}}} \sum_{i<j} b_{i,j},
\]

where

\[
a_{i,j}((X_i, A_i, Y_i), (Z_j)) = \frac{1}{p(X_i)} \pi^b(A_i \mid X_i) \left\{ \mathbb{E}_{h}(Z_j - X_i) - q(X_i) \right\},
\]

\[
b_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j)) = 0.5(a_{i,j} + a_{j,i}).
\]

Then,

\[
\frac{2}{n^{\text{hst}}n^{\text{evl}}} \sum_{i<j} b_{i,j}(X_i, A_i, Y_i, Z_i, (X_j, A_j, Y_j, Z_j))
= \frac{2}{n^{\text{hst}}} \left\{ \sum_{i=1}^{n^{\text{hst}}} \mathbb{E}[b_{i,j} \mid X_i, A_i, Y_i, Z_i] \right\} + o_p(n^{-1/2})
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] + \frac{1}{n^{\text{hst}}} \sum_{i=1}^{n^{\text{hst}}} \mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] + o_p(n^{-1/2})
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \{v(Z_i) - R(\pi^e)\} + o_p(n^{-1/2}).
\]

From the first line to the second line, we used the U-statistics theory [van der Vaart, 1998, Theorem 12.3]. From the third line to the fourth line, based on Theorem [12] we used

\[
\mathbb{E}[a_{j,i} \mid X_i, A_i, Y_i, Z_i] = o_p(n^{-1/2}) + \mathbb{E}[w(A_i, X_i)Y_i \mid X_i = Z_i] \frac{p(X_i)}{p(X_i)} + \mathbb{E}[r(X_i)w(A_i, X_i)Y_i]
= o_p(n^{-1/2}) + v(Z_i) - R(\pi^e),
\]

\[
\mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] = o_p(n^{-1/2}) + \frac{1}{p(X_i)} \pi^b(A_i \mid X_i) \{q(X_i) - q(X_i)\}
= o_p(n^{-1/2}).
\]

Remark 7. \(\mathbb{E}[h(A_i, X_i, Y_i) \mid X_i = Z_i]\) is an abbreviation of \(\{\mathbb{E}[h(A_i, X_i, Y_i) \mid X_i = x]\}_{x = Z_i} \).

Therefore,

\[
\hat{R}_{\text{IPWCSB}} = \mathbb{E}_{n^{\text{hst}}} \left[ \frac{q(X)}{p(X)} \pi^e(A \mid X)^Y \pi^b(A \mid X) \right] + \mathbb{E}_{n^{\text{evl}}}[v(Z)] - R(\pi^e) + o_p(n^{-1/2}).
\]
Step 2  Based on Step 1, we have

\[
\hat{R}_{IPW1} = \frac{1}{k_1k_2} \sum_{b=1}^{k_1} \sum_{c=1}^{k_2} \hat{R}_{b,c}
\]

\[
= \frac{1}{k_1k_2} \sum_{b=1}^{k_1} \sum_{c=1}^{k_2} \left[ \frac{1}{N_o} \sum_{i=1}^{N_o} r(X_{b,i})w(X_{b,i}, A_{b,i})Y_{b,i} + \frac{1}{N_o} \sum_{j=1}^{N_o} v(Z_{c,j}) \right] - R(\pi^o) + o_p(n^{-1/2})
\]

\[
= \frac{1}{k_1N_o} \sum_{b=1}^{k_1} \sum_{i=1}^{N_o} r(X_{b,i})w(X_{b,i}, A_{b,i})Y_{b,i} + \frac{1}{k_2N_o} \sum_{c=1}^{k_2} \sum_{j=1}^{N_o} v(Z_{c,j}) - R(\pi^o) + o_p(n^{-1/2})
\]

\[
= \frac{1}{n} \sum_{i=1}^{n_{\text{hst}}} r(X_i)w(X_i, A_i)Y_i + \frac{1}{n_{\text{evl}}} \sum_{j=1}^{n_{\text{evl}}} v(Z_j) - R(\pi^o) + o_p(n^{-1/2}).
\]

Finally, from stratified sampling CLT, the statement is concluded. \(\square\)

E.3  Proof of Theorem 2

Proof. We denote

\[
\phi_1(x, a, y; r, w, f) = r(x)w(a, x)\{y - f(a, x)\}, \quad \phi_2(z; f) = v(z).
\]

We also denote the union of \(n_{\text{hst}}^1\) and \(n_{\text{evl}}^1\) as \(D_i\) for \(i = 1, 2\), and the number of \(n_{\text{hst}}^1, n_{\text{hst}}^2, n_{\text{evl}}^1, n_{\text{evl}}^2\) as \(n_{11}, n_{21}, n_{12}, n_{22}\). For simplicity, we assume \(n_{11} = n_{12}, n_{21} = n_{22}\).

Then, we have

\[
\sqrt{n} \left\{ E_{n_{\text{hst}}}^1[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)})] + E_{n_{\text{evl}}}^1[\phi_2(Z; \hat{f}^{(1)})] - R(\pi^o) \right\}
\]

\[
= \sqrt{n} \left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{\text{hst}}}^1[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) - \phi_1(X, A, Y; r, w, f)] + \frac{1}{\sqrt{n_{12}}} G_{n_{\text{evl}}}^1[\phi_2(Z; \hat{f}^{(1)}) - \phi_2(Z; f)] \right\}
\]

\[
+ \sqrt{n} \left\{ E[\phi_1(X, A, Y; r, w, f)] - E[\phi_2(Z; f)] \right\}
\]

\[
= \sqrt{n} \left\{ \sum_{i=1}^{n_{\text{hst}}} \phi_1(X, A, Y; r, w, f) + \sum_{j=1}^{n_{\text{evl}}} \phi_2(Z; f) - R(\pi^o) \right\}.
\]

The term (11) is \(o_p(1)\) by Step 1. The term Eq. (12) is also \(o_p(1)\) by Step 2 as follows.

Step 1:  Eq. (11) is \(o_p(1)\).

If we can show that for any \(\epsilon > 0\),

\[
\lim_{n \to \infty} P[|\sqrt{n} \left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{\text{hst}}}^1[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) - \phi_1(X, A, Y; r, w, f)]
\right] + \frac{1}{\sqrt{n_{12}}} G_{n_{\text{evl}}}^1[\phi_2(Z; \hat{f}^{(1)}) - \phi_2(Z; f)]| > \epsilon | D_2] = 0,
\]

then by the bounded convergence theorem, we would have

\[
\lim_{n \to \infty} P[|\sqrt{n} \left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{\text{hst}}}^1[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) - \phi_1(X, A, Y; r, w, f)]
\right] + \frac{1}{\sqrt{n_{12}}} G_{n_{\text{evl}}}^1[\phi_2(Z; \hat{f}^{(1)}) - \phi_2(Z; f)]| > \epsilon] = 0,
\]

yielding the statement.

