Supplementary Materials for "Multiply Robust Causal Inference with Double Negative Control Adjustment for Categorical Unmeasured Confounding"

Xu Shi[1], Wang Miao[2], Jennifer C. Nelson[3], and Eric J. Tchetgen Tchetgen[4]

[1]Department of Biostatistics, University of Michigan
[2]Guanghua School of Management, Peking University
[3]Kaiser Permanente Washington Health Research Institute
[4]Department of Statistics, the Wharton School, University of Pennsylvania

A Proof of Lemma 1

Proof. We first show that Assumption 2 indicates that

$$P(\mathbf{W} \mid \mathbf{Z}, a, x) = P(\mathbf{W} \mid \mathbf{U}, x) P(\mathbf{U} \mid \mathbf{Z}, a, x) \quad (1)$$

$$E[Y \mid \mathbf{Z}, a, x] = E[Y \mid \mathbf{U}, a, x] P(\mathbf{U} \mid \mathbf{Z}, a, x), \quad (2)$$

which intuitively states that in the observed data models, the conditional effect of Z on W is proportional to that of Z on Y, as they share a factor $P(\mathbf{U} \mid \mathbf{Z}, a, x)$ which is the confounding mechanism. First, under Assumption 2 we have $Y(a) \perp\!\!\!\perp (A, Z) \mid (U, X)$. One one hand, we have $f(Y(a) \mid U, X) = f(Y(a) \mid A, Z, U, X) = f(Y \mid A = a, Z, U, X)$. On the other hand, $f(Y(a) \mid U, X) = f(Y(a) \mid A, U, X) = f(Y \mid A = a, U, X)$. We thus have $f(Y \mid A = a, U, X) = f(Y \mid Z, A = a, U, X)$. Therefore $Y \perp\!\!\!\perp Z \mid (U, A, X)$ and

$$E[Y \mid \mathbf{Z}, a, x] = E[Y \mid \mathbf{U}, a, x] P(\mathbf{U} \mid \mathbf{Z}, a, x).$$

Second, under Assumption 2 we also have $W \perp\!\!\!\perp (Z, A) \mid U, X$, therefore

$$P(\mathbf{W} \mid \mathbf{Z}, a, x) = P(\mathbf{W} \mid \mathbf{U}, x) P(\mathbf{U} \mid \mathbf{Z}, a, x)$$

Now, by Assumption 4, because $P(\mathbf{W} \mid \mathbf{U}, x)$ has full column rank with $|W| \geq |U|$, it is left invertible. That is, there is a $|U| \times |W|$ matrix denoted as $P(\mathbf{W} \mid \mathbf{U}, x)^{+}$ such that $P(\mathbf{W} \mid \mathbf{U}, x)^{+} P(\mathbf{W} \mid \mathbf{U}, x) = \mathbb{I}_{|U|}$.

Therefore (1) gives

$$P(\mathbf{U} \mid \mathbf{Z}, a, x) = P(\mathbf{W} \mid \mathbf{U}, x)^{+} P(\mathbf{W} \mid \mathbf{Z}, a, x).$$

Combined with (2) we have

$$E[Y \mid \mathbf{Z}, a, x] = E[Y \mid \mathbf{U}, a, x] P(\mathbf{W} \mid \mathbf{U}, x)^{+} P(\mathbf{W} \mid \mathbf{Z}, a, x).$$

Therefore there exist a $1 \times |W|$ vector $h(a, x)$ such that

$$E[Y \mid \mathbf{Z}, a, x] = h(a, x) P(\mathbf{W} \mid \mathbf{Z}, a, x). \tag{3}$$

In particular, $h(a, x)$ does not depend on $\mathbf{U}$ because neither $E[Y \mid \mathbf{Z}, a, x]$ or $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ depend on $\mathbf{U}$. Similarly, by Assumption 4, $P(\mathbf{U} \mid \mathbf{Z}, a, x)$ has a right inverse denoted as $P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+}$, which satisfies $P(\mathbf{U} \mid \mathbf{Z}, a, x) P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+} = \mathbb{I}_{|U|}$. Multiplying both sides of (1) and (2) by $P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+}$, we have

$$E[Y \mid \mathbf{U}, a, x] = E[Y \mid \mathbf{Z}, a, x] P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+} \tag{4}$$

$$P(\mathbf{W} \mid \mathbf{U}, x) = P(\mathbf{W} \mid \mathbf{Z}, a, x) P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+} \tag{5}$$

Now consider

$$\begin{aligned} E[Y(a) \mid x] =& E[Y \mid \mathbf{U}, a, x] P(\mathbf{U} \mid x) \\ \overset{(4)}{=}& E[Y \mid \mathbf{Z}, a, x] P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+} P(\mathbf{U} \mid x) \\ \overset{(3)}{=}& h(a, x) P(\mathbf{W} \mid \mathbf{Z}, a, x) P(\mathbf{U} \mid \mathbf{Z}, a, x)^{+} P(\mathbf{U} \mid x) \\ \overset{(5)}{=}& h(a, x) P(\mathbf{W} \mid \mathbf{U}, x) P(\mathbf{U} \mid x) \\ =& h(a, x) P(\mathbf{W} \mid x) \end{aligned}$$

Therefore

$$E[Y(a)] = \int_{\mathcal{X}} h(a, x) P(\mathbf{W} \mid x) f(x) dx.$$

Thus we complete the proof of Lemma 1. Below we show Corollary 1.

From (1) we know that

$$\text{rank}(P(\mathbf{W} \mid \mathbf{Z}, a, x)) \leq \min\{\text{rank}(P(\mathbf{W} \mid \mathbf{U}, x)), \text{rank}(P(\mathbf{U} \mid \mathbf{Z}, a, x))\} = |U|.$$

By the Sylvester's rank inequality (Gantmakher, 2000) we also know that

$$|U| = \text{rank}(P(\mathbf{W} \mid \mathbf{U}, x)) + \text{rank}(P(\mathbf{U} \mid \mathbf{Z}, a, x) - |U| \leq \text{rank}(P(\mathbf{W} \mid \mathbf{Z}, a, x)).$$

Therefore

$$\text{rank}(P(\mathbf{W} \mid \mathbf{Z}, a, x)) = |U|.$$

Thus one can learn $|U|$ from the rank of $P(\mathbf{W} \mid \mathbf{Z}, a, x)$, which is observable. In particular, when $|Z| = |W| = |U|$, $P(\mathbf{W} \mid \mathbf{Z}, a, x)$ is invertible under Assumption 4, and the above linear system (3) has a unique solution

$$h(a, x) = E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^{-1}.$$

Thus

$$E[Y(a)] = \int_{\mathcal{X}} E[Y \mid \mathbf{Z}, a, x]P(\mathbf{W} \mid \mathbf{Z}, a, x)^{-1}P(\mathbf{W} \mid x)f(x)dx.$$

In contrast, when $|Z| > |U|$ or $|W| > |U|$, $h(a, x)$ is not unique, but $E[Y(a)]$ is still uniquely identified by

$$E[Y(a)] = \int_{\mathcal{X}} h(a, x)P(\mathbf{W} \mid x)f(x)dx.$$

□

B Proof of Lemma 2

Proof. Because $\Delta = \int_{\mathcal{X}} \Big\{ E[Y(1) \mid X = x] - E[Y(0) \mid X = x] \Big\} f(x)dx$, it suffice to consider $E[Y(1) \mid X = x] - E[Y(0) \mid X = x]$. To simplify notation, conditioning on X is implicit in the following proof. In addition, we let the levels of W and Z be w_i and z_j respectively, $i, j = 0, \ldots, k$. We note that for general polytomous negative controls W and Z of $k+1$ categories, we have

$$
\begin{aligned}
&E[Y(1)] = E[Y \mid \mathbf{Z}, A{=}1]P(\mathbf{W} \mid \mathbf{Z}, A{=}1)^{-1}P(\mathbf{W}) \\
=&E[Y \mid \mathbf{Z}, A{=}1]P(\mathbf{W} \mid \mathbf{Z}, A{=}1)^{-1} \cdot \\
&\Big[P(\mathbf{W} \mid \mathbf{Z}, A{=}1)_{(k+1)\times(k+1)}, P(\mathbf{W} \mid \mathbf{Z}, A{=}0)_{(k+1)\times(k+1)}\Big]_{(k+1)\times 2(k+1)} \begin{pmatrix} P(\mathbf{Z}, A{=}1)_{(k+1)\times 1} \\ P(\mathbf{Z}, A{=}0)_{(k+1)\times 1} \end{pmatrix}_{2(k+1)\times 1} \\
=&E[Y \mid \mathbf{Z}, A{=}1]\Big\{P(\mathbf{Z}, A{=}1)_{(k+1)\times 1} + P(\mathbf{W} \mid \mathbf{Z}, A{=}1)^{-1} \cdot P(\mathbf{W} \mid \mathbf{Z}, A{=}0)P(\mathbf{Z}, A{=}0)\Big\},
\end{aligned}
$$

Thus we can simplify $E[Y(1)] - E[Y(0)]$ as follows.

$$
\begin{aligned}
E[Y(1)] - E[Y(0)] =&E[Y \mid \mathbf{Z}, A{=}1][P(\mathbf{Z}) - P(\mathbf{Z}, A{=}0)] - E[Y \mid \mathbf{Z}, A{=}0][P(\mathbf{Z}) - P(\mathbf{Z}, A{=}1)] \\
&+ E[Y \mid \mathbf{Z}, A{=}1]P(\mathbf{W} \mid \mathbf{Z}, A{=}1)^{-1} \cdot P(\mathbf{W} \mid \mathbf{Z}, A{=}0)P(\mathbf{Z}, A{=}0) \\
&- E[Y \mid \mathbf{Z}, A{=}0]P(\mathbf{W} \mid \mathbf{Z}, A{=}0)^{-1} \cdot P(\mathbf{W} \mid \mathbf{Z}, A{=}1)P(\mathbf{Z}, A{=}1) \\
=&E_Z[\delta_A^Y(Z)] - E(Y \mid \mathbf{Z}, A{=}1)\Big[\mathbb{I} - P^{-1}(\mathbf{W} \mid \mathbf{Z}, A{=}1)P(\mathbf{W} \mid \mathbf{Z}, A{=}0)\Big]P(\mathbf{Z}, A{=}0) \\
&-E(Y \mid \mathbf{Z}, A{=}0)\Big[P^{-1}(\mathbf{W} \mid \mathbf{Z}, A{=}0)P(\mathbf{W} \mid \mathbf{Z}, A{=}1) - \mathbb{I}\Big]P(\mathbf{Z}, A{=}1) \\
=&E_Z[\delta_A^Y(Z)] - \sum_{a\in\{0,1\}} E(Y \mid \mathbf{Z}, 1{-}a) \cdot P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1{-}a) \cdot \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) \cdot P(\mathbf{Z}, a)
\end{aligned}
\tag{6}
$$

where $\mathbb{I}$ is an identity matrix, and $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) = P(\mathbf{W} \mid \mathbf{Z}, A{=}1) - P(\mathbf{W} \mid \mathbf{Z}, A{=}0)$ is a (k+1) by (k+1) matrix. We note that $E_Z[\delta_A^Y(Z)]$ is the g-formula of treatment effect but ignoring the unmeasured confounding U, whereas $\sum_{a\in\{0,1\}} E(Y \mid \mathbf{Z}, 1{-}a) \cdot P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1{-}a) \cdot \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) \cdot P(\mathbf{Z}, a)$ is a bias correction term that adjusts for the bias due to unmeasured confounding using a negative control exposure Z and a negative control outcome W. In the following, we show that when $P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)$ is invertible, we have the following two conclusions:

(1) If there is no unmeasured confounder U, then $E[Y \mid \mathbf{Z}, 1{-}a]P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1{-}a)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z})p(\mathbf{Z} \mid a) = 0$.

In this case, the bias correction term is equal to zero, Eq. 6 reduces to the common effect estimate

$$E[Y(1)] - E[Y(0)] = E_Z[\delta_A^Y(Z)].$$

(2) If the unmeasured confounder U exist (which ensures that $E[\boldsymbol{\delta_A^W}(Z) \mid A, X] \neq 0$), then

$$\begin{aligned}
&E[Y \mid \mathbf{Z}, 1-a]_{1\times(k+1)} P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)_{(k+1)\times(k+1)} \boldsymbol{\delta_A^W}(\mathbf{Z})_{(k+1)\times(k+1)} p(\mathbf{Z} \mid a)_{(k+1)\times 1} \\
=&\Big(\xi_{z_1}^Y(1-a), \xi_{z_2}^Y(1-a), \ldots, \xi_{z_k}^Y(1-a)\Big)_{1\times k} \\
&\begin{pmatrix} \xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\ \xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a) \end{pmatrix}_{k\times k}^{-1} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix}_{k\times 1},
\end{aligned}$$

where

$$\begin{aligned}
\xi_{z_j}^Y(1-a) =& E[Y \mid Z = z_j, 1-a] - E[Y \mid Z = z_0, 1-a], t = 1, \ldots, k; \\
\xi_{z_j}^{w_i}(1-a) =& P(W = w_i \mid Z = z_j, 1-a) - P(W = w_i \mid Z = z_0, 1-a), j = 1, \ldots, k, t = 1, \ldots, k; \\
\delta_A^{w_i}(Z) =& P(W = w_i \mid Z, 1) - P(W = w_i \mid Z, 0), j = 1, \ldots, k.
\end{aligned}$$

Note that column sums of $P(\mathbf{W} \mid \mathbf{Z}, 1-a)$ are all equal to 1. One can show that column sums of $P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)$ are also all equal to 1. This is because for an invertible matrix A with column sums all equal to 1, we have $\mathbb{1}^\intercal A A^{-1} = \mathbb{1}^\intercal A^{-1} = \mathbb{1}^\intercal \mathbb{I} = \mathbb{1}^\intercal$, where $\mathbb{1} = (1, 1, \ldots, 1)^\intercal$. Accordingly, we denote the $(k+1) \times (k+1)$ matrix $P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)$ as $P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) = \begin{pmatrix} 1 - \sum_{i=1}^k c_{i0} & \cdots & 1 - \sum_{i=1}^k c_{ik} \\ c_{10} & \cdots & c_{1k} \\ c_{20} & \cdots & c_{2k} \\ \vdots & & \vdots \\ c_{k0} & \cdots & c_{kk} \end{pmatrix}$,

and as we will show later, we have

$$P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) \boldsymbol{\delta_A^W}(\mathbf{Z}) p(\mathbf{Z} \mid a) = \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} \tag{7}$$

where $m_t = \sum_{j=1}^{k}(c_{t,j} - c_{t,0})E[\delta_A^{w_i}(Z) \mid a]$. When there is no unmeasured confounder, because W is a negative control outcome, we know that $E[\delta_A^{w_i}(Z) \mid a] = 0, \forall j$ and thus $m_t = 0, \forall t$. In this case, the bias adjustment term $\sum_{a\in\{0,1\}} E(Y \mid \mathbf{Z}, 1-a) \cdot P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) \cdot \boldsymbol{\delta_A^W}(\mathbf{Z}) \cdot P(\mathbf{Z}, a)$ is equal to zero. When U actually exists and $P(\mathbf{W} \mid \mathbf{Z}, a)$ is invertible, there exists j such that $E[\delta_A^{w_i}(Z) \mid a] \neq 0$. In this case, we solve for $m_t, t = 1, \ldots, k$ in

$$p(\mathbf{W} \mid \mathbf{Z}, 1-a)P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)\boldsymbol{\delta_A^W}(\mathbf{Z})p(\mathbf{Z} \mid a) = p(\mathbf{W} \mid \mathbf{Z}, 1-a) \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix}. \tag{8}$$

As we will show later, the left hand side of (8) can be simplified as

$$p(\mathbf{W} \mid \mathbf{Z}, 1-a)P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)\boldsymbol{\delta_A^W}(\mathbf{Z})p(\mathbf{Z} \mid a) = \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix}_{(k+1)\times k} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix}_{k\times 1}. \tag{9}$$

As we will show later, the right hand side of (8) can be simplified as

$$p(\mathbf{W} \mid \mathbf{Z}, 1-a)_{(k+1)\times(k+1)} \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} = \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I} \end{pmatrix} \boldsymbol{\xi_Z^W}(1-a)_{k\times k} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix}, \tag{10}$$

where $\boldsymbol{\xi_Z^W}(1-a) = \begin{pmatrix} \xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\ \xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a) \end{pmatrix}$ is a $k \times k$ matrix with element $\xi_{z_j}^{w_i}(1-a) = P(W = w_i \mid Z = z_j, 1-a) - P(W = w_i \mid Z = z_0, 1-a), i, j = 1, \ldots, k$.