To show (14), we show that the conditional mean is 0 and the conditional variance is \(o_p(1)\). Then, (12) is proved by the Chebyshev inequality following the proof of [Kallus & Uehara, 2020].
4). The conditional mean is
\[
\mathbb{E}\left[\sqrt{n}\left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{11}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) - \phi_1(X, A, Y; r, w, f)] + \frac{1}{\sqrt{n_{12}}} G_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \right\} \mid \mathcal{D}_2 \right]\]
\[
= \mathbb{E}\left[\sqrt{n}\left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{11}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) - \phi_1(X, A, Y; r, w, f)] + \frac{1}{\sqrt{n_{12}}} G_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \right\} \mid \mathcal{D}_2, \hat{r}(1), \hat{w}(1), \hat{f}(1) \right]\]
\[
= \mathbb{E}\left[\sqrt{n}\left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{11}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) - \phi_1(X, A, Y; r, w, f)] + \frac{1}{\sqrt{n_{12}}} G_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \right\} \mid \mathcal{D}_2, \hat{r}(1), \hat{w}(1), \hat{f}(1) \right]\]
\[
= 0.
\]
Here, we used a cross-fitting construction. More specifically, regarding the second term, we have
\[
\mathbb{E}[\mathbb{E}_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \mid \mathcal{D}_2, \hat{r}(1), \hat{w}(1), \hat{f}(1)]
\]
\[
= \mathbb{E}[\mathbb{E}_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \mid \mathcal{D}_2, \hat{r}(1), \hat{w}(1), \hat{f}(1)] - \mathbb{E}[\mathbb{E}_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \mid \hat{f}(1)]
\]
\[
= \mathbb{E}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f) \mid \hat{f}(1)] - \mathbb{E}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f) \mid \hat{f}(1)] = 0.
\]

The conditional variance is bounded as
\[
\text{var}\left[\sqrt{n}\left\{ \frac{1}{\sqrt{n_{11}}} G_{n_{11}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) - \phi_1(X, A, Y; r, w, f)] \mid \mathcal{D}_2 \right\} \right]
\]
\[
+ \frac{1}{\sqrt{n_{12}}} G_{n_{12}}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f)] \mid \mathcal{D}_2 \right]\]
\[
= \frac{n}{n_{11}} \text{var}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) - \phi_1(X, A, Y; r, w, f)] \mid \mathcal{D}_2 \right]\]
\[
+ \frac{n}{n_{12}} \text{var}[\phi_2(Z; \hat{f}(1)) - \phi_2(Z; f) \mid \mathcal{D}_2 \right]\]
\[
\leq \frac{n}{n_{11}} \mathbb{E}[\{\hat{r}(1)(X)\hat{w}(1)(A, X)(Y - \hat{f}(1)(A, X)) - r(X)w(A, X)(Y - f(A, X))\}^2 \mid \mathcal{D}_2 \right]\]
\[
+ \frac{n}{n_{12}} \mathbb{E}[\{\hat{v}(1)(Z) - v(Z)\}^2 \mid \mathcal{D}_2 \right] = o_p(1) + o_p(1) = o_p(1).
\]

Here, we used
\[
\frac{n}{n_{11}} \mathbb{E}[\{\hat{r}(1)(X)\hat{w}(1)(A, X)(Y - \hat{f}(1)(A, X)) - r(X)w(A, X)(Y - f(A, X))\}^2 \mid \mathcal{D}_2 \right] = o_p(1). \tag{15}
\]
and
\[
\mathbb{E}[\{\hat{v}(1)(Z) - v(Z)\}^2 \mid \mathcal{D}_2 \right] = o_p(1). \tag{16}
\]

The first equation \([15]\) is proved by
\[
\mathbb{E}[\{\hat{r}(1)\hat{w}(1)(Y - \hat{f}(1)) - r w(Y - f)\}^2 \mid \mathcal{D}_2 \right]\]
\[
= \mathbb{E}[\{\hat{r}(1)\hat{w}(1)(Y - \hat{f}(1)) - \hat{r}(1)\hat{w}(1)(Y - f) + \hat{r}(1)\hat{w}(1)(Y - f) - r w(Y - f)\}^2 \mid \mathcal{D}_2 \right]\]
\[
\leq 2\mathbb{E}[\{\hat{r}(1)\hat{w}(1)(Y - \hat{f}(1)) - \hat{r}(1)\hat{w}(1)(Y - f)\}^2 \mid \mathcal{D}_2 \right] + 2\mathbb{E}[\{\hat{r}(1)\hat{w}(1)(Y - f) - r w(Y - f)\}^2 \mid \mathcal{D}_2 \right]\]
\[
\leq 2C_1C_2||f - \hat{f}(1)||_2^2 + 2 \times 4R_{max}^2||\hat{r}(1)\hat{w}(1) - r w||_2^2 = o_p(1).
\]
Here, we have used a parallelogram law from the second line to the third line. We have use 0 < \hat{r} < C_1, 0 < \hat{w} < C_2, ||f|| < R_{max} according to the Assumption 2 and convergence rate conditions, from the third line to the fourth line. The second equation \([16]\) is proved by Jensen’s inequality.

**Step 2:** Eq. \([12]\) is \(o_p(1)\).
We have
\[ |E[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1)) | \hat{f}(1), \hat{w}(1), \hat{j}] + E[\phi_2(Z; \hat{f}(1)) | \hat{j}] - E[\phi_1(x; r, w, f)] - E[\phi_2(Z; f)]| \]
\[ \leq |E[\{\hat{r}(1)(X)\hat{w}(1)(A, X) - r(X)w(A, X)\} (\hat{f}(1)(A, X) + f(A, X)) | \hat{r}(1), \hat{w}(1), \hat{f}(1)|] + |E[\{\hat{r}(1)(X)\hat{w}(1)(A, X)\} (\hat{f}(1)(A, X) - f(A, X)) | \hat{r}(1), \hat{w}(1), \hat{f}(1)|] + |E[\hat{r}(1)(X)\hat{w}(1)(A, X)\} (\hat{f}(1)(A, X) - f(A, X)) | \hat{r}(1), \hat{w}(1), \hat{f}(1)|] \]
\[ \leq |\hat{r}(1)(X)\hat{w}(1)(A, X) - r(X)w(A, X)||\hat{f}(1)(A, X) - f(A, X)||_2^2 + 0 + 0 = \alpha \beta + 0 = o_p(n^{-1/2}). \]
Here, we have used Hölder’s inequality:
\[ \|fg\|_1 \leq \|f\|_2 \|g\|_2, \]
the relation
\[ E[r(X)w(A, X)|\{\hat{r}(1)(A, X) + f(A, X)\} | \hat{r}(1), \hat{w}(1), \hat{f}(1)] + E[\hat{r}(1)(Z) - v(Z) | \hat{f}(1)] \]
\[ = E[-r(X)w(A, X)\hat{f}(1)(A, X) + \hat{v}(1)\{\hat{r}(1), \hat{w}(1), \hat{f}(1)] + E[-r(X)w(A, X)f(A, X) + v(Z)] \]
\[ = 0 + 0 = 0, \]
and
\[ E[\hat{r}(1)(X)\hat{w}(1)(A, X)\} (\hat{f}(1)(A, X) - f(A, X)) | \hat{r}(1), \hat{w}(1), \hat{f}(1)] = 0. \]

Step 3: By combining everything, we have
\[ E_{n^{2\alpha}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1))] + E_{n^{2\alpha}}[\phi_2(Z; \hat{f}(1))] = R(\pi^x) \]
\[ = E_{n^{2\alpha}}[\phi_1(X, A, Y; r, w, f)] + E_{n^{2\alpha}}[\phi_2(Z; f)] = R(\pi^x) + o_p(1/\sqrt{n}). \]
Then,
\[ \hat{R}_{DRCS} = 0.5E_{n^{2\alpha}}[\phi_1(X, A, Y; \hat{r}(1), \hat{w}(1), \hat{f}(1))] + 0.5E_{n^{2\alpha}}[\phi_2(Z; \hat{f}(1))] + 0.5E_{n^{2\alpha}}[\phi_1(X, A, Y; \hat{r}(2), \hat{w}(2), \hat{f}(2))] + 0.5E_{n^{2\alpha}}[\phi_2(Z; \hat{f}(2))] + 0.5E_{n^{2\alpha}}[\phi_1(X, A, Y; r, w, f)] + 0.5E_{n^{2\alpha}}[\phi_2(Z; f)] + 0.5E_{n^{2\alpha}}[\phi_2(Z; f)] + o_p(1/\sqrt{n}) \]
\[ = E_{n^{2\alpha}}[\phi_1(X, A, Y; r, w, f)] + E_{n^{2\alpha}}[\phi_2(Z; f)] + o_p(1/\sqrt{n}). \]
Finally, by using a stratified sampling CLT [Wooldridge 2001], the statement is concluded based on Assumption [1].

E.4 Proof of Lemma [1]

Proof. We can bound \[ \|\hat{r}(X)\hat{w}(A, X) - r(x)w(A, X)\|_2 = o_p(n^{-p}): \]
\[ \|\hat{r}(X)\hat{w}(A, X) - r(x)w(A, X)\|_2 \leq \|\hat{r}(X)\hat{w}(A, X) - \hat{r}(X)w(A, X)\|_2 + \|\hat{r}(X)w(A, X) - r(x)w(A, X)\|_2 \leq C_1o_p(n^{-p}) + C_2o_p(n^{-p}) = o_p(n^{-p}). \]
Here, we used the assumptions that r(X) is uniformly bounded by C_1 and w(A, X) is uniformly bounded by C_2.

E.5 Proof of Theorem [3]

Proof. Let us define \[ \phi_1(x, a, y; r, w, f) \text{ and } \phi_2(z; f): \]
\[ \phi_1(x, a, y; r, w, f) = r(x)w(a, x)\{y - f(a, x)\}, \phi_2(z; f) = v(z). \] (17)
We also denote the union of \( n_1^{\text{hist}} \) and \( n_1^{\text{evl}} \) by \( D_1 \) for \( i = 1, 2 \), and the number of \( n_2^{\text{hist}}, n_1^{\text{evl}}, n_2^{\text{evl}} \) by \( n_{11}, n_{21}, n_{12}, n_{22} \). For simplicity, we assume \( n_{11} = n_{12}, n_{21} = n_{22} \).

Then, we have
\[
\{ \mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)} )] + \mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; \hat{f}^{(1)})] - R(\pi^*) \}
= \left\{ \frac{1}{\sqrt{n_{11}}} \mathbb{G}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) - \phi_1(X, A, Y; r^t, w^t, f^t)] + \frac{1}{\sqrt{n_{12}}} \mathbb{G}_{n_1^{\text{evl}}}[\phi_2(Z; \hat{f}^{(1)}) - \phi_2(Z; f^t)] \right\}
+ \{ \mathbb{E}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) \mid \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}] - \mathbb{E}[\phi_2(Z; \hat{f}^{(1)}) \mid \hat{f}^{(1)}] \}
- \mathbb{E}[\phi_1(X, A, Y; r^t, w^t, f^t)] - \mathbb{E}[\phi_2(Z; f^t)]
+ \{ \mathbb{E}_{n_2^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + \mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; f^t)] - R(\pi^*) \}. \tag{18}
\]

The term \( \boxed{(18)} \) is \( o_p(1/\sqrt{n}) \) by Step 1 in the previous theorem noting that what we have used is \( \| \hat{r}(X) \hat{w}(A, X) - w^t(A, X)r^t(X) \| = o_p(1), \| \hat{f}(A, X) - f^t(A, X) \| = o_p(1) \). The term Eq. (19) is also \( o_p(1) \) by Step 1 as we will show soon.

**Step 1:** Eq. (19) is \( o_p(1) \). We have
\[
\mathbb{E}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}) \mid \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}] + \mathbb{E}[\phi_2(Z; \hat{f}^{(1)}) \mid \hat{f}^{(1)}] - \mathbb{E}[\phi_1(x; r, w, f)] - \mathbb{E}[\phi_2(Z; f)]
\leq \mathbb{E}[\{ \| \hat{r}^{(1)}(X) \| \hat{w}^{(1)}(A, X) - r^t(X)w^t(A, X) \|] - f^t(A, X)] - f^t(A, X) \mid \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}] \}
+ \mathbb{E}[\| \hat{r}^{(1)}(X) \| \hat{w}^{(1)}(A, X) - f^t(A, X)] \mid \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)}] + \mathbb{E}[\| \hat{r}^{(1)}(X) \| \hat{w}^{(1)}(A, X) - f^t(A, X)] \mid \hat{r}^{(1)}, \hat{w}^{(1)}].
\]

Here, if \( f^t(a, x) = f(a, x) \), we have
\[
o_p(1)o_p(1) + o_p(1) + 0 = o_p(1) = o_p(1).
\]
if \( r^t(x)w^t(a, x) = r(x)w(a, x) \), we have
\[
o_p(1)o_p(1) + 0 + o_p(1) = o_p(1) = o_p(1).
\]
Therefore, Eq. (19) is \( o_p(1) \).

**Step 2:** By combining together, we have
\[
\mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)})] + \mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; \hat{f}^{(1)})] - R(\pi^*)
= \mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + \mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; f^t)] - R(\pi^*) + o_p(1).
\]

Then,
\[
\hat{R}_{\text{ORCS}} = 0.5\mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; \hat{r}^{(1)}, \hat{w}^{(1)}, \hat{f}^{(1)})] + 0.5\mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; \hat{f}^{(1)})]
+ 0.5\mathbb{E}_{n_2^{\text{hist}}}[\phi_1(X, A, Y; \hat{r}^{(2)}, \hat{w}^{(2)}, \hat{f}^{(2)})] + 0.5\mathbb{E}_{n_2^{\text{evl}}}[\phi_2(Z; \hat{f}^{(2)})]
= 0.5\mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + 0.5\mathbb{E}_{n_1^{\text{evl}}}[\phi_2(Z; f^t)]
+ 0.5\mathbb{E}_{n_2^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + 0.5\mathbb{E}_{n_2^{\text{evl}}}[\phi_2(Z; f^t)] + o_p(1)
= \mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + \mathbb{E}_{n_1^{\text{hist}}}[\phi_2(Z; f^t)] + o_p(1).
\]

Then, the statement is concluded because
\[
\mathbb{E}\left[ \mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + \mathbb{E}_{n_2^{\text{evl}}}[\phi_2(Z; f^t)] \right] = R(\pi^*)
\]

based on the double robust structure and
\[
\mathbb{E}_{n_1^{\text{hist}}}[\phi_1(X, A, Y; r^t, w^t, f^t)] + \mathbb{E}_{n_2^{\text{evl}}}[\phi_2(Z; f^t)] = R(\pi^*) + o_p(1)
\]

from the law of large numbers based on Assumption.
E.6 Proof of Theorem 14

Proof. We can prove similarly as in the proof of Theorem 1. Therefore, we omit the proof.

E.7 Proof of Theorem 14

Proof of Theorem 14. For the ease of notation, we prove the case \( n^{\text{bst}} = n^{\text{evl}} \) noting the kernel estimator is linearized as in Theorem 13 and the generalization is easy. We have

\[
\left\| \frac{\hat{q}(x)}{p(x)} - \frac{q(x)}{p(x)} - \hat{e}_h(x) \right\|_\infty = o_p(n^{-1/2}),
\]

where

\[
\hat{e}_h(x) = \frac{1}{p(x)} \{ \hat{q}(x) - q(x) \} - \frac{q(x)}{p^2(x)} \{ \hat{p}_h(x) - p(x) \}.
\]

This is proved by Theorem 12. Then,

\[
\hat{R}_{\text{IPWCSB}} = \mathbb{E}_{n^{\text{bst}}} \left[ \frac{\hat{q}_b(X) \pi^e(A \mid X) Y}{\hat{p}_b(X) \pi^b(A \mid X)} \right] = \frac{1}{n^{\text{bst}}} \sum_{i=1}^{n^{\text{bst}}} \pi^e(A_i \mid X_i) Y_i \{ r(X_i) + \hat{e}_h(X_i) \}
\]

\[
= \mathbb{E}_{n^{\text{bst}}} \left[ \frac{q(X) \pi^e(A \mid X) Y}{p(X) \pi^b(A \mid X)} \right] + \frac{1}{n^{\text{bst}} n^{\text{evl}}} \sum_{i=1}^{n^{\text{bst}}} \sum_{j=1}^{n^{\text{evl}}} a_{i,j} + \frac{2}{n^{\text{bst}} n^{\text{evl}}} \sum_{i<j} b_{i,j},
\]

where

\[
a_{i,j}((X_i, A_i, Y_i), (Z_j)) = \frac{1}{p(X_i)} \frac{\pi^e(A_i \mid X_i) Y_i}{\pi^b(A_i \mid X_i)} \{ K_h(Z_j - X_i) - q(X_i) \}
\]

\[
- \frac{q(X_i) \pi^e(A_i \mid X_i) Y_i}{p^2(X_i) \pi^b(A_i \mid X_i)} \{ K_h(X_j - X_i) - p(X_i) \},
\]

\[
b_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j)) = 0.5 \{ a_{i,j} + a_{j,i} \}.
\]