Because $\begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix}$ has rank k with an identity matrix $\mathbb{I}_{k\times k}$, and $p(\mathbf{W} \mid \mathbf{Z}, 1-a)$ is invertible, we know that the lefthand side of the above Eq. (10) has rank k. Since for a $(k+1) \times k$ matrix A and a $k \times k$ matrix B, we have $\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}$, we know that $\boldsymbol{\xi_Z^W}(1-a)$ has to have rank k. Therefore, $\boldsymbol{\xi_Z^W}(1-a)$ is invertible.

Combining (9) and (10) we arrive at the following linear equations

$$\begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} \xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\ \xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a) \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} = \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix},$$

the solution to which is

$$\begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} = \begin{pmatrix} \xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\ \xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a) \end{pmatrix}^{-1} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix}. \tag{11}$$

Finally, we have

$$\begin{aligned} & E[Y \mid \mathbf{Z}, 1-a] P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) p(\mathbf{Z} \mid a) \\ =& E[Y \mid \mathbf{Z}, 1-a] \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} \text{ by Eq. (7)} \\ =& \Big(E[Y \mid Z = z_0, 1-a], E[Y \mid Z = z_1, 1-a], \ldots, E[Y \mid Z = z_k, 1-a] \Big) \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix} \\ =& \Big(\xi_{z_1}^{Y}(1-a), \xi_{z_2}^{Y}(1-a), \ldots, \xi_{z_k}^{Y}(1-a) \Big) \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix}, \end{aligned}$$

where $\xi^{Y}_{z_j}(1-a) = E[Y \mid Z=z_j, 1-a] - E[Y \mid Z=z_0, 1-a], t = 1, \ldots, k$. By Eq. (11) we have

$$
\begin{aligned}
&E[Y \mid \mathbf{Z}, 1-a]P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(\mathbf{Z})p(\mathbf{Z} \mid a)\\
=&\Big(\xi^{Y}_{z_1}(1-a), \xi^{Y}_{z_2}(1-a), \ldots, \xi^{Y}_{z_k}(1-a)\Big)\\
&\begin{pmatrix}
\xi^{w_1}_{z_1}(1-a) & \xi^{w_1}_{z_2}(1-a) & \cdots & \xi^{w_1}_{z_k}(1-a)\\
\xi^{w_2}_{z_1}(1-a) & \xi^{w_2}_{z_2}(1-a) & \cdots & \xi^{w_2}_{z_k}(1-a)\\
\vdots & \vdots & \ddots & \vdots\\
\xi^{w_k}_{z_1}(1-a) & \xi^{w_k}_{z_2}(1-a) & \cdots & \xi^{w_k}_{z_k}(1-a)
\end{pmatrix}^{-1}
\begin{pmatrix}
E[\delta^{w_1}_{A}(Z) \mid a]\\
E[\delta^{w_2}_{A}(Z) \mid a]\\
\vdots\\
E[\delta^{w_k}_{A}(Z) \mid a]
\end{pmatrix}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&E_Z[\delta^{Y}_{A}(Z)] - \sum_{a\in\{0,1\}} E(Y \mid \mathbf{Z}, 1-a) \cdot P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) \cdot \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(\mathbf{Z}) \cdot P(\mathbf{Z} \mid a)P(a)\\
=&E_Z[\delta^{Y}_{A}(Z)]\\
&-E_{A,Z}[\Big(\xi^{Y}_{z_1}(1-A), \xi^{Y}_{z_2}(1-A), \ldots, \xi^{Y}_{z_k}(1-A)\Big)\\
&\begin{pmatrix}
\xi^{w_1}_{z_1}(1-A) & \xi^{w_1}_{z_2}(1-A) & \cdots & \xi^{w_1}_{z_k}(1-A)\\
\xi^{w_2}_{z_1}(1-A) & \xi^{w_2}_{z_2}(1-A) & \cdots & \xi^{w_2}_{z_k}(1-A)\\
\vdots & \vdots & \ddots & \vdots\\
\xi^{w_k}_{z_1}(1-A) & \xi^{w_k}_{z_2}(1-A) & \cdots & \xi^{w_k}_{z_k}(1-A)
\end{pmatrix}^{-1}
\begin{pmatrix}
\delta^{w_1}_{A}(Z)\\
\delta^{w_2}_{A}(Z)\\
\vdots\\
\delta^{w_k}_{A}(Z)
\end{pmatrix}]\\
&\equiv E_Z[\delta^{Y}_{A}(Z)] - E_{A,Z}[\boldsymbol{R}(1-A)\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z)],
\end{aligned}
$$

where $\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(a)_{k\times k} = \begin{pmatrix}
\xi^{w_1}_{z_1}(a) & \xi^{w_1}_{z_2}(a) & \cdots & \xi^{w_1}_{z_k}(a)\\
\xi^{w_2}_{z_1}(a) & \xi^{w_2}_{z_2}(a) & \cdots & \xi^{w_2}_{z_k}(a)\\
\vdots & \vdots & \ddots & \vdots\\
\xi^{w_k}_{z_1}(a) & \xi^{w_k}_{z_2}(a) & \cdots & \xi^{w_k}_{z_k}(a)
\end{pmatrix}$, $\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z) = \begin{pmatrix}
\delta^{w_1}_{A}(Z)\\
\delta^{w_2}_{A}(Z)\\
\vdots\\
\delta^{w_k}_{A}(Z)
\end{pmatrix}$,
$\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z}}(a)_{k\times 1} = \Big(\xi^{Y}_{z_1}(a), \xi^{Y}_{z_2}(a), \ldots, \xi^{Y}_{z_k}(a)\Big)^{\intercal}$, and $\boldsymbol{R}(a)_{1\times k} = \boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z}}(a)^{\intercal}\Big(\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(a)\Big)^{-1}$.

Proof of Eq. (7):

$$
\begin{aligned}
&P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z})p(\mathbf{Z} \mid a) \\
&= \begin{pmatrix} 1-\sum_{i=1}^{k} c_{i0} & \cdots & 1-\sum_{i=1}^{k} c_{ik} \\ c_{10} & \cdots & c_{1k} \\ c_{20} & \cdots & c_{2k} \\ \vdots & & \vdots \\ c_{k0} & \cdots & c_{kk} \end{pmatrix}_{(k+1)\times(k+1)} \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix}_{(k+1)\times k} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix}_{k\times 1} \\
&= \begin{pmatrix} -\Big(\sum_{i=1}^{k} c_{i1} - c_{i0}\Big) & -\Big(\sum_{i=1}^{k} c_{i2} - c_{i0}\Big) & \cdots & -\Big(\sum_{i=1}^{k} c_{ik} - c_{i0}\Big) \\ c_{11}-c_{10} & c_{12}-c_{10} & \cdots & c_{1k}-c_{10} \\ c_{21}-c_{20} & c_{22}-c_{20} & \cdots & c_{2k}-c_{20} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1}-c_{k0} & c_{k2}-c_{k0} & \cdots & c_{kk}-c_{k0} \end{pmatrix} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix} \\
&= \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} c_{11}-c_{10} & c_{12}-c_{10} & \cdots & c_{1k}-c_{10} \\ c_{21}-c_{20} & c_{22}-c_{20} & \cdots & c_{2k}-c_{20} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k1}-c_{k0} & c_{k2}-c_{k0} & \cdots & c_{kk}-c_{k0} \end{pmatrix} \begin{pmatrix} E[\delta_A^{w_1}(Z) \mid a] \\ E[\delta_A^{w_2}(Z) \mid a] \\ \vdots \\ E[\delta_A^{w_k}(Z) \mid a] \end{pmatrix} \\
&= \begin{pmatrix} -\mathbb{1}_{1\times k} \\ \mathbb{I}_{k\times k} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \end{pmatrix},
\end{aligned}
$$

where $m_t = \sum_{j=1}^{k}(c_{t,j} - c_{t,0})E[\delta_A^{w_i}(Z) \mid a]$.

Proof of of Eq (9):

$$
\begin{aligned}
&p(\mathbf{W} \mid \mathbf{Z}, 1-a) P^{-1}(\mathbf{W} \mid \mathbf{Z}, 1-a) \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) p(\mathbf{Z} \mid a)=\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(\mathbf{Z}) p(\mathbf{Z} \mid a) \\
&=\begin{pmatrix}
-\sum_{i=1}^{k} \delta^{W=w_i}(Z=z_0) & -\sum_{i=1}^{k} \delta^{W=w_i}(Z=z_1) & \cdots & -\sum_{i=1}^{k} \delta^{W=w_i}(Z=z_k) \\
\delta^{w_1}(Z=z_0) & \delta^{w_1}(Z=z_1) & \cdots & \delta^{w_1}(Z=z_k) \\
\delta^{w_2}(Z=z_0) & \delta^{w_2}(Z=z_1) & \cdots & \delta^{w_2}(Z=z_k) \\
\vdots & \vdots & \ddots & \vdots \\
\delta^{w_k}(Z=z_0) & \delta^{w_k}(Z=z_1) & \cdots & \delta^{w_k}(Z=z_k)
\end{pmatrix}
\begin{pmatrix}
P(Z=z_0 \mid a) \\
P(Z=z_1 \mid a) \\
\vdots \\
P(Z=z_k \mid a)
\end{pmatrix} \\
&=\begin{pmatrix}
-\sum_{i=1}^{k} E[\delta_A^{w_i}(Z) \mid a] \\
E[\delta_A^{w_1}(Z) \mid a] \\
E[\delta_A^{w_2}(Z) \mid a] \\
\vdots \\
E[\delta_A^{w_k}(Z) \mid a]
\end{pmatrix}_{(k+1) \times 1}
=\begin{pmatrix}
-\mathbb{1}_{1 \times k} \\
\mathbb{I}_{k \times k}
\end{pmatrix}_{(k+1) \times k}
\begin{pmatrix}
E[\delta_A^{w_1}(Z) \mid a] \\
E[\delta_A^{w_2}(Z) \mid a] \\
\vdots \\
E[\delta_A^{w_k}(Z) \mid a]
\end{pmatrix}_{k \times 1} .
\end{aligned} \tag{12}
$$

Proof of Eq (10): Because $p(\mathbf{W} \mid \mathbf{Z}, 1-a)$ has column sums all equal to one, similar to Eq. (7) we have

$$
\begin{aligned}
&p(\mathbf{W} \mid \mathbf{Z}, 1-a)_{(k+1) \times (k+1)}
\begin{pmatrix}
-\mathbb{1}_{1 \times k} \\
\mathbb{I}_{k \times k}
\end{pmatrix}_{(k+1) \times k} \\
&=\begin{pmatrix}
1-\sum_{i=1}^{k} P(W=w_i \mid Z=z_0, 1-a) & \cdots & 1-\sum_{i=1}^{k} P(W=w_i \mid Z=z_k, 1-a) \\
P(W=w_1 \mid Z=z_0, 1-a) & \cdots & P(W=w_1 \mid Z=z_k, 1-a) \\
P(W=w_2 \mid Z=z_0, 1-a) & \cdots & P(W=w_2 \mid Z=z_k, 1-a) \\
\vdots & & \vdots \\
P(W=w_k \mid Z=z_0, 1-a) & \cdots & P(W=w_k \mid Z=z_k, 1-a)
\end{pmatrix}
\begin{pmatrix}
-\mathbb{1}_{1 \times k} \\
\mathbb{I}_{k \times k}
\end{pmatrix} \\
&=\begin{pmatrix}
\sum_{i=1}^{k} \xi_{z_1}^{w_i}(1-a) & -\sum_{i=1}^{k} \xi_{z_2}^{w_i}(1-a) & \cdots & -\sum_{i=1}^{k} \xi_{z_k}^{w_i}(1-a) \\
\xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\
\xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\
\vdots & \vdots & \ddots & \vdots \\
\xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a)
\end{pmatrix}
=\begin{pmatrix}
-\mathbb{1}_{1 \times k} \\
\mathbb{I}_{k \times k}
\end{pmatrix}
\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1-a)
\end{aligned}
$$

where $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1-a) = \begin{pmatrix} \xi_{z_1}^{w_1}(1-a) & \xi_{z_2}^{w_1}(1-a) & \cdots & \xi_{z_k}^{w_1}(1-a) \\ \xi_{z_1}^{w_2}(1-a) & \xi_{z_2}^{w_2}(1-a) & \cdots & \xi_{z_k}^{w_2}(1-a) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{z_1}^{w_k}(1-a) & \xi_{z_2}^{w_k}(1-a) & \cdots & \xi_{z_k}^{w_k}(1-a) \end{pmatrix}$, and $\xi_{z_j}^{w_i}(1-a) = P(W=w_i \mid Z = z_j, 1-a) - P(W=w_i \mid Z=z_0, 1-a), i, j = 1, \ldots, k$.

B.1 An alternative illustration of the representation

In this section, we illustrate with a toy example that the scaling factor $R(A, X)$ depends on A when there is an A-U interaction in the outcome model $E[Y \mid A, U, X]$. We illustrate this in the following example where Z, W, and U are binary. Note that the following models are not required for the identification and estimation results in our paper.

By Assumption 2 we know that $E[Y \mid A, Z, U, X] = E[Y \mid A, U, X]$ and $E[W \mid A, Z, U, X] = E[W \mid U, X]$. Let $\alpha(X)$, $\beta(X)$, and $\gamma(X)$ denote any arbitrary function of the observed confounders X. Because A, Z, and U are binary, we have the following nonparametric representation of the underlying true data generating models

$$\begin{aligned} &E[Y \mid A, Z, U, X] \overset{NCE}{=} E[Y \mid A, U, X] = \alpha_0(X) + \alpha_A(X)A + \alpha_U(X)U + \alpha_{AU}(X)AU \\ &E[W \mid A, Z, U, X] \overset{NCO}{=} E[W \mid U, X] = \beta_0(X) + \beta_U(X)U \\ &E[U \mid A, Z, X] = \gamma_0(X) + \gamma_A(X)A + \gamma_Z(X)Z + \gamma_{AZ}(X)AZ. \end{aligned} \tag{13}$$

From the true models, we can derive the observed model as follows

$$\begin{aligned} E[Y \mid A, Z, X] &= \alpha_0(X) + \alpha_A(X)A \\ &\quad + \big[\alpha_U(X) + \alpha_{AU}(X)A\big]\big[\gamma_0(X) + \gamma_A(X)A + \gamma_Z(X)Z + \gamma_{AZ}(X)AZ\big] \\ E[W \mid A, Z, X] &= \beta_0(X) + \beta_U(X)\big[\gamma_0(X) + \gamma_A(X)A + \gamma_Z(X)Z + \gamma_{AZ}(X)AZ\big]. \end{aligned} \tag{14}$$

Therefore by definition

$$\begin{aligned} R(A, X) =& \frac{\big[\alpha_U(X) + \alpha_{AU}(X)A\big]\big[\gamma_Z(X) + \gamma_{AZ}(X)A\big]}{\beta_U(X)\big[\gamma_Z(X) + \gamma_{AZ}(X)A\big]} \\ =& \frac{\big[\alpha_U(X) + \alpha_{AU}(X)A\big]}{\beta_U(X)} = \frac{\alpha_U(X)}{\beta_U(X)} + \frac{\alpha_{AU}(X)}{\beta_U(X)}A \end{aligned} \tag{15}$$

We can see that, if there is no A-U interaction in the outcome model $E[Y \mid A, U, X]$, i.e., $\alpha_{AU}(X) = 0$, then $R(A, X) = \alpha_U(X)/\beta_U(X)$, which only depends on X. In this case, $R(X)$ accounts for the different scales of the effects of U on Y and U on W in $\Delta_{\text{bias}} = E[R(X)\delta_A^W(Z,X)]$. In contrast, if there is A-U interaction in $E[Y \mid A,U,X]$, i.e., $\alpha_{AU}(X) \neq 0$, then $R(A,X)$ depends on A, which further accounts for the effect modification by U in the outcome model. In this case, the bias adjustment term should be $\Delta_{\text{bias}} = E[R(1-A,X)\delta_A^W(Z,X)]$ which we illustrate as follows.

To simplify notation, hereafter we ignore covariates X. We have $\delta_A^W(Z) = \beta_U\left[\gamma_A + \gamma_{AZ}Z\right]$ and $\delta_A^W(Z)R(A) = [\gamma_A + \gamma_{AZ}Z][\alpha_U + \alpha_{AU}A]$. Therefore

$$\Delta_{\text{bias}} = E[\delta_A^W(Z)R(1-A)] = E[\gamma_A + \gamma_{AZ}Z][\alpha_U + \alpha_{AU}] - \underline{\alpha_{AU}E[(\gamma_A + \gamma_{AZ}Z)A]}. \tag{16}$$

Now we compare the true ATE and the naive ATE without accounting for unmeasured confounder. When there is A-U interaction in $E[Y \mid A,U,X]$, the true ATE is given by

$$\text{true ATE} = \alpha_A + \underline{\alpha_{AU}E[U]}, \tag{17}$$

whereas what we can obtain from fitting the observed data model $E[Y \mid A,Z]$ is

$$\Delta_{\text{confounded}} = E[\delta_A^Y(Z)] = \alpha_A + E[\gamma_A + \gamma_{AZ}Z][\alpha_U + \alpha_{AU}] + \underline{\alpha_{AU}E[\gamma_0 + \gamma_Z Z]} \tag{18}$$

Note that $E[U] = E[(\gamma_0 + \gamma_Z Z) + (\gamma_A + \gamma_{AZ}Z)A]$. Therefore, from the underlined parts of (16)-(18), we can see that using $R(1-A)$ allows us to account for the effect modification by U, in the scenario where $\alpha_{AU} \neq 0$.

□

C Generalization to polytomous negative controls

In this section, we generalize our results to allow for polytomous negative controls Z and W, and U, i.e., k is any positive integer. Similar to the binary case in Section 3, we first characterize the EIF for Δ in the nonparametric model. We then propose to use the EIF to construct an estimating equation to obtain a multiply robust and locally efficient estimator of Δ which requires estimating the distribution of the observed data under a parametric (or semiparametric) working model and then evaluating the EIF under such working model.

C.1 Efficient influence function in the nonparametric model

Recall that Lemma 2 provides an alternative representation of Δ given by

$$\Delta = \Delta_{\text{confounded}} - \Delta_{\text{bias}},$$
$$\Delta_{\text{confounded}} = E[\delta_A^Y(Z,X)],\ \Delta_{\text{bias}} = E[\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)],$$

where $\boldsymbol{R}(a,x) = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(a,x)^\intercal \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)^{-1}$ is a $1 \times k$ vector with $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(a,x) = \{\xi_{z_1}^Y(a,x), \xi_{z_2}^Y(a,x), \ldots, \xi_{z_k}^Y(a,x)\}^\intercal$ and $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)$ is a $k \times k$ matrix with $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(a,x)_{i,j} = \xi_{z_j}^{w_i}(a,x)$, and $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(z,x) = \{\delta_A^{w_1}(z,x), \delta_A^{w_2}(z,x), \ldots, \delta_A^{w_k}(z,x)\}^\intercal$. In addition, let $\boldsymbol{\Gamma}_W = \{\mathbb{1}(W = w_1), \mathbb{1}(W = w_2), \ldots, \mathbb{1}(W = w_k)\}^\intercal$ denote a $k \times 1$ vector generalizing the binary W, with $\boldsymbol{\Gamma}_{Wi} = \mathbb{1}(W = w_i)$. Let $\boldsymbol{\Pi}(Z \mid A,X) = \{\mathbb{1}(Z = z_1)/f(Z = z_1 \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X), \mathbb{1}(Z = z_2)/f(Z = z_2 \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X), \ldots, \mathbb{1}(Z = z_k)/f(Z = z_k \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X)\}^\intercal$ denote a $k \times 1$ vector generalizing $(2Z-1)/f(Z \mid A,X)$ in the binary case, with $\boldsymbol{\Pi}(Z \mid A,X)_j = \mathbb{1}(Z = z_j)/f(Z = z_j \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X)$. We begin by noting that the EIF for $\Delta_{\text{confounded}}$ in the general case is still given by Eq. (11) of the main manuscript. The following theorem is a natural generalization of Theorem 1 to the case of polytomous Z, W, and U, which reduces to Theorem 1 when $k = 1$. It is proved in Appendix F.

Theorem C.1. *Under Assumptions 1 – 4, the efficient influence function of the bias correction term Δ_{bias} in* *the nonparametric model* $\mathcal{M}_{nonpar}$ *is*

$$
\begin{aligned}
EIF_{\Delta_{bias}} =& E[\boldsymbol{R}(1-A,X) \mid Z,X] \cdot \frac{2A-1}{f(A \mid Z,X)}\Big(\boldsymbol{\Gamma}_W - \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)A - E[\boldsymbol{\Gamma}_W \mid A=0,Z,X]\Big) \\
&+\boldsymbol{\Pi}(Z \mid A,X)^{\intercal}\frac{f(1-A \mid X)}{f(A \mid X)}\Big\{\Big[Y - E[Y \mid Z=0,A,X] - \boldsymbol{R}(A,X)(\boldsymbol{\Gamma}_W - \\
&\quad E[\boldsymbol{\Gamma}_W \mid Z=0,A,X])\Big]\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)^{-1}\Big\} \cdot E[\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X) \mid 1-A,X] \\
&+\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X) - \Delta_{bias}.
\end{aligned}
$$

Thus, the efficient influence function of Δ *is given by*

$$
EIF_{\Delta} = EIF_{\Delta_{confounded}} - EIF_{\Delta_{bias}},
$$

and the semiparametric efficiency bound in $\mathcal{M}_{nonpar}$ *for estimating the ATE is* $E[EIF_{\Delta}(O)^2]^{-1}$.

C.2 Multiply robust estimation of Δ

In this section, we propose a multiply robust and locally efficient estimator using the EIF_{Δ} of Theorem C.1 as an estimating equation and evaluating it under a working model of the observed data distribution. Specifically, let $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{W}}(X)$ be a $k \times k$ matrix with $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{W}}(X)_{i,j} = \eta_{Az_j}^{w_i}(X)$ denoting the joint effect of A and $\mathbb{1}(Z = z_j)$ under the restriction that for $i, j = 1, \ldots, k$

$$
\eta_{Az_j}^{w_i}(X)A\mathbb{1}(Z=z_j) = [\xi_{z_j}^{w_i}(A,X) - \xi_{z_j}^{w_i}(A=0,X)]\mathbb{1}(Z=z_j) = [\delta_A^{w_i}(z=z_j,X) - \delta_A^{w_i}(Z=z_0,X)]A\mathbb{1}(Z=z_j).
$$

It is straightforward to verify that for $i = 1, \ldots, k$

$$
E[\boldsymbol{\Gamma}_{Wi} \mid A,Z,X] = E[\boldsymbol{\Gamma}_{Wi} \mid A=0,Z=z_0,X] + \delta_A^{w_i}(Z=z_0,X)A + \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A=0,X)\boldsymbol{\Gamma}_Z + \boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X)A\,\boldsymbol{\Gamma}_Z, \tag{19}
$$

where $\boldsymbol{\Gamma}_Z = \{\mathbb{1}(Z = z_1), \mathbb{1}(Z = z_2), \ldots, \mathbb{1}(Z = z_k)\}^{\intercal}$, $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A,X)$ is the i-th row of $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)$ with $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A,X)\boldsymbol{\Gamma}_Z = \sum_{j=1}^{k} \xi_{z_j}^{w_i}(A = 0,X)\mathbb{1}(Z = z_j)$, $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X)$ is the i-th row of $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{W}}(X)$ with $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X)A\,\boldsymbol{\Gamma}_Z = \sum_{j=1}^{k} \eta_{Az_j}^{w_i}(X)A\mathbb{1}(Z = z_j)$. Likewise we have

$$
E[Y \mid Z,A,X] = E[Y \mid Z=0,A,X] + \boldsymbol{R}(A,X)\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)\boldsymbol{\Gamma}_Z, \tag{20}
$$

where $\boldsymbol{R}(A,X)\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)\boldsymbol{\Gamma}_Z = \sum_{j=1}^{k} \boldsymbol{R}(A,X)\boldsymbol{\xi}_{\boldsymbol{z_j}}^{\boldsymbol{W}}(A,X)\mathbb{1}(Z = z_j)$ and $\boldsymbol{\xi}_{\boldsymbol{z_j}}^{\boldsymbol{W}}(A,X)$ is the j-th column of $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)$.

Similar to Section 3.3, we specify parametric working model $f(A, Z \mid X; \alpha^{A,Z})$, $E[W \mid A = 0, Z = 0, X; \beta^{W0}]$, $E[Y \mid Z = 0, A, X; \beta^{Y}]$, $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A, X; \beta^{WZ})$, $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X; \beta^{WA})$, $\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{W}}(X; \beta^{WAZ})$, and $\boldsymbol{R}(A, X; \beta^{R})$, with β^{WAZ} a common subset of β^{WZ} and β^{WA}. We estimate the indexing parameters as follows. Let $\hat{\alpha}^{A,Z}_{\text{mle}}$, $\hat{\beta}^{W0}_{\text{mle}}$, and $\hat{\beta}^{Y}_{\text{mle}}$ solve

$$\mathbb{P}_n\Big\{U_{\alpha^{A,Z}}(\hat{\alpha}^{A,Z}_{\text{mle}})\Big\} = \mathbb{P}_n\Big\{\frac{\partial}{\partial \alpha^{A,Z}}\Big|_{\alpha^{A,Z}=\hat{\alpha}^{A,Z}_{\text{mle}}} \log f(A, Z \mid X; \alpha^{A,Z})\Big\} = 0,$$
$$\mathbb{P}_n\Big\{U_{\beta^{W0}}(\hat{\beta}^{W0}_{\text{mle}})\Big\} = \mathbb{P}_n\Big\{\frac{\partial}{\partial \beta^{W0}}\Big|_{\beta^{W0}=\hat{\beta}^{W0}_{\text{mle}}} \mathbb{1}(A = 0, Z = z_0) \log f(W \mid A = 0, Z = z_0, X; \beta^{W0})\Big\} = 0,$$
$$\mathbb{P}_n\Big\{U_{\beta^{Y}}(\hat{\beta}^{Y}_{\text{mle}})\Big\} = \mathbb{P}_n\Big\{\frac{\partial}{\partial \beta^{Y}}\Big|_{\beta^{Y}=\hat{\beta}^{Y}_{\text{mle}}} \mathbb{1}(Z = z_0) \log f(Y \mid A, Z = z_0, X; \beta^{Y})\Big\} = 0 \text{ respectively.}$$

In addition we obtain $\hat{\beta}^{WA}_{\text{dr}}$, $\hat{\beta}^{WZ}_{\text{dr}}$, and $\hat{\beta}^{R}_{\text{dr}}$ by solving the following g-estimating equations generalized to polytomous case evaluated at the above estimated nuisance models

$$\mathbb{P}_n\Big\{U_{\beta^{W_iA},\beta^{W_iZ}}(\hat{\beta}^{W_iA}_{\text{dr}}, \hat{\beta}^{W_iZ}_{\text{dr}})\Big\} = \mathbb{P}_n\Big\{\Big[g_0^{(i)}(A, Z, X) - E[g_0^{(i)}(A, Z, X) \mid X; \hat{\alpha}^{A,Z}_{\text{mle}}]\Big]\Big[\mathbf{\Gamma}_{Wi} - E[\mathbf{\Gamma}_{Wi} \mid A, Z, X; \hat{\beta}^{W0}_{\text{mle}}, \beta^{W_iZ}, \beta^{W_iA}]\Big]\Big\} = 0, \; i = 1, \ldots, k,$$

$$\mathbb{P}_n\Big\{U_{\beta^{R}}(\hat{\beta}^{R}_{\text{dr}})\Big\} = \mathbb{P}_n\Big\{\Big[g_1(A, Z, X) - E[g_1(A, Z, X) \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}}]\Big]\Big[Y - E[Y \mid Z = 0, A, X; \hat{\beta}^{Y}_{\text{mle}}] - \boldsymbol{R}(A, X; \beta^{R})(\mathbf{\Gamma}_W - E[\mathbf{\Gamma}_W \mid Z = 0, A, X; \hat{\beta}^{W0}_{\text{mle}}, \hat{\beta}^{WA}_{\text{dr}}])\Big]\Big\} = 0,$$

where $g_0^{(i)}(A, Z, X)$ is a vector of $\dim(\beta^{W_iA}) + \dim(\beta^{W_iZ}) - \dim(\beta^{W_iAZ})$ functions, $(\beta^{W_iA}, \beta^{W_iZ}, \beta^{W_iAZ})$ is the subset of $(\beta^{WA}, \beta^{WZ}, \beta^{WAZ})$ corresponding to the i-th level of W, $E[\mathbf{\Gamma}_{Wi} \mid A, Z, X]$ is parameterized by Eq. (19), $E[\mathbf{\Gamma}_W \mid A, Z = 0, X]$ is a vector of $E[\mathbf{\Gamma}_{Wi} \mid A, Z = 0, X]$, $i = 1, \ldots, k$, and $g_1(A, Z, X)$ is a $k \times 1$ vector of $\dim(\beta^{R})$ functions. It can be shown that $\hat{\beta}^{WA}_{\text{dr}}$ and $\hat{\beta}^{WZ}_{\text{dr}}$ are CAN under the union model $\mathcal{M}_2 \cup \mathcal{M}_3$, and $\hat{\beta}^{R}_{\text{dr}}$ is CAN under the union model $\mathcal{M}_1 \cup \mathcal{M}_3$ (Robins and Rotnitzky, 2001; Wang and Tchetgen Tchetgen, 2018). The outcome model $E[Y \mid Z, A, X; \hat{\beta}^{Y}_{\text{mle}}, \hat{\beta}^{WZ}_{\text{dr}}, \hat{\beta}^{R}_{\text{dr}}]$ is then obtained using Eq. (20).