Then,

\[
\frac{2}{n^{\text{bst}} n^{\text{evl}}} \sum_{i<j} b_{i,j} = \mathbb{E}_{n^{\text{bst}}} \left[ \sum_{i=1}^{n^{\text{bst}}} \mathbb{E}[b_{i,j} \mid X_i, A_i, Y_i, Z_i] \right] + o_p(n^{-1/2})
\]

\[
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \mathbb{E}[a_{i,i} \mid X_i, A_i, Y_i, Z_i] + \frac{1}{n^{\text{bst}}} \sum_{i=1}^{n^{\text{bst}}} \mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] + o_p(n^{-1/2})
\]

\[
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \{ v(Z_i) - r(X_i) v(X_i) \} + o_p(n^{-1/2}).
\]

From the first line to the second line, we have used a U-statistics theory [van der Vaart, 1998, Chapter 12]. From the third line to the fourth line, we have used

\[
\mathbb{E}[a_{i,i} \mid Z_i, X_i, A_i, Y_i] = o_p(n^{-1/2}) + \left\{ \mathbb{E} \left[ \frac{1}{p(X_i)} \frac{\pi^e(A_i \mid X_i) Y_i}{\pi^b(A_i \mid X_i)} \mid X_i = Z_i \right] - \mathbb{E} \left[ \frac{q(X_i) \pi^e(A_i \mid X_i) Y_i}{p^2(X_i) \pi^b(A_i \mid X_i)} \mid X_i \right] \right\} p(X_i)
\]

\[
= o_p(n^{-1/2}) + v(Z_i) - r(X_i) v(X_i),
\]

\[
\mathbb{E}[a_{i,j} \mid Z_i, X_i, A_i, Y_i] = o_p(n^{-1/2}).
\]
Therefore,
\[
\hat{R}_{IPWCSB} = \mathbb{E}_{n^{\text{hast}}} \left[ \frac{q(X)}{p(X)} \left\{ \pi^e(A \mid X)Y - \pi^h(A \mid X) - v(X) \right\} \right] + \mathbb{E}_{n^{\text{evl}}} [v(Z)] + o_p(n^{-1/2}).
\]

The final statement is concluded by CLT.

E.8 Proof of Theorem 5

Proof. For the ease of the notation, we prove the case \(n^{\text{hast}} = n^{\text{evl}}\) noting the kernel estimator is linearized as in Theorem 12 and the generalization is easy. Here, we concatenate \(X\) and \(A\) as \(D\). We also write \(p(x)\pi^h(a \mid x)\) as \(u(d)\).

\[
\left\| \hat{q}_n(x) - q(x) - \hat{\epsilon}_h(d) \right\|_{\infty} = o(n^{-1/2}),
\]

where
\[
\hat{\epsilon}_h(d) = \frac{1}{u(d)} \{ \hat{q}_n(x) - q(x) \} - \frac{q(x)}{u(x)} (\hat{u}_h(d) - u(d)).
\]

This is proved by Theorem 12. Then,
\[
\hat{R}_{IPWCS} = \mathbb{E}_{n^{\text{hast}}} \left[ \frac{\hat{q}_n(X) \pi^e(A \mid X)Y}{\hat{p}_n(X) \pi^e_h(A \mid X)} \right]
\]
\[
= \frac{1}{n^{\text{hast}}} \sum_{i=1}^{n^{\text{hast}}} \pi^e(A_i \mid X_i)Y_i \left\{ \frac{\hat{q}_n(X_i)}{u(D_i)} + \hat{\epsilon}_h(X_i) \right\}
\]
\[
= \mathbb{E}_{n^{\text{hast}}} \left[ \frac{q(X) \pi^e(A \mid X)Y}{p(X) \pi^h(A \mid X)} \right] + \frac{1}{n^{\text{hast}}n^{\text{evl}}} \sum_{i=1}^{n^{\text{hast}}} \sum_{j=1}^{n^{\text{evl}}} a_{i,j}
\]
\[
= \mathbb{E}_{n^{\text{hast}}} \left[ \frac{q(X) \pi^e(A \mid X)Y}{p(X) \pi^h(A \mid X)} \right] + \frac{2}{n^{\text{hast}}n^{\text{evl}}} \sum_{i<j} b_{i,j},
\]

where
\[
a_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j)) = \frac{1}{p(X_i)} \pi^e(A_i \mid X_i)Y_i \{(K_h(Z_j - X_i) - q(X_i))
\]
\[
- \frac{q(X_i)}{u^2(D_i)} \pi^h(A_i \mid X_i)Y_i \{K_h(D_j - D_i) - u(D_i)}
\]
\[
b_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j)) = 0.5\{a_{i,j} + a_{j,i}\}.
\]

Then,
\[
\frac{2}{n^{\text{hast}}n^{\text{evl}}} \sum_{i<j} b_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j))
\]
\[
= \frac{2}{n^{\text{hast}}} \left\{ \mathbb{E}_{n^{\text{hast}}} \left[ a_{i,j} \mid X_i, A_i, Y_i, Z_i \right] \right\} + o_p(n^{-1/2})
\]
\[
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] + \frac{1}{n^{\text{hast}}} \sum_{i=1}^{n^{\text{hast}}} \mathbb{E}[a_{i,j} \mid X_i, A_i, Y_i, Z_i] + o_p(n^{-1/2})
\]
\[
= \frac{1}{n^{\text{evl}}} \sum_{i=1}^{n^{\text{evl}}} \{v(Z_i) - r(X_i)w(X_i, A_i)f(D_i)\} + o_p(n^{-1/2}).
\]

From the first line to the second line, we used the U-statistics theory (van der Vaart [1998] Chapter 12). From the third line to the fourth line, we used
\[
\mathbb{E}[a_{i,j} \mid Z_i, X_i, A_i, Y_i]
\]
\[
= o_p(n^{-1/2}) + \mathbb{E} \left[ \frac{1}{p(X_i)} \pi^e(A_i \mid X_i)Y_i \mid X_i = Z_i \right] p(X_i) - \mathbb{E} \left[ \frac{q(X_i)}{u^2(D_i)} \pi^h(A_i \mid X_i)Y_i \mid D_i = D_i \right] u(D_i)
\]
\[
= o_p(n^{-1/2}) + v(Z_i) - r(X_i)w(X_i, A_i)f(D_i),
\]
\[
\mathbb{E}[a_{i,j} \mid Z_i, X_i, A_i, Y_i] = o_p(n^{-1/2}).
\]
Therefore,
\[
\hat{R}_{HPS}^{\text{WCS}} = E_{n_{\text{hst}}} \left[ \frac{q(X)}{p(X)} \pi^n(A | X) \{ Y - f(A, X) \} \right] + E_{n_{\text{evl}}} [v(Z)] + o_p(n^{-1/2}).
\]

The final statement is concluded by CLT. 