Finally the proposed multiply robust estimator solves $\mathbb{P}_n\Big\{EIF_\Delta(O; \hat{\Delta}_{\text{mr}}, \hat{\theta}^\intercal)\Big\} = 0$, where $EIF_\Delta(O; \Delta, \hat{\theta})$ is EIF_Δ evaluated at $\hat{\theta} = \{(\hat{\alpha}^{A,Z}_{\text{mle}})^\intercal, (\hat{\beta}^{Y}_{\text{mle}})^\intercal, (\hat{\beta}^{W0}_{\text{mle}})^\intercal, (\hat{\beta}^{WZ}_{\text{dr}})^\intercal, (\hat{\beta}^{WA}_{\text{dr}})^\intercal, (\hat{\beta}^{R}_{\text{dr}})^\intercal\}^\intercal$. That is

$$\hat{\Delta}_{\text{mr}} = \hat{\Delta}_{\text{confounded,mr}} - \hat{\Delta}_{\text{bias,mr}},$$

where

$$
\begin{aligned}
\hat{\Delta}_{\text{confounded,mr}} = \mathbb{P}_n\Big\{ & \frac{2A-1}{f(A \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}})}\big(Y - E[Y \mid A, Z, X; \hat{\beta}^{Y}_{\text{mle}}, \hat{\beta}^{WZ}_{\text{dr}}, \hat{\beta}^{R}_{\text{dr}}]\big) + \\
& \big(E[Y \mid A{=}1, Z, X; \hat{\beta}^{Y}_{\text{mle}}, \hat{\beta}^{WZ}_{\text{dr}}, \hat{\beta}^{R}_{\text{dr}}] - E[Y \mid A{=}0, Z, X; \hat{\beta}^{Y}_{\text{mle}}, \hat{\beta}^{WZ}_{\text{dr}}, \hat{\beta}^{R}_{\text{dr}}]\big)\Big\} \\
\hat{\Delta}_{\text{bias,mr}} = \mathbb{P}_n\Big\{ & E[\boldsymbol{R}(1-A, X) \mid Z, X; \hat{\beta}^{R}_{\text{dr}}, \hat{\alpha}^{A,Z}_{\text{mle}}]\frac{2A-1}{f(A \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}})}\Big(\boldsymbol{\Gamma}_W - E[\boldsymbol{\Gamma}_W \mid Z, X; \hat{\beta}^{W0}_{\text{mle}}, \hat{\beta}^{WA}_{\text{dr}}, \hat{\beta}^{WZ}_{\text{dr}}]\Big) \\
& + \boldsymbol{\Pi}(Z \mid A, X; \hat{\alpha}^{A,Z}_{\text{mle}})^{\intercal}\frac{f(1-A \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})}{f(A \mid X; \hat{\alpha}^{A,Z}_{\text{mle}})}\Big\{\Big[Y - E[Y \mid Z=0, A, X; \hat{\beta}^{Y}_{\text{mle}}] - \boldsymbol{R}(A, X; \hat{\beta}^{R}_{\text{dr}})(\boldsymbol{\Gamma}_W - \\
& E[\boldsymbol{\Gamma}_W \mid Z=0, A, X; \hat{\beta}^{W0}_{\text{mle}}, \hat{\beta}^{WA}_{\text{dr}}])\Big]\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A, X; \hat{\beta}^{WZ}_{\text{dr}})^{-1}\Big\}E[\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z, X) \mid 1-A, X; \hat{\beta}^{WA}_{\text{dr}}, \hat{\alpha}^{A,Z}_{\text{mle}}] \\
& + \boldsymbol{R}(1-A, X; \hat{\beta}^{R}_{\text{dr}})\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z, X; \hat{\beta}^{WA}_{\text{dr}})\Big\}.
\end{aligned}
$$

Note that $E[\boldsymbol{R}(1-A, X) \mid Z, X; \hat{\beta}^{R}_{\text{dr}}, \hat{\alpha}^{A,Z}_{\text{mle}}] = \sum_a \boldsymbol{R}(1-a, X; \hat{\beta}^{R}_{\text{dr}})f(a \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}})$ is evaluated under $f(A \mid Z, X; \hat{\alpha}^{A,Z}_{\text{mle}})$, and $E[\delta^{W}_{A}(Z, X) \mid 1-A, X; \hat{\beta}^{WA}_{\text{dr}}, \hat{\alpha}^{A,Z}_{\text{mle}}] = \sum_z \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(z, X; \hat{\beta}^{WA}_{\text{dr}})f(z \mid 1-A, X; \hat{\alpha}^{A,Z}_{\text{mle}})$ is evaluated under $f(Z \mid 1-A, X; \hat{\alpha}^{A,Z}_{\text{mle}})$.

The following theorem generalizes Theorem 2 to polytomous case and is proved in Appendix G. The submodels $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are defined as in Section 3.1 except that instead of scalars, $E[\boldsymbol{\Gamma}_W \mid A, Z, X; \beta^{W}]_{k\times 1}$, $\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z, X; \beta^{WA})_{k\times 1}$, $\boldsymbol{R}(A, X; \beta^{R})_{1\times k}$, and $\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A, X; \beta^{WZ})_{k\times k}$ are now vectors and matrices.

Theorem C.2. *Suppose Assumptions 1 – 4 and standard regularity conditions stated in Appendix G hold, then $\sqrt{n}(\hat{\Delta}_{mr} - \Delta)$ is regular and asymptotic linear under $\mathcal{M}_{union}$ with influence function*

$$
IF_{union}(O; \Delta, \theta^*) = EIF_{\Delta}(O; \Delta, \theta^*) - \frac{\partial EIF_{\Delta}(O; \Delta, \theta)}{\partial \theta^{\intercal}}\Big|_{\theta^*} E\Big\{\frac{\partial U_{\theta}(O; \theta)}{\partial \theta^{\intercal}}\Big|_{\theta^*}\Big\}^{-1} U_{\theta}(O; \theta^*),
$$

and thus $\sqrt{n}(\hat{\Delta}_{mr} - \Delta) \to_d N(0, \sigma^2_{\Delta})$ where $\sigma^2_{\Delta}(\Delta, \theta^*) = E[IF_{union}(O; \Delta, \theta^*)^2]$, θ^* denotes the probability limit of $\hat{\theta}$, and $U_{\theta}(O; \theta) = (U^{\intercal}_{\alpha^{A,Z}}, U^{\intercal}_{\beta^{Y}}, U^{\intercal}_{\beta^{W0}}, U^{\intercal}_{\beta^{WA}, \beta^{WZ}}, U^{\intercal}_{\beta^{R}})^{\intercal}$, where $U_{\beta^{WA}, \beta^{WZ}}$ is the collection of $U_{\beta^{W_iA}, \beta^{W_iZ}}, i = 1, \ldots, k$. Furthermore, $\hat{\Delta}_{mr}$ is locally semiparametric efficient in the sense that it achieves the semiparametric efficiency bound for Δ in $\mathcal{M}_{union}$ at the intersection submodel $\mathcal{M}_{intersect} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$ where $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are all correctly specified.

D Proof of Theorem 1 (efficient influence function in $\mathcal{M}_{\text{nonpar}}$ for binary case)

In this section, we show that the efficient influence function in $\mathcal{M}_{\text{nonpar}}$ for

$$\Delta = \int_{\mathcal{X}} \Big\{ E_Z[\delta_A^Y(Z,X) \mid X=x] - E_{A,Z}[R(1-A,X)\delta_A^W(Z,X) \mid X=x] \Big\} f(x)dx$$

where

$$\begin{aligned}
\delta_A^Y(z,x) &= E[Y \mid A=1, Z=z, X=x] - E[Y \mid A=0, Z=z, X=x]; \\
\delta_A^W(z,x) &= E[W \mid A=1, Z=z, X=x] - E[W \mid A=0, Z=z, X=x]; \\
\xi_Z^Y(a,x) &= E[Y \mid A=a, Z=1, X=x] - E[Y \mid Z=0, A=a, X=x]; \\
\xi_Z^W(a,x) &= E[W \mid A=a, Z=1, X=x] - E[W \mid A=a, Z=0, X=x]; \\
R(1-a,x) &= \frac{\xi_Z^Y(1-a,x)}{\xi_Z^W(1-a,x)}
\end{aligned}$$

is

$$\begin{aligned}
&\text{IF}_\Delta(Y,W,A,Z,X) \\
=&\frac{2A-1}{f(A \mid Z,X)}\Big(Y - \delta_A^Y(Z,X)A - E[Y \mid A=0,Z,X]\Big) + \delta_A^Y(Z,X) \\
-&\frac{2A-1}{f(A \mid Z,X)}\Big(W - \delta_A^W(Z,X)A - E[W \mid A=0,Z,X]\Big) \cdot E[R(1-A,X) \mid Z,X] \\
-&\frac{2Z-1}{f(Z \mid A,X)}\Big[Y - E[Y \mid Z=0,A,X] - R(A,X)\Big(W - E[W \mid Z=0,A,X]\Big)\Big] \cdot \\
&\frac{1}{\xi_Z^W(A,X)} E[\delta_A^W(Z,X) \mid 1-A, X] \frac{f(1-A \mid X)}{f(A \mid X)} \\
-&R(1-A,X)\delta_A^W(Z,X) - \Delta.
\end{aligned}$$

Proof. Let $f(Y,W,A,Z,X;\theta)$ denote a one-dimensional regular parametric submodel of $\mathcal{M}_{\text{nonpar}}$ indexed by θ, under which $\Delta_\theta = E_\theta[\delta_{A,\theta}^Y(Z,X)] - E_\theta[R_\theta(1-A,X)\delta_{A,\theta}^W(Z,X)]$. The efficient influence function in $\mathcal{M}_{\text{nonpar}}$ is defined as the unique mean zero, finite variance random variable D satisfying

$$\frac{\partial}{\partial\theta}\Big|_{\theta=0} \Delta_\theta = E[D \cdot S(Y,W,A,Z,X)],$$

where $S(\cdot)$ is the score function of the path $f(Y,W,A,Z,X;\theta)$ at $\theta = 0$, and $\frac{\partial}{\partial\theta}\Big|_{\theta=0}\Delta_\theta$ is the pathwise derivative of Δ. To find D, we derive the following pathwise derivatives. First, for $\delta^Y_{A,\theta}(Z,X) = E_\theta[Y \mid A=1,Z,X] - E_\theta[Y \mid A=0,Z,X]$, we have

$$\begin{aligned}\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta^Y_{A,\theta}(Z,X) =&E\Big[\frac{2A-1}{f(A\mid Z,X)}(Y-E[Y\mid A,Z,X])S(Y,A,Z,X)\Big|Z,X\Big]\\ =&E\Big[\frac{2A-1}{f(A\mid Z,X)}(Y-\delta^Y_A(Z,X)A-E[Y\mid A{=}0,Z,X])S(Y,A,Z,X)\Big|Z,X\Big]\end{aligned} \tag{21}$$

Second, for $\delta^W_{A,\theta}(Z,X) = P_\theta[W \mid A=1,Z,X] - P_\theta[W \mid A=0,Z,X]$, we have

$$\begin{aligned}\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta^W_{A,\theta}(Z,X) =&E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-E[W\mid A,Z,X])S(W,A,Z,X)\Big|Z,X\Big]\\ =&E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-\delta^W_A(Z,X)A-E[W\mid A{=}0,Z,X])S(W,A,Z,X)\Big|Z,X\Big]\end{aligned} \tag{22}$$

Third, for $\xi^Y_{Z;\theta}(1{-}A,X) = E_\theta[Y \mid 1{-}A, Z=1, X] - E_\theta[Y \mid 1{-}A, Z=0, X]$, we have

$$\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^Y_{Z;\theta}(1{-}A,X) =E\Big[\frac{2Z-1}{f(Z\mid 1{-}A,X)}(Y-E[Y\mid Z,1{-}A,X])S(Y,Z,1{-}A,X)\Big|1{-}A,X\Big] \tag{23}$$

Forth, for $\xi^W_{Z;\theta}(1{-}A,X) = P_\theta[W \mid 1{-}A, Z=1, X] - P_\theta[W \mid 1{-}A, Z=0, X]$, we have

$$\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^W_{Z;\theta}(1{-}A,X) =E\Big[\frac{2Z-1}{f(Z\mid 1{-}A,X)}(W-E[W\mid Z,1{-}A,X])S(W,Z,1{-}A,X)\Big|1{-}A,X\Big] \tag{24}$$

Lastly, using Eq. (23) and (24), we have

$$\begin{aligned}\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1{-}A,X) =&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\frac{\xi^Y_{Z;\theta}(1{-}A,X)}{\xi^W_{Z;\theta}(1{-}A,X)}\\ =&\frac{\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^Y_{Z;\theta}(1{-}A,X)\xi^W_Z(1{-}A,X)-\xi^Y_Z(1{-}A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^W_{Z;\theta}(1{-}A,X)}{[\xi^W_Z(1{-}A,X)^2]}\\ =&\frac{1}{\xi^W_Z(1{-}A,X)}\cdot E\Big[\frac{2Z-1}{f(Z\mid 1{-}A,X)}(Y-E[Y\mid Z,1{-}A,X])S(Y,Z,1{-}A,X)\Big|1{-}A,X\Big]\\ &-\frac{R(1{-}A,X)}{\xi^W_Z(1{-}A,X)}\cdot E\Big[\frac{2Z-1}{f(Z\mid 1{-}A,X)}(W-E[W\mid Z,1{-}A,X])S(W,Z,1{-}A,X)\Big|1{-}A,X\Big]\\ =&E\Big[\frac{1}{\xi^W_Z(1{-}A,X)}\cdot\frac{2Z-1}{f(Z\mid 1{-}A,X)}(Y-E[Y\mid Z,1{-}A,X])S(Y,Z,1{-}A,X)\Big|1{-}A,X\Big]\\ &-E\Big[\frac{R(1{-}A,X)}{\xi^W_Z(1{-}A,X)}\cdot\frac{2Z-1}{f(Z\mid 1{-}A,X)}(W-E[W\mid Z,1{-}A,X])S(W,Z,1{-}A,X)\Big|1{-}A,X\Big]\\ =&E\Big[\frac{2Z-1}{f(Z\mid 1{-}A,X)}[Y-E[Y\mid Z,1{-}A,X]-R(1{-}A,X)(W-E[W\mid Z,1{-}A,X])]\cdot\\ &\frac{1}{\xi^W_Z(1{-}A,X)}S(Y,W,Z,1{-}A,X)\Big|1{-}A,X\Big],\end{aligned} \tag{25}$$

where the last equation holds because for any function f, $E[f(Y,Z,A,X)S(W \mid Y,Z,A,X)] = 0$ and $E[f(W,Z,A,X)S(Y \mid W,Z,A,X)] = 0$.