E.9 Proof of Theorem 6

Proof. For the ease of the notation, we prove the case \( n_{\text{hst}} = n_{\text{evl}} \) noting the kernel estimator is linearized as in Theorem 13 and and generalization is easy. Here, \( \hat{v}_h(a, x) \) is defined as

\[
\hat{v}_h(a, x) = \frac{\hat{p}_h(a, x)}{\hat{u}_h(a, x)}, \quad \hat{p}_h(a, x) = \frac{1}{n_{\text{hst}}} \sum_{i=1}^{n_{\text{hst}}} Y_i K_h(\{X_i, A_i\} - \{x, a\})
\]

\[
\hat{u}_h(a, x) = \frac{1}{n_{\text{hst}}} \sum_{i=1}^{n_{\text{hst}}} K_h(\{X_i, A_i\} - \{x, a\}).
\]

We have

\[
\left\| \frac{\hat{p}_h(a, x)}{\hat{u}_h(a, x)} - \frac{f(a, x)u(a, x)}{u(a, x)} - \hat{e}_h(x, a) \right\|_\infty = o_p(n^{-1/2}),
\]

where

\[
\hat{e}_h(x, a) = \frac{1}{u(a, x)} \{ \hat{p}_h(a, x) - f(a, x)u(a, x) \} - \frac{f(a, x)}{u(a, x)} \{ \hat{u}_h(a, x) - u(a, x) \}.
\]

This is proved by Theorem 12. Then, we have

\[
\hat{R}_{DM} = E_{n_{\text{evl}}} [\hat{v}_h(Z, A)]
\]

\[
= E_{n_{\text{evl}}} [v(Z, A)] + \frac{1}{n_{\text{hst}} n_{\text{evl}}} \sum_{i=1}^{n_{\text{evl}}} \sum_{j=1}^{n_{\text{hst}}} a_{i,j}
\]

\[
= E_{n_{\text{evl}}} [v(Z, A)] + \frac{2}{n_{\text{hst}} n_{\text{evl}}} \sum_{i<j} b_{i,j},
\]

where

\[
a_{i,j}((Z_i, A_i), (X_j, A_j, Y_j)) = \frac{1}{u(Z_i, A_i)} \{ Y_j K_h(\{X_j, A_j\} - \{Z_i, A_i\}) - u(Z_i, A_i) f(Z_i, A_i) \} - 
\]

\[
\frac{f(Z_i, A_i)}{u(Z_i, A_i)} \{ K_h(\{X_j, A_j\} - \{Z_i, A_i\}) - u(Z_i, A_i) \},
\]

\[
b_{i,j}((X_i, A_i, Y_i, Z_i), (X_j, A_j, Y_j, Z_j)) = 0.5 \{ a_{i,j} + a_{j,i} \}.
\]

Then,

\[
\frac{2}{n_{\text{hst}} n_{\text{evl}}} \sum_{i<j} b_{i,j}(X_i, A_i, Y_i, Z_i, (X_j, A_j, Y_j, Z_j))
\]

\[
= \frac{2}{n_{\text{hst}}} \left\{ \sum_{i=1}^{n_{\text{hst}}} E[b_{i,j} | X_i, A_i, Y_i, Z_i] \right\} + o_p(n^{-1/2})
\]

\[
= \frac{1}{n_{\text{evl}}} \sum_{i=1}^{n_{\text{evl}}} E[a_{i,i} | X_i, A_i, Y_i, Z_i] + \frac{1}{n_{\text{hst}}} \sum_{i=1}^{n_{\text{hst}}} E[a_{i,j} | X_i, A_i, Y_i, Z_i] + o_p(n^{-1/2})
\]

\[
= \frac{1}{n_{\text{evl}}} \sum_{i=1}^{n_{\text{evl}}} \{ Y_i - f(X_i, A_i) \} r(X_i) w(X_i, A_i) + o_p(n^{-1/2}).
\]
From the first line to the second line, we used a U-statistics theory \cite{van_der_Vaart_1998} Chapter 12. From the third line to the fourth line, we used
\[ E[a_{i,j} | Z_i, X_i, A_i, Y_i] = \mathcal{O}(n^{-1/2}) + \left\{ \frac{Y_i}{u(X_i, A_i)} - \frac{f(X_i, A_i)}{u(X_i, A_i)} \right\} q(X_i) \pi^c(A_i | X_i) \]
\[ = \mathcal{O}(n^{-1/2}) + Y_i r(X_i) w(A_i, X_i) - r(X_i) f(X_i, A_i) w(A_i, X_i), \]
\[ E[a_{i,j} | Z_i, X_i, A_i, Y_i] = \mathcal{O}(n^{-1/2}). \]
This is proved by Theorem 12. Therefore,
\[ R_{DM} = E_{a \sim b} [r(X) w(A, X) \{ Y - f(A, X) \}] + E_{v \sim v} [v(Z)] + \mathcal{O}(n^{-1/2}). \]
The final statement is concluded by CLT.

**E.10 Proof of Theorem 7**

**Proof.** We prove the statement following Zhou et al. \cite{Zhou_2018}. Though the proof is very similar, for completeness, we sketch the proof the case \( \rho = 0.5 \). Because the estimator is asymptotically linear, the generalization is easy as in Theorem 13.

Define two scores:
\[ \hat{\Gamma}_i = r_w(D_i) (X_i, A_i) \{ Y_i - f(D_i)(X_i, A_i) \} + [f(a_1, X_i), \ldots, f(a_\alpha, X_i)]^\top, \]
\[ \Gamma_i = r_w(X_i, A_i) \{ Y_i - f(X_i, A_i) \} + [f(a_1, X_i), \ldots, f(a_\alpha, X_i)]^\top, \]
where \( D_i \) is an indicator which cross-fold estimator is used and \( \alpha \) is a dimension of the action, \( r_w(x) = r(x) / \pi^b(a | x) \). Then, we have \( R_{DMCS}(\pi) = \frac{2}{n} \sum_{i=1}^{n/2} (\pi(Z_i), \hat{\Gamma}_i) \). Here, we define the estimator with oracle efficient influence function
\[ \hat{R}(\pi) = \frac{2}{n} \sum_{i=1}^{n/2} \langle \pi(Z_i), \hat{\Gamma}_i \rangle. \]

In addition, define
\[ \Delta(\pi_a, \pi_b) = R(\pi_a) - R(\pi_b), \quad \hat{\Delta}(\pi_a, \pi_b) = \hat{R}(\pi_a) - \hat{R}(\pi_b), \]
\[ \Delta(\pi_a, \pi_b) = \hat{R}(\pi_a) - \hat{R}(\pi_b). \]

**Step 1:** First, following Zhou et al. \cite{Zhou_2018} Theorem 2), we prove the following. Let \( \hat{\pi} \in \arg \min_{\pi \in \Pi} \hat{R}(\pi) \). Then for any \( \delta > 0 \), with probability at least \( 1 - 2\delta \),
\[ R(\hat{\pi}) - R(\pi^*) \leq \mathcal{O} \left( \left\{ k(\Pi) + \sqrt{\log(1/\delta)} \right\} \sqrt{\frac{\Upsilon_*}{n}} \right), \]
where
\[ \Upsilon_* = \sup_{\pi \in \Pi} E[(\Gamma_i, \pi(Z_i))^2] \]
\[ = \sup_{\pi \in \Pi} E[r(X_i)^2 w_\pi^2(X_i, A_i) \{ Y_i - f(X_i, A_i) \}^2] + E[E_v(Z_i) \{ f(a, x) \}^2], \]
when \( w_\pi(a, x) = \pi(a, x) / \pi^b(a, x) \), \( v_\pi(x) = E_{\pi(a|x)} [f(a, x) | x] \).

This is proved as follows. We have
\[ R(\hat{\pi}) - R(\pi^*) \leq \sup_{\pi_a, \pi_b \in \Pi} |\hat{\Delta}(\pi_a, \pi_b) - \Delta(\pi_a, \pi_b)|. \]
Then, by using a Chaining argument as Lemma 1 \cite{Zhou_2018}, we can bound an expectation of \( \sup_{\pi_a, \pi_b \in \Pi} |\hat{\Delta}(\pi_a, \pi_b) - \Delta(\pi_a, \pi_b)| \) via Rademacher complexity. Then, as in Lemma 2 \cite{Zhou_2018}, the high probability bound is obtained via Talagrand inequality. Then, we have
\[ R(\hat{\pi}) - R(\pi^*) \leq \mathcal{O} \left( \left\{ k(\Pi) + \sqrt{\log(1/\delta)} \right\} \sqrt{\frac{\Upsilon_*}{n}} \right), \]
where
\[ Y'_* = \sup_{\pi_a, \pi_b \in \Pi} \mathbb{E}[(\Gamma_1, \pi_a(Z_i) - \pi_b(Z_i))^2]. \]

This concludes the above statement because
\[ Y_* = \sup_{\pi_a, \pi_b \in \Pi} \mathbb{E}[(\Gamma_1, \pi_a(Z_i) - \pi_b(Z_i))^2] \]
\[ = \sup_{\pi_a, \pi_b \in \Pi} \mathbb{E}[r(X_i)^2w_{\pi_a} - w_{\pi_b}]^2\{Y_i - f(X_i, A_i)\}^2 + \mathbb{E}\{v_{\pi_a}(Z_i) - v_{\pi_b}(Z_i)\}^2 \]
\[ \leq \sup_{\pi} 2\mathbb{E}[r(X_i)^2w_{\pi}^2(X_i, A_i)\{Y_i - f(X_i, A_i)\}^2 + 2\mathbb{E}\{v_{\pi}(Z_i)\}^2]. \]

Step 2: Assume \( \kappa(\Pi) < \infty \), then
\[ \sup_{\pi_a, \pi_b \in \Pi} |\hat{\Delta}(\pi_a, \pi_b) - \Delta(\pi_a, \pi_b)| = o_p(n^{-1/2}). \]

The proof of this statement is based on the double structure of the influence function and cross-fitting. We omit the proof because it is long, and almost the same as Lemma 3 (Zhou et al., 2018).

Step 3: Finally, based on Theorem 3 (Zhou et al., 2018), we have
\[ R(\hat{\pi}) - R(\pi^*) \leq \sup_{\pi_a, \pi_b \in \Pi} |\hat{\Delta}(\pi_a, \pi_b) - \Delta(\pi_a, \pi_b)| + \sup_{\pi_a, \pi_b \in \Pi} |\tilde{\Delta}(\pi_a, \pi_b) - \Delta(\pi_a, \pi_b)| \]
\[ \leq O_p \left( k(\Pi) + \sqrt{\log(1/\delta)} \right) \frac{Y_*}{n}. \]

This means there exists an integer \( N_\delta \) such that with probability at least \( 1 - 2\delta \), for all \( n \geq N_\delta \):
\[ R(\hat{\pi}) - R(\pi^*) \leq \left( k(\Pi) + \sqrt{\log(1/\delta)} \right) \frac{Y_*}{n}. \]

Remark 8. In the general case,
\[ T_* = \sup_{\pi \in \Pi} \rho^{-1} \mathbb{E}[r(X_i)^2w_{\pi}^2(X_i, A_i)\{Y_i - f(X_i, A_i)\}^2] + (1 - \rho)^{-1} \mathbb{E}\{v_{\pi}(Z_i)\}^2]. \]

F OPE with Known Distribution of Evaluation Data

In this section, we consider a special case where \( q(x) \) is known.

By applying (6) in Section 2.3, we obtain the efficiency bound under nonparametric model defined, which is defined as \( \bar{\pi}(\pi^*) = \mathbb{E}[r^2(X)w^2(A, X)\var[Y | A, X]]. \)

As the estimator \( \hat{R}_{DRCS}(\pi^*) \) in Section 4, we construct an estimator with cross-fitting. Instead of (6), we use an estimator defined as \( \mathbb{E}_{\pi_{\tilde{z}}} [\tilde{r}(k)(X)\tilde{w}(k)(A, X)\{Y - \tilde{f}(k)(A, X)\}] + \mathbb{E}_{q(z)\pi^*(a|z)}[\tilde{f}(k)(a, z)] \). The algorithm is almost the same as before. To estimate \( r(x) \), we can simply use density estimation for \( p(x) \) because \( q(x) \) is known and the integration in \( \mathbb{E}_{q(z)\pi^*(a|z)}[\tilde{f}(k)(a, z)] \) can be taken exactly because \( q(x) \) and \( \pi^*(a | x) \) are known. Let us denote this estimator as \( \hat{R}_{DRCS} \). We can show that \( \hat{R}_{DRCS}(\pi^*) \) achieves the efficiency bound.

Theorem 14 (Efficiency of \( \hat{R}_{DRCS} \)). For \( k \in \{1, \ldots, \xi\} \), assume there exists \( p > 0, q > 0, p+q \geq 1/2 \) such that \( \|\tilde{r}(k)(X)\tilde{w}(k)(A, X) - r(X)w(A, X)\|^2 = o_p(n^{-p}) \) and \( \|\tilde{f}(k)(A, X) - f(A, X)\|^2 = o_p(n^{-q}) \). Then, we have \( \sqrt{n}\text{Var}(\hat{R}_{DRCS}(\pi^*) - R(\pi^*)) \xrightarrow{d} \mathcal{N}(0, \bar{\pi}(\pi^*)). \)

This asymptotic variance is equal to the asymptotic variance when \( \rho = 0 \) as shown in Remark 2 because the case \( \rho = 0 \) implies that we have infinite data from \( q(x) \).
G Algorithm for Off-Policy Learning with Cross-Fitting

In the proposed method of O-PLE under a covariate shift, we train an evaluation policy by using an estimator \( \hat{R}_{\text{DRCS}}(\pi^*) \), which is constructed via cross-fitting. In this section, we introduce an algorithm where we use a linear-in-parameter model with kernel functions to approximate a new policy. For \( x \in \mathcal{X} \), a linear-in-parameter model is defined as
\[
\mathbb{P}(a \mid x; \sigma^2) = \frac{\exp(g(a, x; \sigma^2))}{\sum_{c \in A} \exp(g(c, x; \sigma^2))},
\]
where \( g(a, x; \sigma^2) = \beta_a^T \varphi(x; \sigma^2) + \beta_0(a), \varphi(x; \sigma^2) = [\varphi_1(x; \sigma^2), \ldots, \varphi_m(x; \sigma^2)]^T \), \( \varphi_m(x; \sigma^2) \) is the Gaussian kernel defined as \( \varphi_u(x; \sigma^2) = \exp\left(-\frac{\|x-c_u\|^2}{2\sigma^2}\right) \), 1 \( \leq u \leq m \), where \( \{c_1, \ldots, c_m\} \) is \( m \) chosen points from \( \{X_i\}_{i=1}^{n}\text{test} \), \( \beta_a \in \mathbb{R}^m \), and \( \beta_0(a) \in \mathbb{R} \). In optimization, we put a regularization term \( R(\{\beta_a, \beta_0(a)\}) \) and train a new policy as
\[
\hat{\pi} = \arg\max_{\pi} \hat{R}_{\text{DRCS}}(\pi) + \mathcal{R}(\{\beta, \beta_0(a)\}),
\]
where \( \lambda > 0 \). The parameters \( \sigma^2 \) and \( \lambda \) are hyper-parameters selected via cross-validation. Thus, in the proposed method, we use the cross-fitting and cross-validation. We describe the algorithm in Algorithm 2 with \( K \) fold cross-fitting and \( L \) fold cross-validation. For brevity, in the algorithm, let us assume \( n_{\text{test}}/\xi, n_{\text{test}}/L, n_{\text{eval}}/\xi, n_{\text{eval}}/L \in \mathbb{N} \). In Algorithm 2 we express the objective function with hyper-parameters \( \sigma^2 \) and \( \lambda \) as
\[
\mathbb{E}_{n_{\text{test}}} \left[ \hat{R}(X) \frac{\pi(A|X; \sigma^2)}{\hat{\pi}(A|X)} (Y - \hat{f}(A, X)) \right] + \mathbb{E}_{n_{\text{eval}}} \left[ \mathbb{E}_{r(a|Z)} \left[ \hat{f}(a, Z) \pi(a \mid Z; \sigma^2) \right] + \mathcal{R}(\{\beta_a, \beta_0(a)\}) \right].
\]

H Details of Experiments in Section 7.1

First, we show the description of the datasets in Table. All datasets are downloaded from [https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/](https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/)

In addition to the results shown in Section 7.1, we show the performances of IPWCS and DM estimator with nuisance functions estimated by the kernel Ridge regression, which are referred as IPWCS-R and DM-R. In addition, for OPE, we also show the results with the different sample size.