In the following, we consider the pathwise derivative of $E[\delta_A^Y(Z,X)]$ and $E[\delta_A^W(Z,X)\cdot R(1-A,X)]$ respectively. By Eq. (21), the pathwise derivative of $E[\delta_A^Y(Z,X)]$ is given by

$$\begin{aligned}\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\delta_{A,\theta}^Y(Z,X)] =& E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_A^Y(Z,X)] + E[\delta_A^Y(Z,X)S(Z,X)]\\ =& E\{[\frac{2A-1}{f(A\mid Z,X)}(Y-\delta_A^Y(Z,X)A - E[Y\mid A=0,Z,X])+\\ & \delta_A^Y(Z,X)]\cdot S(Y,A,Z,X)\}.\end{aligned}$$

Because $E[f(Y,A,Z,X)S(W \mid Y,A,Z,X)] = 0$ for any function f, we have

$$\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\delta_\theta^Y(Z,X)] = E\Big\{[\frac{2A-1}{f(A\mid Z,X)}(Y-\delta_A^Y(Z,X)A - E[Y\mid A=0,Z,X]) + \delta_A^Y(Z,X)]S(Y,W,A,Z,X)\Big\}.$$

Now we consider the pathwise derivative of $E[\delta_A^W(Z,X)\cdot R(1-A,X)]$. Note that

$$\begin{aligned}&\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\delta_{A,\theta}^W(Z,X)\cdot R_\theta(1-A,X)]\\ =& E[R(1-A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(Z,X)] + E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)] + E[\delta_A^W(Z,X)R(1-A,X)S(A,Z,X)].\end{aligned} \tag{26}$$

Thus we consider the $E[R(1-A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(Z,X)]$ and $E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]$ respectively. First, we consider $E[R(1-A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(Z,X)]$ as follows. By Eq. (22)

$$\begin{aligned}&E[R(1-A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(Z,X)] = \int\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(z,x)R(1-a,x)f(a,z,x)dadzdx\\ =&\int E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-E[W\mid A,Z,X])S(W,A,Z,X)\Big|Z=z,X=x\Big]R(1-a,x)f(a,z,x)dadzdx\\ =&\int E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-E[W\mid A,Z,X])E[R(1-A,X)\mid Z=z,X=x]S(W,A,Z,X)\Big|Z=z,X=x\Big]f(z,x)dzdx\\ =&E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-E[W\mid A,Z,X])e_R(Z,X)S(W,A,Z,X)\Big],\end{aligned} \tag{27}$$

where $e_R(z,x) = E[R(1-A,X) \mid Z=z, X=x]$. Because $E[f(W,A,Z,X)S(Y \mid W,A,Z,X)] = 0$ for any function f, we have

$$E[R(1-A,X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^W(Z,X)] = E\Big[\frac{2A-1}{f(A\mid Z,X)}(W-E[W\mid A,Z,X])e_R(Z,X)S(Y,W,A,Z,X)\Big]. \tag{28}$$

Second, we consider $E[\frac{\partial}{\partial\theta}\Big|_{\theta=0} R_\theta(1-A,X)\delta_A^W(Z,X)]$ as follows. By Eq. (25)

$$
\begin{aligned}
& E[\frac{\partial}{\partial\theta}\Big|_{\theta=0} R_\theta(1-A,X)\delta_A^W(Z,X)] \\
=& \int \frac{\partial}{\partial\theta}\Big|_{\theta=0} R_\theta(1-a,x)\delta_A^W(z,x)f(a,z,x)dadzdx \\
=& \int E\Big[\frac{2Z-1}{f(Z\mid 1-A,X)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big] \\
& \frac{1}{\xi_Z^W(1-A,X)}S(Y,W,Z,1-A,X)\Big|1-A=1-a,X=x\Big]\cdot\delta_A^W(z,x)f(a,z,x)dadzdx
\end{aligned}
$$

Note that the above can not be combined with Eq. (26) and (28) since the score is evaluated at $S(Y,W,Z,1-A,X)$ rather than $S(Y,W,Z,A,X)$, which we solve as follows. Denote $E\Big[\frac{2Z-1}{f(Z|1-A,X)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\frac{1}{\xi_Z^W(1-A,X)}S(Y,W,Z,1-A,X)\Big|1-A=1-a,X=x\Big]$ as $h(1-a,x)$, then we have

$$
\begin{aligned}
& \int h(1-a,x)\delta_A^W(z,x)f(a,z,x)dadzdx \\
=& \int h(1-a,x)\delta_A^W(z,x)[f(z,x)-f(1-a,z,x)]dadzdx \\
=& \int h(1-a,x)\delta_A^W(z,x)f(z\mid x)f(x)dadzdx - \int h(1-a,x)\delta_A^W(z,x)f(z\mid 1-a,x)dzf(1-a,x)dadx \\
=& \underbrace{\int h(1-a,x)E[\delta_A^W(z,x)\mid x]daf(x)dx}_{\mathcal{A}} - \underbrace{\int h(1-a,x)E[\delta_A^W(z,x)\mid 1-a,x]f(1-a,x)dadx}_{\mathcal{B}}.
\end{aligned}
$$

We consider simplifying $\mathcal{A}$ as follows. Because $E[\delta_A^W(Z,X)\mid x]$ is a function of x, we have

$$
\begin{aligned}
& \int h(1-a,x)E[\delta_A^W(Z,X)\mid x]daf(x)dx \\
=& \int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big] \\
& \frac{1}{\xi_Z^W(1-a,x)}S(Y,W,Z,1-a,x)\Big|1-a,x\Big]E[\delta_A^W(Z,X)\mid x]daf(x)dx \\
=& \int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big] \\
& \frac{1}{\xi_Z^W(1-a,x)}E[\delta_A^W(Z,X)\mid x]S(Y,W,Z,1-a,x)\mid 1-a,x\Big]daf(x)dx \\
=& \int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big] \\
& \frac{1}{\xi_Z^W(1-a,x)}\frac{E[\delta_A^W(Z,X)\mid x]}{f(1-a\mid x)}S(Y,W,Z,1-a,x)\mid 1-a,x\Big]f(1-a,x)dadx.
\end{aligned}
\tag{29}
$$

We consider simplifying $\mathcal{B}$ as follows. Because $E[\delta_A^W(Z,X)\mid 1-a,x]$ is a function of $1-a$ and x, we have

$$\begin{aligned}
&\int h(1-a,x)E[\delta_A^W(Z,X)\mid 1-a,x]f(1-a,x)dadx\\
=&\int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\\
&\frac{1}{\xi_Z^W(1-a,x)}S(Y,W,Z,1-a,x)\Big|1-a,x\Big]E[\delta_A^W(Z,X)\mid 1-a,x]f(1-a,x)dadx\\
=&\int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\\
&\frac{1}{\xi_Z^W(1-a,x)}E[\delta_A^W(Z,X)\mid 1-a,x]S(Y,W,Z,1-a,x)\Big|1-a,x\Big]f(1-a,x)dadx.
\end{aligned} \tag{30}$$

Combining Eq. (29) and (30) we have

$$\begin{aligned}
&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]\\
=&\int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\\
&\frac{1}{\xi_Z^W(1-a,x)}S(Y,W,Z,1-a,x)\Big|1-a,x\Big]\delta_A^W(Z,X)f(a,z,x)dadzdx\\
=&\int E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\\
&\frac{1}{\xi_Z^W(1-a,x)}\left(\frac{E[\delta_A^W(Z,X)\mid x]}{f(1-a\mid x)}-E[\delta_A^W(Z,X)\Big|1-a,x]\right)\cdot S(Y,W,Z,1-a,x)\Big|1-a,x\Big]f(1-a,x)dadx.
\end{aligned} \tag{31}$$

Let $g(1-a,x)$ denote

$$\begin{aligned}
&E\Big[\frac{2Z-1}{f(Z\mid 1-a,x)}\Big[Y-E[Y\mid Z,1-A,X]-R(1-A,X)(W-E[W\mid Z,1-A,X])\Big]\\
&\frac{1}{\xi_Z^W(1-a,x)}\left(\frac{E[\delta_A^W(Z,X)\mid x]}{f(1-a\mid x)}-E[\delta_A^W(Z,X)\Big|1-a,x]\right)\cdot S(Y,W,Z,1-a,x)\Big|1-a,x\Big]
\end{aligned}$$

then Eq. (31) is

$$E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]=\int g(1-a,x)f(1-a,x)dadx.$$

Because A is a binary variable taking on values 0 and 1, we have

$$\begin{aligned}
\int g(1-a,x)f(1-a,x)dadx &= \int g(1-0,x)P(A=1-0\mid x)+g(1-1,x)P(A=1-1\mid x)f(x)dx\\
&=\int g(a,x)f(a\mid x)daf(x)dx=\int g(a,x)f(a,x)dadx.
\end{aligned}$$

Therefore Eq. (31) becomes

$$
\begin{aligned}
&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)] = \int g(a,x)f(a,x)dadx\\
=&E\Big[\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid Z,A,X]-R(A,X)(W-E[W\mid Z,A,X])\Big]\\
&\frac{1}{\xi_Z^W(A,X)}\left(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\right)S(Y,W,A,Z,X)].
\end{aligned}
$$

Because $E[f(Y,A,Z,X)S(W\mid Y,A,Z,X)]=0$, and similarly $E[f(W,A,Z,X)S(Y\mid W,A,Z,X)]=0$ for any function f, we have

$$
\begin{aligned}
&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]\\
=&E\Big[\frac{2Z-1}{f(Z\mid A,X)}(Y-E[Y\mid A,Z,X])\frac{1}{\xi_Z^W(A,X)}\Big(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\Big)\\
&[S(Y,A,Z,X)+S(W\mid Y,A,Z,X)]\Big]\\
-&E\Big[\frac{2Z-1}{f(Z\mid A,X)}(W-E[W\mid A,Z,X])\frac{R(A,X)}{\xi_Z^W(A,X)}\Big(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\Big)\\
&[S(W,A,Z,X)+S(Y\mid W,A,Z,X)]\Big]\\
=&E[\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid A,Z,X]-R(A,X)(W-E[W\mid A,Z,X])\Big]\\
&\frac{1}{\xi_Z^W(A,X)}\Big(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\Big)S(Y,W,A,Z,X)].
\end{aligned}
$$

In addition, because Z is binary, we have $E[Y\mid A,Z,X]=\xi_Z^Y(A,X)Z+E[Y\mid Z=0,A,X]=R(A,X)\xi_Z^W(A,X)Z+E[Y\mid Z=0,A,X]$, and $E[W\mid A,Z,X]=\xi_Z^W(A,X)Z-E[W\mid Z=0,A,X]$. Therefore

$$
\begin{aligned}
&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]\\
=&E[\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-R(A,X)\xi_Z^W(A,X)Z-E[Y\mid Z=0,A,X]-R(A,X)(W-\xi_Z^W(A,X)Z-\\
&E[W\mid Z=0,A,X])\Big]\frac{1}{\xi_Z^W(A,X)}\Big(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\Big)S(Y,W,A,Z,X)]\\
=&E[\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid Z=0,A,X]-R(A,X)(W-E[W\mid Z=0,A,X])\Big]\\
&\frac{1}{\xi_Z^W(A,X)}\Big(\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]\Big)S(Y,W,A,Z,X)].
\end{aligned}
$$

Note that

$$\begin{aligned}\frac{E[\delta_A^W(Z,X)\mid X]}{f(A\mid X)}-E[\delta_A^W(Z,X)\mid A,X]=&\int_{\mathcal{Z}}\frac{\delta_A^W(z,x)}{f(a\mid x)}[f(z\mid x)-f(z,a\mid x)]dz\\=&\int_{\mathcal{Z}}\delta_A^W(z,x)\frac{f(z,1-a\mid x)}{f(a\mid x)}dz\\=&\int_{\mathcal{Z}}\delta_A^W(z,x)\frac{f(z\mid 1-a,x)f(1-a\mid x)}{f(a\mid x)}dz\\=&E[\delta_A^W(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}.\end{aligned}$$

Therefore, we finally arrive at

$$\begin{aligned}&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}R_\theta(1-A,X)\delta_A^W(Z,X)]\\=&E[\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid Z=0,A,X]-R(A,X)(W-E[W\mid Z=0,A,X])\Big]\\&\frac{1}{\xi_Z^W(A,X)}\Big(E[\delta_A^W(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big)S(Y,W,A,Z,X)].\end{aligned}\tag{32}$$

Combining Eq. (26), (28), and (32) we have

$$\begin{aligned}&\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\delta_{A,\theta}^W(Z,X)\cdot R_\theta(1-A,X)]\\=&E[\Big\{\delta_A^W(Z,X)R(1-A,X)+\frac{2A-1}{f(A\mid Z,X)}(W-\delta_A^W(Z,X)A-E[W\mid A=0,Z,X])E[R(1-A,X)\mid Z,X]\\&+\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid Z=0,A,X]-R(A,X)(W-E[W\mid Z=0,A,X])\Big]\cdot\\&\frac{1}{\xi_Z^W(A,X)}\Big(E[\delta_A^W(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big)\Big\}S(Y,W,A,Z,X)].\end{aligned}$$

Therefore, the efficient influence function in $\mathcal{M}_{\text{nonpar}}$ for $\Delta=E[\delta_A^Y(Z,X)]-E[\delta_A^W(Z,X)\cdot R(1–A,X)]$ is given by

$$\begin{aligned}\text{IF}_\Delta(Y,W,A,Z,X)=&\frac{2A-1}{f(A\mid Z,X)}\Big(Y-\delta_A^Y(Z,X)A-E[Y\mid A=0,Z,X]\Big)\\&-\frac{2A-1}{f(A\mid Z,X)}\Big(W-\delta_A^W(Z,X)A-E[W\mid A=0,Z,X]\Big)E[R(1-A,X)\mid Z,X]\\&-\frac{2Z-1}{f(Z\mid A,X)}\Big[Y-E[Y\mid Z=0,A,X]-R(A,X)\Big(W-E[W\mid Z=0,A,X]\Big)\Big]\cdot\\&\frac{1}{\xi_Z^W(A,X)}\Big(E[\delta_A^W(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big)\\&+\delta_A^Y(Z,X)-R(1-A,X)\delta_A^W(Z,X)-\Delta.\end{aligned}$$

□

E Proof of Theorem 2

Proof. Under the regularity conditions given in Theorem 3.2 of Newey and McFadden (1994), the estimated nuisance parameters

$$\hat{\theta} = \{(\hat{\alpha}^{A,Z}_{\text{mle}})^{\intercal}, (\hat{\beta}^{Y}_{\text{mle}})^{\intercal}, (\hat{\beta}^{W0}_{\text{mle}})^{\intercal}, (\hat{\beta}^{WZ}_{\text{dr}})^{\intercal}, (\hat{\beta}^{WA}_{\text{dr}})^{\intercal}, (\hat{\beta}^{R}_{\text{dr}})^{\intercal}\}^{\intercal}$$

from solving the moment function vector $\mathbb{P}_n\{U_\theta(O;\theta)\} = 0$ are asymptotically normal and converge at $o(n^{-1/2})$ rate to its probability limit

$$\theta^* = \{(\alpha^{A,Z}_{*})^{\intercal}, (\beta^{Y}_{*})^{\intercal}, (\beta^{W0}_{*})^{\intercal}, (\beta^{WZ}_{*})^{\intercal}, (\beta^{WA}_{*})^{\intercal}, (\beta^{R}_{*})^{\intercal}\}^{\intercal}$$

regardless of whether the corresponding nuisance models are correctly specified.