In Tables 6–7, we show the additional experimental results with the same setting as Section 7. In this setting, the sample size is fixed at 800.

In Tables 8–9, we show the additional experimental results with 500 samples. The other setting is the same as Section 7.1.

In Tables 10–11, we show the additional experimental results with 300 samples. The other setting is the same as Section 7.1.

We also add the OPE and OPL experiment with the pendigits dataset in Tables 12 and 13. In this experiment, the sample size is fixed at 800.

In Tables 14–15, we show the additional experimental results with 1,000 samples for the satimage and pendigits datasets. We could not conduct experiments for the vehicle dataset because it only has 800 samples. The other setting is the same as Section 7.1.

For OPE, we highlight in bold the best two estimators in each case. For OPL, we highlight in bold the best one estimator in each case. The proposed DRCS estimator performs well in many datasets. The DM estimator also works well, but the performance dramatically drops when the model is misspecified.

**Remark 9** (Self-normalization). We can add self-normalization for improving the performances of the proposed estimator [Swaminathan & Joachims (2015a)]. There can be several ways for incorporating the density ratio estimator in the self-normalization. Because the proposition of self-normalization method is not main topic of this paper, we omit the details.
Algorithm 2 Off-policy learning using $\hat{R}_{\text{DRC5}}(\pi^*)$ with $\xi$-fold cross-fitting.

**Input:** $\xi$: the number of the cross-fitting for constructing $\hat{R}_{\text{DRC5}}(\pi^*)$. $L$: the number of the cross-validation for constructing the optimal policy. $\Pi$: a hypothesis class of $\pi^*$. $\{\sigma_1^2, \ldots, \sigma_n^2\}$: candidates of $\sigma^2$. $\{\lambda_1, \ldots, \lambda_n\}$: candidates of $\lambda$.

Take a $\xi$-fold random partition $(I_k)^{\xi}_{k=1}$ of observation indices $[n_{\text{hst}}] = \{1, \ldots, n_{\text{hst}}\}$ such that the size of each fold $I_k$ is $n_{\text{hst}}^k = n_{\text{hst}}/\xi$.

Take a $\xi$-fold random partition $(J_k)^{\xi}_{k=1}$ of observation indices $[n_{\text{evl}}] = \{1, \ldots, n_{\text{evl}}\}$ such that the size of each fold $J_k$ is $n_{\text{evl}}^k = n_{\text{evl}}/\xi$.

For each $k \in [\xi] = \{1, \ldots, \xi\}$, define $I_k := \{1, \ldots, n_{\text{hst}}\} \setminus I_k$ and $J_k := \{1, \ldots, n_{\text{evl}}\} \setminus J_k$.

Define $S_k = \{(X_i, A_i, Y_i)\}_{i \in I_k}$.

for $k \in [K]$ do

Construct nuisance estimators $\hat{\pi}^b_k(a \mid X)$, $\hat{r}_k(x)$, and $\hat{f}_k(a, x)$ using $S_k$.

end for

Take a $L$-fold random partition $(I_{\ell})^{L}_{\ell=1}$ of observation indices $[n_{\text{hst}}] = \{1, \ldots, n_{\text{hst}}\}$ such that the size of each fold $I_{\ell}$ is $n_{\text{hst}}^\ell = n_{\text{hst}}/L$.

Take a $L$-fold random partition $(J_l)^{L}_{l=1}$ of observation indices $[n_{\text{evl}}] = \{1, \ldots, n_{\text{evl}}\}$ such that the size of each fold $J_l$ is $n_{\text{evl}}^l = n_{\text{evl}}/L$.

For each $\ell, l \in [L] = \{1, \ldots, L\}$, define $I_{\ell} := \{1, \ldots, n_{\text{hst}}\} \setminus I_{\ell}$ and $J_l := \{1, \ldots, n_{\text{evl}}\} \setminus J_l$.

for $\hat{\sigma}^2 \in \{\sigma_1^2, \ldots, \sigma_n^2\}$ do

for $\hat{\lambda} \in \{\lambda_1, \ldots, \lambda_n\}$ do

Define $\text{Score}_{\hat{\sigma}^2, \hat{\lambda}} = 0$.

for $\ell \in [L]$ do

Obtain $\hat{\pi}$ by solving the following optimization problem:

$$\hat{\pi} = \arg \max_{\pi \in \Pi} \mathbb{E}_{n_{\text{hst}}^{\ell}} \left[ \hat{r}(X) \frac{\pi(A \mid X; \hat{\sigma}^2)}{\hat{\pi}^b(A \mid X)} \{Y - \hat{f}(A, X)\} + \mathbb{E}_{n_{\text{evl}}^{\ell}} \mathbb{E}_{\pi^*(a \mid Z)} \left[ \hat{f}(a, Z) \pi(a \mid Z; \hat{\sigma}^2) \right] + \hat{\lambda} \mathcal{R} \left( \{\beta_a, \beta_{0,a}\} \right) \right],$$

where $\mathbb{E}_{n_{\text{hst}}^{\ell}}$ denotes an empirical approximation using $i \in I_{\ell}$, $\mathbb{E}_{n_{\text{evl}}^{\ell}}$ denotes a sample approximation using $j \in J_{l}$, and $\hat{\pi}^b$, $\hat{r}$, and $\hat{f}$ are the corresponding nuisance estimators chosen from $\hat{\pi}^b_k, \hat{r}_k, \hat{f}_k$.

Update the score $\text{Score}_{\hat{\sigma}^2, \hat{\lambda}}$ by

$$\text{Score}_{\hat{\sigma}^2, \hat{\lambda}} = \text{Score}_{\hat{\sigma}^2, \hat{\lambda}} + \mathbb{E}_{n_{\text{hst}}^{\ell}} \left[ \hat{r}(X) \frac{\hat{\pi}(A \mid X; \hat{\sigma}^2)}{\hat{\pi}^b(A \mid X)} \{Y - \hat{f}(A, X)\} + \mathbb{E}_{n_{\text{evl}}^{\ell}} \mathbb{E}_{\pi^*(a \mid Z)} \left[ \hat{f}(a, Z) \pi(a \mid Z; \hat{\sigma}^2) \right] + \hat{\lambda} \mathcal{R} \left( \{\beta_a, \beta_{0,a}\} \right) \right],$$

where $\mathbb{E}_{n_{\text{hst}}^{\ell}}$ denotes an empirical approximation using $i \in I_{\ell}$, and $\mathbb{E}_{n_{\text{evl}}^{\ell}}$ denotes a sample approximation using $j \in J_{l}$.