The main step of the proof is to show that $EIF_\Delta(O)$ is an unbiased estimating equation for Δ under $\mathcal{M}_{\text{union}}$. This is completed by first showing that that $\beta^{WA}_* = \beta^{WA}$, $\beta^{WZ}_* = \beta^{WZ}$ under $\mathcal{M}_2 \cup \mathcal{M}_3$, and $\beta^R_* = \beta^R$ under $\mathcal{M}_1 \cup \mathcal{M}_3$ in Section E.1; then showing that Δ^*_{mr}, the probability limit of $\hat{\Delta}_{\text{mr}}$, satisfies $E[\Delta^*_{\text{mr}}] = \Delta$ in Section E.2.

Now we derive the asymptotic distribution of $\hat{\Delta}_{\text{mr}}$. Assuming that the regularity conditions given in Corollary 1, Chapter 8 of Manski (1988) hold for $EIF_\Delta(O;\Delta,\theta)$ and $U_\theta(O;\theta)$, it follows from standard Taylor expansion of $\sqrt{n}\mathbb{P}_n\{EIF_\Delta(O;\Delta,\theta^*)\} = 0$ that

$$\begin{aligned}0 =& \sqrt{n}\mathbb{P}_n\{EIF_\Delta(O;\Delta,\theta^*)\} + \frac{\partial EIF_\Delta(O;\Delta,\theta)}{\partial \Delta^{\intercal}}\Big|_{\Delta}\sqrt{n}(\hat{\Delta}_{\text{mr}} - \Delta) \\ &+ \sqrt{n}\mathbb{P}_n\left\{\frac{\partial EIF_\Delta(O;\Delta,\theta)}{\partial\theta^{\intercal}}\Big|_{\theta^*} E\Big\{-\frac{\partial U_\theta(O;\theta)}{\partial\theta^{\intercal}}\Big|_{\theta^*}\Big\}^{-1} U_\theta(O;\theta^*)\right\} + o_p(1),\end{aligned}$$

where $\frac{\partial EIF_\Delta(O;\Delta,\theta)}{\partial\Delta^{\intercal}}\Big|_{\Delta} = -1$. Therefore

$$\sqrt{n}(\hat{\Delta}_{\text{mr}} - \Delta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IF_{\text{union}}(O_i;\Delta,\theta^*) + o_p(1),$$

where

$$IF_{\text{union}}(O;\Delta,\theta^*) = EIF_\Delta(O;\Delta,\theta^*) + \frac{\partial EIF_\Delta(O;\Delta,\theta)}{\partial\theta^{\intercal}}\Big|_{\theta^*} E\Big\{-\frac{\partial U_\theta(O;\theta)}{\partial\theta^{\intercal}}\Big|_{\theta^*}\Big\}^{-1} U_\theta(O;\theta^*),$$

O_i stands for the i-th observation and θ^* is the probability limit of $\hat{\theta}$. By Slutsky's Theorem and the Central

Limit Theorem we have $\sqrt{n}(\hat{\Delta}_{\text{mr}} - \Delta) \rightarrow_d N(0, \sigma^2_\Delta)$, where $\sigma^2_\Delta(\Delta, \theta^*) = E[IF_{\text{union}}(O; \Delta, \theta^*)^2]$.

At the intersection submodel $\mathcal{M}_{\text{intersect}}$ where all models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are correctly specified, $\theta^* = \theta$ and we have that

$$\frac{\partial EIF_\Delta(O; \Delta, \theta)}{\partial \theta^\intercal}\Big|_{\theta^*} = 0,$$

and thus

$$IF_{\text{union}}(O; \Delta, \theta^* = \theta) = EIF_\Delta(O; \Delta, \theta^* = \theta).$$

Therefore if all models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ are correctly specified, $\hat{\Delta}_{\text{mr}}$ achieves the semiparametric efficient bound under model $\mathcal{M}_{\text{union}}$.

E.1 Proof that $\beta^{WA}_* = \beta^{WA}$, $\beta^{WZ}_* = \beta^{WZ}$ under $\mathcal{M}_2 \cup \mathcal{M}_3$, and $\beta^R_* = \beta^R$ under $\mathcal{M}_1 \cup \mathcal{M}_3$

To simplify notation, we let $\text{D}^*_{\delta^W_A(Z,X)}$, $\text{D}^*_{R(1-A,X)}$, $R^*(1-A, X)$, $E^*[Y \mid Z = 0, A, X]$, $E^*[W \mid A=0, Z = 0, X]$, $\delta^{W^*}_A(Z, X)$, $\xi^{W^*}_Z(A, X)$, $E^*[W \mid Z = 0, A, X]$, $f^*(A \mid Z, X)$, $f^*(Z \mid A, X)$, and $f^*(A \mid X)$ denote the probability limit of the estimated nuisance models. Similarly, we let Δ^*_{mr}, $\Delta^*_{\text{confounded,dr}}$, and $\Delta^*_{\text{bias,mr}}$ denote the probability limit of the estimated parameters of interest.

We start with showing that $\delta^{W^*}_A(Z, X) = \delta^W_A(Z, X)$ and $\xi^{W^*}_Z(A, X) = \xi^W_Z(A, X)$ under $\mathcal{M}_2 \cup \mathcal{M}_3$, and $R^*(A, X) = R(A, X)$ under $\mathcal{M}_1 \cup \mathcal{M}_3$. Note that $\delta^W_A(Z, X)$, $\xi^W_Z(A, X)$, and $R(A, X)$ do not by themselves give rise to a likelihood, and estimation of these components relies on construction of estimating equations that depends on other components of the full data likelihood such as $f(A, Z \mid X)$ and $E[W \mid A = 0, Z = 0, X]$ which can be estimated by the MLE. Therefore, we show such doubly robust property by showing that the constructed estimating equations are unbiased with mean zero under the union models $\mathcal{M}_2 \cup \mathcal{M}_3$ (or $\mathcal{M}_1 \cup \mathcal{M}_3$).

First, we show that $\delta^{W^*}_A(Z, X) = \delta^W_A(Z, X)$ and $\xi^{W^*}_Z(A, X) = \xi^W_Z(A, X)$ under $\mathcal{M}_2 \cup \mathcal{M}_3$. Under $\mathcal{M}_2$ where $f(A, Z \mid X; \alpha^{A,Z})$, $\xi^W_Z(A, X; \beta^{WZ})$ and $\delta^W_A(Z, X; \beta^{WA})$ are correctly specified, we have $\alpha^{A,Z}_* = \alpha^{A,Z}$, $f^*(A, Z \mid X) = f(A, Z \mid X)$, and thus $E^*[g_0(A, Z, X) \mid X] = E[g_0(A, Z, X) \mid X]$ for any function $g_0(A, Z, X)$. Recall that $\hat{\beta}^{WA}_{\text{dr}}$ and $\hat{\beta}^{WZ}_{\text{dr}}$ solves $\mathbb{P}_n\Big\{U_{\beta^{WA},\beta^{WZ}}(\hat{\beta}^{WA}_{\text{dr}}, \hat{\beta}^{WZ}_{\text{dr}})\Big\} = 0$ with $\lim_{n\rightarrow\infty} \mathbb{P}_n\Big\{U_{\beta^{WA},\beta^{WZ}}(\hat{\beta}^{WA}_{\text{dr}}, \hat{\beta}^{WZ}_{\text{dr}})\Big\} = E[U_{\beta^{WA},\beta^{WZ}}(\beta^{WA}_*, \beta^{WZ}_*)]$. Therefore we consider $E[U_{\beta^{WA},\beta^{WZ}}(\beta^{WA}_*, \beta^{WZ}_*)]\Big|_{\beta^{WA}_*=\beta^{WA},\beta^{WZ}_*=\beta^{WZ}}$ under $\mathcal{M}_2$ where $\xi^W_Z(A, X; \beta^{WZ})$ and $\delta^W_A(Z, X; \beta^{WA})$ are correctly specified, i.e. $\xi^W_Z(A, X) = \xi^W_Z(A, X; \beta^{WZ})$ and $\delta^W_A(Z, X) = \delta^W_A(Z, X; \beta^{WA})$, we have

$$
\begin{aligned}
&E[U_{\beta^{WA},\beta^{WZ}}(\beta^{WA},\beta^{WZ})]\\
=&E\{\Big[g_0(A,Z,X)-E^*[g_0(A,Z,X)\mid X]\Big]\Big[W-E[W\mid A,Z,X;\beta_*^{W0},\beta^{WZ},\beta^{WA}]\Big]\}\\
=&E\{\Big[g_0(A,Z,X)-E[g_0(A,Z,X)\mid X]\Big]\Big[E[W\mid A=0,Z=0,X]-E[W\mid A=0,Z=0,X;\beta_*^{W0}]+\\
&[\xi_Z^W(A=0,X)-\xi_Z^W(A=0,X;\beta^{WZ})]Z+[\delta_A^W(Z=0,X)-\delta_A^W(Z=0,X;\beta^{WA})]A+\\
&[\eta_{AZ}^W(X)-\eta_{AZ}^W(X;\beta^{WAZ})]AZ\Big]\}\\
=&E\{\Big[g_0(A,Z,X)-E[g_0(A,Z,X)\mid X]\Big]\Big[E[W\mid A=0,Z=0,X]-E[W\mid A=0,Z=0,X;\beta_*^{W0}]\Big]\}\\
=&0
\end{aligned}
$$

because $E\Big[\{g_0(A,Z,X)-E[g_0(A,Z,X)\mid X]\}h(X)\Big]=0$ for any function h. Thus, under $\mathcal{M}_2$ where $\xi_Z^W(A,X;\beta^{WZ})$ and $\delta_A^W(Z,X;\beta^{WA})$ are correctly specified, $\mathbb{P}_n\Big\{U_{\beta^{WA},\beta^{WZ}}(\hat\beta_{\text{dr}}^{WA},\hat\beta_{\text{dr}}^{WZ})\Big|_{\hat\beta_{\text{dr}}^{WA}=\beta^{WA},\hat\beta_{\text{dr}}^{WZ}=\beta^{WZ}}\Big\}$ converges to zero, i.e. (β^{WA},β^{WZ}) is a solution to the probability limit of $\mathbb{P}_n\Big\{U_{\beta^{WA},\beta^{WZ}}(\hat\beta_{\text{dr}}^{WA},\hat\beta_{\text{dr}}^{WZ})\Big\}=0$. Thus $\beta_*^{WA}=\beta^{WA}$, and $\beta_*^{WZ}=\beta^{WZ}$, and thus $\delta_A^{W^*}(Z,X)=\delta_A^W(Z,X)$ and $\xi_Z^{W^*}(A,X)=\xi_Z^W(A,X)$ under $\mathcal{M}_2$.

Similar arguments apply to the scenario under $\mathcal{M}_3$. Under $\mathcal{M}_3$ where working models $R(A,X;\beta^R)$, $E[Y\mid Z=0,A,X;\beta^Y]$, $\xi_Z^W(A,X;\beta^{WZ})$, $\delta_A^W(Z,X;\beta^{WA})$, and $E[W\mid A=0,Z=0,X;\beta^W]$ are correctly specified, we have $\beta_*^{W0}=\beta^{W0}$ and thus $E^*[W\mid A=0,Z=0,X]=E[W\mid A=0,Z=0,X]$. In addition, we again have $\xi_Z^W(A,X)=\xi_Z^W(A,X;\beta^{WZ})$ and $\delta_A^W(Z,X)=\delta_A^W(Z,X;\beta^{WA})$. Now consider

$$
\begin{aligned}
&E[U_{\beta^{WA},\beta^{WZ}}(\beta^{WA},\beta^{WZ})]\\
=&E\{\Big[g_0(A,Z,X)-E^*[g_0(A,Z,X)\mid X]\Big]\Big[W-E[W\mid A,Z,X;\beta_*^{W0},\beta^{WZ},\beta^{WA}]\Big]\}\\
=&E\{\Big[g_0(A,Z,X)-E^*[g_0(A,Z,X)\mid X]\Big]\Big[E[W\mid A=0,Z=0,X]-E[W\mid A=0,Z=0,X;\beta_*^{W0}]+\\
&[\xi_Z^W(A=0,X)-\xi_Z^W(A=0,X;\beta^{WZ})]Z+[\delta_A^W(Z=0,X)-\delta_A^W(Z=0,X;\beta^{WA})]A+\\
&[\eta_{AZ}^W(X)-\eta_{AZ}^W(X;\beta^{WAZ})]AZ\Big]\}\\
=&E\{\Big[g_0(A,Z,X)-E^*[g_0(A,Z,X)\mid X]\Big]\Big[E[W\mid A=0,Z=0,X]-E[W\mid A=0,Z=0,X;\beta_*^{W0}]\Big]\}\\
=&0
\end{aligned}
$$

because $E[W\mid A=0,Z=0,X;\beta_*^{W0}]=E[W\mid A=0,Z=0,X]$. Therefore $\delta_A^{W^*}(Z,X)=\delta_A^W(Z,X)$

and $\xi_Z^{W*}(A,X) = \xi_Z^W(A,X)$ under $\mathcal{M}_3$. In addition, we have that

$$\begin{aligned} E^*[W \mid A,Z,X] =& E^*[W \mid A=0,Z=0,X] + \xi_Z^{W*}(A=0,X)Z + \delta_A^{W*}(Z=0,X)A + \eta_{AZ}^{W*}(X)AZ \\ =& E[W \mid A=0,Z=0,X] + \xi_Z^{W}(A=0,X)Z + \delta_A^{W}(Z=0,X)A + \eta_{AZ}^{W}(X)AZ \\ =& E[W \mid A,Z,X]. \end{aligned} \tag{33}$$