end for

end for

end for

end for

Obtain $\hat{\pi}$ by solving the following optimization problem:

$$\hat{\pi} = \arg \max_{\pi \in \Pi} \mathbb{E}_{n_{\text{hst}}} \left[ \hat{r}(X) \frac{\pi(A \mid X; \hat{\sigma}^2)}{\hat{\pi}^b(A \mid X)} \{Y - \hat{f}(A, X)\} + \mathbb{E}_{n_{\text{evl}}} \mathbb{E}_{\pi^*(a \mid Z)} \left[ \hat{f}(a, Z) \pi(a \mid Z; \hat{\sigma}^2) \right] + \hat{\lambda} \mathcal{R} \left( \{\beta_a, \beta_{0,a}\} \right) \right],$$

where $\hat{\sigma}^2 = \arg \max_{\sigma^2} \text{Score}_{\hat{\sigma}^2, \hat{\lambda}}.$

32
Table 6: Off-policy evaluation with the satimage dataset with 800 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.107</td>
<td>67.448</td>
<td>0.042</td>
<td>0.045</td>
<td>0.073</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.096</td>
<td>74.740</td>
<td>0.134</td>
<td>0.093</td>
<td>0.177</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.154</td>
<td>58.031</td>
<td>0.336</td>
<td>0.022</td>
<td>0.372</td>
</tr>
</tbody>
</table>

Table 7: Off-policy evaluation with the vehicle dataset with 800 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.029</td>
<td>218390.000</td>
<td>0.038</td>
<td>0.079</td>
<td>0.040</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.019</td>
<td>329825.704</td>
<td>0.095</td>
<td>0.130</td>
<td>0.089</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.037</td>
<td>173603.802</td>
<td>0.213</td>
<td>0.089</td>
<td>0.210</td>
</tr>
</tbody>
</table>

Table 8: Off-policy evaluation with the satimage dataset with 500 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.112</td>
<td>729.208</td>
<td>0.049</td>
<td>0.077</td>
<td>0.097</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.087</td>
<td>790.188</td>
<td>0.146</td>
<td>0.130</td>
<td>0.173</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.179</td>
<td>453.553</td>
<td>0.335</td>
<td>0.047</td>
<td>0.374</td>
</tr>
</tbody>
</table>

Table 9: Off-policy evaluation with the vehicle dataset with 500 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.029</td>
<td>104311.242</td>
<td>0.049</td>
<td>0.077</td>
<td>0.097</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.014</td>
<td>186170.520</td>
<td>0.146</td>
<td>0.130</td>
<td>0.173</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.034</td>
<td>82883.403</td>
<td>0.335</td>
<td>0.047</td>
<td>0.374</td>
</tr>
</tbody>
</table>

Table 10: Off-policy evaluation with the satimage dataset with 300 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.103</td>
<td>765.985</td>
<td>0.026</td>
<td>0.077</td>
<td>0.097</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.074</td>
<td>40.273</td>
<td>0.126</td>
<td>0.130</td>
<td>0.173</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.169</td>
<td>4367.009</td>
<td>0.297</td>
<td>0.047</td>
<td>0.374</td>
</tr>
</tbody>
</table>

Table 11: Off-policy evaluation with the vehicle dataset with 300 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE</th>
<th>IPWCS MSE</th>
<th>DM MSE</th>
<th>IPWCS-R MSE</th>
<th>DM-R MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7πd + 0.3πu</td>
<td>0.036</td>
<td>78064.888</td>
<td>0.029</td>
<td>0.077</td>
<td>0.097</td>
</tr>
<tr>
<td>0.4πd + 0.6πu</td>
<td>0.020</td>
<td>108655.809</td>
<td>0.126</td>
<td>0.130</td>
<td>0.173</td>
</tr>
<tr>
<td>0.0πd + 1.0πu</td>
<td>0.063</td>
<td>59301.622</td>
<td>0.297</td>
<td>0.047</td>
<td>0.374</td>
</tr>
</tbody>
</table>
### Table 12: Off-policy evaluation with pendigits dataset with 800 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE std</th>
<th>IPWCS MSE std</th>
<th>DM MSE std</th>
<th>IPWCS-R MSE std</th>
<th>DM-R MSE std</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.7\pi^d + 0.3\pi^u$</td>
<td>0.118 0.020</td>
<td>1074.278 838.074</td>
<td>0.083 0.035</td>
<td>0.052 0.045</td>
<td>0.089 0.014</td>
</tr>
<tr>
<td>$0.4\pi^d + 0.6\pi^u$</td>
<td><strong>0.110</strong> 0.026</td>
<td>1328.069 1045.287</td>
<td>0.089 0.053</td>
<td>0.056 0.040</td>
<td>0.231 0.026</td>
</tr>
<tr>
<td>$0.0\pi^d + 1.0\pi^u$</td>
<td><strong>0.314</strong> 0.086</td>
<td>231.043 217.068</td>
<td>0.050 0.049</td>
<td><strong>0.116</strong> 0.187</td>
<td>0.511 0.037</td>
</tr>
</tbody>
</table>

### Table 13: Off-policy learning with pendigits dataset with 800 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS RWD STD</th>
<th>IPWCS RWD STD</th>
<th>DM RWD STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.7\pi^d + 0.3\pi^u$</td>
<td><strong>0.683</strong> 0.030</td>
<td>0.241 0.048</td>
<td>0.507 0.060</td>
</tr>
<tr>
<td>$0.4\pi^d + 0.6\pi^u$</td>
<td><strong>0.678</strong> 0.039</td>
<td>0.252 0.032</td>
<td>0.445 0.096</td>
</tr>
<tr>
<td>$0.0\pi^d + 1.0\pi^u$</td>
<td><strong>0.409</strong> 0.067</td>
<td>0.204 0.031</td>
<td>0.212 0.041</td>
</tr>
</tbody>
</table>

### Table 14: Off-policy evaluation with the satimage dataset with 1,000 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE std</th>
<th>IPWCS MSE std</th>
<th>DM MSE std</th>
<th>IPWCS-R MSE std</th>
<th>DM-R MSE std</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.7\pi^d + 0.3\pi^u$</td>
<td>0.111 0.024</td>
<td>58.724 91.964</td>
<td><strong>0.052</strong> 0.034</td>
<td><strong>0.050</strong> 0.069</td>
<td>0.067 0.019</td>
</tr>
<tr>
<td>$0.4\pi^d + 0.6\pi^u$</td>
<td><strong>0.090</strong> 0.026</td>
<td>118.317 188.729</td>
<td>0.173 0.097</td>
<td><strong>0.099</strong> 0.087</td>
<td>0.170 0.039</td>
</tr>
<tr>
<td>$0.0\pi^d + 1.0\pi^u$</td>
<td><strong>0.145</strong> 0.038</td>
<td>82.801 103.326</td>
<td>0.369 0.106</td>
<td><strong>0.018</strong> 0.026</td>
<td>0.395 0.046</td>
</tr>
</tbody>
</table>

### Table 15: Off-policy evaluation with the pendigits dataset with 1,000 samples

<table>
<thead>
<tr>
<th>Behavior Policy</th>
<th>DRCS MSE std</th>
<th>IPWCS MSE std</th>
<th>DM MSE std</th>
<th>IPWCS-R MSE std</th>
<th>DM-R MSE std</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.7\pi^d + 0.3\pi^u$</td>
<td>0.118 0.021</td>
<td>1299.936 829.752</td>
<td>0.094 0.029</td>
<td><strong>0.040</strong> 0.040</td>
<td><strong>0.090</strong> 0.012</td>
</tr>
<tr>
<td>$0.4\pi^d + 0.6\pi^u$</td>
<td><strong>0.106</strong> 0.021</td>
<td>1483.730 1014.923</td>
<td>0.256 0.063</td>
<td><strong>0.067</strong> 0.078</td>
<td>0.241 0.023</td>
</tr>
<tr>
<td>$0.0\pi^d + 1.0\pi^u$</td>
<td><strong>0.313</strong> 0.091</td>
<td>300.599 216.541</td>
<td>0.496 0.064</td>
<td><strong>0.099</strong> 0.167</td>
<td>0.531 0.033</td>
</tr>
</tbody>
</table>