Second, we show that $R^*(A,X) = R(A,X)$ under $\mathcal{M}_1 \cup \mathcal{M}_3$. Under $\mathcal{M}_1$ where working models $f(A,Z \mid X;\alpha^{A,Z})$ and $R(A,X;\beta^R)$ are correctly specified, we have $\alpha_*^{A,Z} = \alpha^{A,Z}$, $f^*(A,Z \mid X) = f(A,Z \mid X)$, and thus $E^*[g_1(A,Z,X) \mid A,X] = E[g_1(A,Z,X) \mid A,X]$ for any function $g_1(A,Z,X)$. Recall that $\hat{\beta}_{\text{dr}}^R$ solves $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = 0$ with $\lim_{n\to\infty}\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = E[U_{\beta^R}(\beta_*^R)]$. Now consider $E[U_{\beta^R}(\beta_*^R)]\Big|_{\beta_*^R=\beta^R}$ under $\mathcal{M}_1$ where $R(A,X;\beta^R)$ is correctly specified, i.e. $R(A,X) = R(A,X;\beta^R)$, we have

$$\begin{aligned} &E[U_{\beta^R}(\beta^R)] = E\Big\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A,X]\Big]\Big[Y - E^*[Y \mid Z=0,A,X] - \\ &R(A,X;\beta^R)(W - E^*[W \mid Z=0,A,X])\Big]\Big\} \\ =& E\Big\{\Big[g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\Big]\Big[\{R(A,X) - R(A,X;\beta^R)\}\xi_Z^W(A,X)Z + \\ &\{E[Y \mid Z=0,A,X] - E^*[Y \mid Z=0,A,X]\} + \{E[W \mid Z=0,A,X] - E^*[W \mid Z=0,A,X]\}R(A,X;\beta^R)\Big]\Big\} \\ =& E\Big\{\Big[g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\Big]\Big[\{E[Y \mid Z=0,A,X] - E^*[Y \mid Z=0,A,X]\} + \\ &\{E[W \mid Z=0,A,X] - E^*[W \mid Z=0,A,X]\}R(A,X;\beta^R)\Big]\Big\} \\ =& 0 \end{aligned}$$

because $E\Big[\{g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\}h(A,X)\Big] = 0$ for any function h. Thus, under $\mathcal{M}_1$ where $R(A,X;\beta^R)$ is correctly specified, $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big|_{\hat{\beta}_{\text{dr}}^R=\beta^R}\Big\}$ converges to zero, i.e. β^R is a solution to the probability limit of $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = 0$. Thus $\beta_*^R = \beta^R$ and $R^*(A,X) = R(A,X)$ under $\mathcal{M}_1$.

Similar arguments apply to the scenario under $\mathcal{M}_3$. Under $\mathcal{M}_3$ where working models $R(A,X;\beta^R)$, $E[Y \mid Z=0,A,X;\beta^Y]$ and $E[W \mid A,Z,X;\beta^W]$ are correctly specified, we have $E^*[Y \mid Z=0,A,X] =$

$E[Y \mid Z=0, A, X]$ and by (33) we have $E^*[W \mid A, Z, X] = E[W \mid A, Z, X]$. We again consider

$$
\begin{aligned}
&E[U_{\beta^R}(\beta^R)] = E\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A, X]\Big]\Big[Y - E^*[Y \mid Z=0, A, X] - \\
&R(A,X;\beta^R)(W - E^*[W \mid Z=0, A, X])\Big]\} \\
=&E\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A, X]\Big]\Big[\{R(A,X) - R(A,X;\beta^R)\}\xi_Z^W(A,X)Z + \\
&\{E[Y \mid Z=0, A, X] - E^*[Y \mid Z=0, A, X]\} + \{E[W \mid Z=0, A, X] - E^*[W \mid Z=0, A, X]\}R(A,X;\beta^R)\Big]\} \\
=&E\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A, X]\Big]\Big[\{E[Y \mid Z=0, A, X] - E^*[Y \mid Z=0, A, X]\} + \\
&\{E[W \mid Z=0, A, X] - E^*[W \mid Z=0, A, X]\}R(A,X;\beta^R)\Big]\} \\
=&0
\end{aligned}
$$

because $E^*[Y \mid Z=0, A, X] = E[Y \mid Z=0, A, X]$ and $E^*[W \mid Z=0, A, X] = E[W \mid Z=0, A, X]$. Therefore $\beta_*^R = \beta^R$ and $R^*(A,X) = R(A,X)$ under $\mathcal{M}_3$.

E.2 Proof that $E[\Delta_{\mathbf{mr}}^*] = \Delta$ under $\mathcal{M}_{\mathbf{union}}$

Using the results in Section E.1, we now show that $E[\Delta_{\text{mr}}^*] = \Delta$ under $\mathcal{M}_{\text{union}}$. To this end, we consider $E[\Delta_{\text{confounded}}^*]$, $E[\mathrm{D}^*_{\boldsymbol{\delta_A^W}(Z,X)}]$, and $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]$ under $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ respectively.

Under $\mathcal{M}_1$ where working models $f(A, Z \mid X; \alpha^{A,Z})$ and $R(A, X; \beta^R)$ are correctly specified, we have $f^*(A, Z \mid X) = f(A, Z \mid X)$ and $R^*(A,X) = R(A,X)$. First we consider

$$
E[\Delta_{\text{confounded}}^*] = E\Big[\frac{2A-1}{f^*(A \mid Z, X)}\Big(E[Y \mid A, Z, X] - E^*[Y \mid A, Z, X]\Big) + E^*[Y \mid A{=}1, Z, X] - E^*[Y \mid A{=}0, Z, X]\Big].
$$

Note that for any function of A, Z and X, denoted as $h(A,Z,X)$, we have

$$
E[\frac{2A-1}{f(A \mid Z, X)}h(A,Z,X)] = E[h(1,Z,X) - h(0,Z,X)]. \tag{34}
$$

Accordingly, when $f^*(A \mid Z, X) = f(A \mid Z, X)$, we have

$$
\begin{aligned}
E[\Delta_{\text{confounded}}^*] =& E\Big[\frac{2A-1}{f(A \mid Z, X)}\Big(E[Y \mid A, Z, X] - E^*[Y \mid A, Z, X]\Big) + E^*[Y \mid A{=}1, Z, X] - E^*[Y \mid A{=}0, Z, X]\Big] \\
=& E[\delta_A^Y(Z,X) - \delta_A^{Y^*}(Z,X) + \delta_A^{Y^*}(Z,X)] = \Delta_{\text{confounded}}.
\end{aligned}
$$

Second, consider

$$
\begin{aligned}
E[\mathrm{D}^*_{\delta_A^W(Z,X)}] =& E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(W - E^*[W \mid A,Z,X]\Big)\sum_A R^*(1-A,X)f^*(A \mid Z,X)\Big] \\
=& E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(E[W \mid A,Z,X] - E^*[W \mid A,Z,X]\Big)\sum_A R^*(1-A,X)f^*(A \mid Z,X)\Big].
\end{aligned}
$$

When $f^*(A \mid Z,X) = f(A \mid Z,X)$, by Eq. (34) we have

$$
\begin{aligned}
E[\mathrm{D}^*_{\delta_A^W(Z,X)}] =& E\Big[\frac{2A-1}{f(A \mid Z,X)}\Big(E[W \mid A,Z,X] - E^*[W \mid A,Z,X]\Big)E[R^*(1-A,X) \mid Z,X]\Big] \\
=& E\Big[\Big(\delta_A^W(Z,X) - \delta_A^{W^*}(Z,X)\Big)E[R^*(1-A,X) \mid Z,X]\Big].
\end{aligned}
$$

Because we also have $R^*(A,X) = R(A,X)$,

$$
\begin{aligned}
& E[\mathrm{D}^*_{\delta_A^W(Z,X)} + R^*(1-A,X)\delta_A^{W^*}(Z,X)] \\
=& E\Big[\Big(\delta_A^W(Z,X) - \delta_A^{W^*}(Z,X)\Big)E[R(1-A,X) \mid Z,X] + R(1-A,X)\delta_A^{W^*}(Z,X)\Big] = \Delta_{\text{bias}}.
\end{aligned}
$$

Third, consider

$$
\begin{aligned}
E[\mathrm{D}^*_{R(1-A,X)}] =& E\Big\{\frac{2Z-1}{f^*(Z \mid A,X)}\frac{1}{\xi_Z^{W^*}(A,X)}\Big(\sum_Z \delta_A^{W^*}(Z,X)f^*(Z \mid 1-A,X)\frac{f^*(1-A \mid X)}{f^*(A \mid X)}\Big) \\
& \Big[Y - E^*[Y \mid Z=0,A,X] - R^*(A,X)\Big(W - E^*[W \mid Z=0,A,X]\Big)\Big]\Big\} \\
=& E\Big\{\frac{2Z-1}{f^*(Z \mid A,X)}\frac{1}{\xi_Z^{W^*}(A,X)}\Big(\sum_Z \delta_A^{W^*}(Z,X)f^*(Z \mid 1-A,X)\frac{f^*(1-A \mid X)}{f^*(A \mid X)}\Big) \\
& \Big[\{R(A,X) - R^*(A,X)\}\xi_Z^W(A,X)Z + \{E[Y \mid Z=0,A,X] - E^*[Y \mid Z=0,A,X]\} \\
& + \{E[W \mid Z=0,A,X] - E^*[W \mid Z=0,A,X]\}R^*(A,X)\Big]\Big\}.
\end{aligned}
$$

When $f^*(Z \mid A,X) = f(Z \mid A,X)$, by similar argument as Eq. (34) we have

$$
E[\mathrm{D}^*_{R(1-A,X)}] = E\Big\{\frac{\xi_Z^W(A,X)}{\xi_Z^{W^*}(A,X)}[R(A,X)-R^*(A,X)]\cdot[\sum_Z \delta_A^{W^*}(Z,X)f^*(Z \mid 1-A,X)]\frac{f^*(1-A \mid X)}{f^*(A \mid X)}\Big\}. \tag{35}
$$

We can see that when $R^*(A,X) = R(A,X)$, $E[\mathrm{D}^*_{R(1-A,X)}] = 0$.

In summary, under $\mathcal{M}_1$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}] =& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\delta^W_A(Z,X)} + R^*(1-A,X)\delta^{W^*}_A(Z,X)] + E[\mathrm{D}^*_{R(1-A,X)}]\} \\
=& \Delta_{\text{confounded}} - \{\Delta_{\text{bias}} + 0\} = \Delta
\end{aligned}
$$

Under $\mathcal{M}_2$ where $f(A,Z \mid X;\alpha^{A,Z})$, $\xi^W_Z(A,X;\beta^{WZ})$ and $\delta^W_A(Z,X;\beta^{WA})$ are correctly specified, we have $f^*(A,Z \mid X) = f(A,Z \mid X)$, $\xi^{W^*}_Z(A,X) = \xi^W_Z(A,X)$, and $\xi^{W^*}_A(A,X) = \xi^W_A(A,X)$. Particularly, $f^*(A \mid X) = f(A \mid X)$. First we consider

$$
E[\Delta^*_{\text{confounded}}] = E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E^*[Y \mid A,Z,X]\Big) + E^*[Y \mid A=1,Z,X] - E^*[Y \mid A=0,Z,X]\Big].
$$

When $f^*(A \mid Z,X) = f(A \mid Z,X)$, by Eq. (34) we have

$$
\begin{aligned}
E[\Delta^*_{\text{confounded}}] =& E\Big[\frac{2A-1}{f(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E^*[Y \mid A,Z,X]\Big) + E^*[Y \mid A=1,Z,X] - E^*[Y \mid A=0,Z,X]\Big] \\
=& E[\delta^Y_A(Z,X) - \delta^{Y^*}_A(Z,X) + \delta^{Y^*}_A(Z,X)] = \Delta_{\text{confounded}}.
\end{aligned}
$$

Second, consider

$$
\begin{aligned}
E[\mathrm{D}^*_{\delta^W_A(Z,X)}] =& E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(W - E^*[W \mid A,Z,X]\Big)\sum_A R^*(1-A,X)f^*(A \mid Z,X)\Big] \\
=& E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(E[W \mid A,Z,X] - E^*[W \mid A,Z,X]\Big)\sum_A R^*(1-A,X)f^*(A \mid Z,X)\Big].
\end{aligned}
$$

When $f^*(A \mid Z,X) = f(A \mid Z,X)$, by Eq. (34) we have

$$
\begin{aligned}
E[\mathrm{D}^*_{\delta^W_A(Z,X)}] =& E\Big[\frac{2A-1}{f(A \mid Z,X)}\Big(E[W \mid A,Z,X] - E^*[W \mid A,Z,X]\Big)E[R^*(1-A,X) \mid Z,X]\Big] \\
=& E\Big[\Big(\delta^W_A(Z,X) - \delta^{W^*}_A(Z,X)\Big)E[R^*(1-A,X) \mid Z,X]\Big] \\
=& 0
\end{aligned}
$$

because $\xi^{W^*}_A(A,X) = \xi^W_A(A,X)$.

Third, consider

$$
\begin{aligned}
E[\mathrm{D}^*_{R(1-A,X)}] =& E\Big\{\frac{2Z-1}{f^*(Z\mid A,X)}\frac{1}{\xi_Z^{W^*}(A,X)}\Big(\sum_Z \delta_A^{W^*}(Z,X)f^*(Z\mid 1-A,X)\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big)\\
&\Big[Y-E^*[Y\mid Z=0,A,X]-R^*(A,X)\Big(W-E^*[W\mid Z=0,A,X]\Big)\Big]\Big\}\\
=&E\Big\{\frac{2Z-1}{f^*(Z\mid A,X)}\frac{1}{\xi_Z^{W^*}(A,X)}\Big(\sum_Z \delta_A^{W^*}(Z,X)f^*(Z\mid 1-A,X)\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big)\\
&\Big[\{R(A,X)-R^*(A,X)\}\xi_Z^W(A,X)Z+\{E[Y\mid Z=0,A,X]-E^*[Y\mid Z=0,A,X]\}\\
&+\{E[W\mid Z=0,A,X]-E^*[W\mid Z=0,A,X]\}R^*(A,X)\Big]\Big\}.
\end{aligned}
$$

When $f^*(Z\mid A,X)=f(Z\mid A,X)$, by similar argument as Eq. (34) we have

$$
E[\mathrm{D}^*_{R(1-A,X)}]=E\Big\{\frac{\xi_Z^W(A,X)}{\xi_Z^{W^*}(A,X)}[R(A,X)-R^*(A,X)]\cdot[\sum_Z \delta_A^{W^*}(Z,X)f^*(Z\mid 1-A,X)]\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big\}. \tag{36}
$$

Note that when the model for $f(A\mid X)$ is correctly specified, i.e., $f^*(A\mid X)=f(A\mid X)$, in Appendix H we show that for any function $h(Y,W,A,Z,X)$, we have

$$
E[h(Y,W,A,Z,X)\frac{f(1-A\mid X)}{f(A\mid X)}]=E[h(Y,W,Z,1-A,X)].
$$

Let $h(Y,W,A,Z,X)=\frac{\xi_Z^W(A,X)}{\xi_Z^{W^*}(A,X)}[R(A,X)-R^*(A,X)]\cdot[\sum_Z \delta_A^{W^*}(Z,X)f^*(Z\mid 1-A,X)]$, then Eq. (36) is equivalent to

$$
E[\mathrm{D}^*_{R(1-A,X)}]=E\Big\{\frac{\xi_Z^W(1-A,X)}{\xi_Z^{W^*}(1-A,X)}[R(1-A,X)-R^*(1-A,X)]\cdot[\sum_Z \delta_A^{W^*}(Z,X)f^*(Z\mid 1-(1-A),X)]\Big\}.
$$

In this case, because we also have that $\xi_Z^{W^*}(A,X)=\xi_Z^W(A,X)$, and $\xi_A^{W^*}(A,X)=\xi_A^W(A,X)$,

$$
\begin{aligned}
&E[\mathrm{D}^*_{R(1-A,X)}+R^*(1-A,X)\delta_A^{W^*}(Z,X)]\\
=&E\Big\{[R(1-A,X)-R^*(1-A,X)]E[\delta_A^W(Z,X)\mid A,X]+R^*(1-A,X)\delta_A^W(Z,X)\Big\}=\Delta_{\text{bias}}.
\end{aligned}
$$

In summary, under $\mathcal{M}_2$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}] =& E[\Delta^*_{\text{confounded}}]-\{E[\mathrm{D}^*_{\delta_A^W(Z,X)}]+E[\mathrm{D}^*_{R(1-A,X)}+R^*(1-A,X)\delta_A^{W^*}(Z,X)]\}\\
=&\Delta_{\text{confounded}}-\{0+\Delta_{\text{bias}}\}=\Delta
\end{aligned}
$$

Under $\mathcal{M}_3$ where $R(A, X; \beta^R)$, $E[Y \mid Z = 0, A, X; \beta^Y]$, $\xi_Z^W(A, X; \beta^{WZ})$, $\delta_A^W(Z, X; \beta^{WA})$, and $E[W \mid A = 0, Z = 0, X; \beta^W]$ are correctly specified, we have $R^*(A, X) = R(A, X)$, $E^*[Y \mid Z = 0, A, X] = E[Y \mid Z = 0, A, X]$, $\xi_Z^{W^*}(A, X) = \xi_Z^W(A, X)$, $\delta_A^{W^*}(Z, X) = \delta_A^W(Z, X)$, and $E^*[W \mid A = 0, Z = 0, X] = E[W \mid A = 0, Z = 0, X]$. First we consider

$$E[\Delta^*_{\text{confounded}}] = E\Big[\frac{2A-1}{f^*(A \mid Z, X)}\Big(E[Y \mid A, Z, X] - E^*[Y \mid A, Z, X]\Big) + E^*[Y \mid A{=}1, Z, X] - E^*[Y \mid A{=}0, Z, X]\Big].$$

Note that

$$\begin{aligned} E^*[Y \mid Z, A, X] =& E^*[Y \mid Z = 0, A, X] + R^*(A, X)\xi_Z^{W^*}(A, X)Z \\ =& E[Y \mid Z = 0, A, X] + R(A, X)\xi_Z^W(A, X)Z = E[Y \mid Z, A, X], \end{aligned}$$

therefore we have

$$\begin{aligned} E[\Delta^*_{\text{confounded}}] =& E\Big[\frac{2A-1}{f(A \mid Z, X)}\Big(E[Y \mid A, Z, X] - E[Y \mid A, Z, X]\Big) + E[Y \mid A{=}1, Z, X] - E[Y \mid A{=}0, Z, X]\Big] \\ =& E\{E[Y \mid A{=}1, Z, X] - E[Y \mid A{=}0, Z, X]\} = \Delta_{\text{confounded}}. \end{aligned}$$

Second, consider

$$\begin{aligned} E[\mathrm{D}^*_{\delta_A^W(Z,X)}] =& E\Big[\frac{2A-1}{f^*(A \mid Z, X)}\Big(W - E^*[W \mid A, Z, X]\Big)\sum_A R^*(1{-}A, X)f^*(A \mid Z, X)\Big] \\ =& E\Big[\frac{2A-1}{f^*(A \mid Z, X)}\Big(E[W \mid A, Z, X] - E^*[W \mid A, Z, X]\Big)\sum_A R^*(1{-}A, X)f^*(A \mid Z, X)\Big] \\ =& 0 \end{aligned}$$

because $E^*[W \mid A, Z, X] = E[W \mid A, Z, X]$ by (33).

Third, consider

$$
\begin{aligned}
E[\mathrm{D}^*_{R(1-A,X)}] =& E\Big\{ \frac{2Z-1}{f^*(Z \mid A,X)} \frac{1}{\xi_Z^{W^*}(A,X)} \Big(\sum_Z \delta_A^{W^*}(Z,X) f^*(Z \mid 1-A, X) \frac{f^*(1-A \mid X)}{f^*(A \mid X)} \Big) \\
& \Big[Y - E^*[Y \mid Z=0, A, X] - R^*(A,X) \Big(W - E^*[W \mid Z=0,A,X] \Big) \Big] \Big\} \\
=& E\Big\{ \frac{2Z-1}{f^*(Z \mid A,X)} \frac{1}{\xi_Z^{W^*}(A,X)} \Big(\sum_Z \delta_A^{W^*}(Z,X) f^*(Z \mid 1-A, X) \frac{f^*(1-A \mid X)}{f^*(A \mid X)} \Big) \\
& \Big[\{R(A,X) - R^*(A,X)\} \xi_Z^W(A,X) Z + \{E[Y \mid Z=0,A,X] - E^*[Y \mid Z=0,A,X]\} \\
& + \{E[W \mid Z=0,A,X] - E^*[W \mid Z=0,A,X]\} R^*(A,X) \Big] \Big\} \\
=& 0
\end{aligned}
$$

because $R^*(A,X) = R(A,X)$, $E^*[Y \mid Z=0,A,X] = E[Y \mid Z=0,A,X]$, $\delta_A^{W^*}(Z=0,X) = \delta_A^{W^*}(Z=0,X)$, and $E^*[W \mid A=0, Z=0, X] = E[W \mid A=0, Z=0, X]$

Thus, under $\mathcal{M}_3$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}] =& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\delta_A^W(Z,X)}] + E[\mathrm{D}^*_{R(1-A,X)}] + E[R^*(1-A,X)\delta_A^{W^*}(Z,X)]\} \\
=& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\delta_A^W(Z,X)}] + E[\mathrm{D}^*_{R(1-A,X)}] + E[R(1-A,X)\delta_A^{W}(Z,X)]\} \\
=& \Delta_{\text{confounded}} - \{0 + 0 + \Delta_{\text{bias}}\} = \Delta
\end{aligned}
$$

In summary, $E[\Delta^*_{\text{mr}}] = \Delta$ under $\mathcal{M}_{\text{union}} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$.

E.3 Variance estimator

Here we provide a consistent estimator of the asymptotic variance $\sigma^2_\Delta(\Delta, \theta^*)$ by writing our problem in the form of standard M-estimation. Recall that $\hat{\theta}$ are the estimated nuisance parameters that solve $\mathbb{P}_n\left\{U_\theta(O;\hat{\theta})\right\} = 0$, and $\hat{\Delta}_{\text{mr}}$ is the proposed multiply robust estimator that solves $\mathbb{P}_n\Big\{EIF_\Delta(O;\hat{\Delta}_{\text{mr}},\hat{\theta})\Big\} = 0$. Let $\gamma = (\theta, \Delta)^\intercal$ denote the vector of all parameters of dimension k, $\psi(\gamma) = \{U_\theta(O;\theta)^\intercal, EIF_\Delta(O;\Delta,\theta)\}^\intercal$, and let $G_n(\gamma) = \mathbb{P}_n\Big\{\psi(\gamma)\Big\} = \frac{1}{n}\sum_{i=1}^n \psi(O_i;\gamma)$ denote a $k \times 1$ vector of estimating functions where the k-th element is the estimating function for Δ, then $\hat{\gamma} = (\hat{\theta}, \hat{\Delta}_{\text{mr}})$ is the solution to the estimating equations $G_n(\gamma) = 0$. Let $A_n(\hat{\gamma}) = -\frac{\partial G_n(\gamma)}{\partial \gamma^\intercal}\Big|_{\gamma=\hat{\gamma}} = -\frac{1}{n}\sum_{i=1}^n \left\{\frac{\partial}{\partial \gamma^\intercal}\psi(O_i;\hat{\gamma})\right\}$ and $B_n(\hat{\gamma}) = \frac{1}{n}\sum_{i=1}^n \psi(O_i;\hat{\gamma})\psi(O_i;\hat{\gamma})^\intercal$. We define the empirical sandwich estimator as follows

$$
\widehat{Var}(\hat{\gamma}) = A_n(\hat{\gamma})^{-1} B_n(\hat{\gamma}) \left(A_n(\hat{\gamma})^{-1}\right)^\intercal .
$$

Then a consistent estimator for the asymptotic variance of $\hat{\Delta}_{\text{mr}}$ corresponds to $\widehat{Var}(\hat{\gamma})_{k,k}$, the (k,k)-th element of $\widehat{Var}(\hat{\gamma})$. In practice, one can also apply the nonparametric bootstrap to estimate the variance.

□

F Proof of Theorem C.1 (efficient influence function in $\mathcal{M}_{\text{nonpar}}$ for polytomous case)

Proof. Let $f(Y, W, A, Z, X; \theta)$ denote a one-dimensional regular parametric submodel of $\mathcal{M}_{\text{nonpar}}$ indexed by θ, under which $\Delta_\theta = E_\theta[\delta^Y_{A,\theta}(Z,X)] - E_\theta[\boldsymbol{R}_\theta(1-A,X)\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A},\theta}(Z,X)]$. The efficient influence function in $\mathcal{M}_{\text{nonpar}}$ is defined as the unique mean zero, finite variance random variable D satisfying

$$\frac{\partial}{\partial\theta}\Big|_{\theta=0}\Delta_\theta = E[D\cdot S(Y, W, A, Z, X)],$$

where $S(\cdot)$ is the score function of the path $f(Y, W, A, Z, X; \theta)$ at $\theta = 0$, and $\frac{\partial}{\partial\theta}\Big|_{\theta=0}\Delta_\theta$ is the pathwise derivative of Δ. To find D, we derive the following pathwise derivatives. First of all, we have the same result as Eq. (21) for the pathwise derivative of $\delta^Y_A(Z,X) = E[Y \mid Z, A{=}1, X] - E[Y \mid Z, A = 0, X]$, which is

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta^Y_{A,\theta}(Z,X)\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(Y - E[Y\mid A,Z,X])S(Y,A,Z,X)\Big|Z,X\Big]\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(Y - \delta^Y_A(Z,X)A - E[Y\mid A{=}0,Z,X])S(Y,A,Z,X)\Big|Z,X\Big].
\end{aligned}$$

Accordingly, the pathwise derivative of $E[\delta^Y_A(Z,X)]$ is given by

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\delta^Y_{A,\theta}(Z,X)]\\
=&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta^Y_A(Z,X)] + E[\delta^Y_A(Z,X)S(Z,X)]\\
=&E[\frac{2A-1}{f(A\mid Z,X)}(Y - E[Y\mid A,Z,X])S(Y,A,Z,X)] + E[\delta^Y_A(Z,X)S(Z,X)]\\
=&E\{[\frac{2A-1}{f(A\mid Z,X)}(Y - \delta^Y_A(Z,X)A - E[Y\mid A{=}0,Z,X]) + \delta^Y_A(Z,X)]\cdot S(Y,A,Z,X)\}.
\end{aligned}$$

Second, for $\delta^{w_i}_A(Z,X) = E[\mathbb{1}(W=w_i)\mid A{=}1,Z,X] - E[\mathbb{1}(W=w_i)\mid A{=}0,Z,X]$ we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta^{W=w_i}_\theta(Z,X)\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(\mathbb{1}(W=w_i) - E[\mathbb{1}(W=w_i)\mid A,Z,X])S(W,A,Z,X)\Big|Z,X\Big]\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(\mathbb{1}(W=w_i) - \delta^{w_i}_A(Z,X)A - E[\mathbb{1}(W=w_i)\mid A{=}0,Z,X])S(W,A,Z,X)\Big|Z,X\Big].
\end{aligned}$$

Generalizing from $\mathbb{1}(W=w_i)$ to vector $\mathbf{\Gamma}_W$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A},\theta}(Z,X)_{k\times 1}\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(\mathbf{\Gamma}_W-E[\mathbf{\Gamma}_W\mid A,Z,X])_{k\times 1}S(W,A,Z,X)\Big|Z,X\Big]\\
=&E\Big[\frac{2A-1}{f(A\mid Z,X)}(\mathbf{\Gamma}_W-\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)A-E[\mathbf{\Gamma}_W\mid A{=}0,Z,X])_{k\times 1}S(W,A,Z,X)\Big|Z,X\Big].
\end{aligned}\tag{37}$$

Third, for $\xi^{Y}_{z_j}(1-A,X)=E[Y\mid Z{=}z_j,1-A,X]-E[Y\mid Z{=}z_0,1-A,X]$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^{Y}_{z_j,\theta}(1-A,X)\\
=&E\Big[\Big(\frac{I(Z{=}z_j)}{f(Z{=}z_j\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big)\Big(Y-E[Y\mid Z,1-A,X]\Big)S(Y,Z,1-A,X)\Big|1-a\Big].
\end{aligned}$$

Generalizing from z_j to vector $\mathbf{Z}$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z},\theta}(1-A,X)^{\intercal}_{1\times k}\\
=&E\Big\{[\Big(\frac{\mathbb{1}(Z=z_k)}{f(Z=z_k\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big),\Big(\frac{\mathbb{1}(Z=z_k-1)}{f(Z=z_k-1\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big),\\
&\ldots,\Big(\frac{\mathbb{1}(Z=z_1)}{f(Z=z_1\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big)]_{1\times k}\Big(Y-E[Y\mid Z,1-A,X]\Big)S(Y,Z,1-A,X)\Big|1-A,X\Big\}\\
=&E\Big\{\mathbf{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\Big(Y-E[Y\mid Z,1-A,X]\Big)S(Y,Z,1-A,X)\Big|1-A,X\Big\},
\end{aligned}\tag{38}$$

where $\mathbf{\Pi}(Z\mid A,X)=\{\frac{\mathbb{1}(Z=z_1)}{f(Z=z_1|A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0|A,X)},\frac{\mathbb{1}(Z=z_2)}{f(Z=z_2|A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0|A,X)},\ldots,\frac{\mathbb{1}(Z=z_k)}{f(Z=z_k|A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0|A,X)}\}^{\intercal}$ denote a $k\times 1$ vector generalizing $\frac{2Z-1}{f(Z|A,X)}$ in the binary case, with $\mathbf{\Pi}(Z\mid A,X)_j=\frac{\mathbb{1}(Z=z_j)}{f(Z=z_j|A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0|A,X)}$.

Forth, for $\xi^{w_i}_{z_j}(1-A,X)=E[\mathbb{1}(W=w_i)\mid Z{=}z_j,1-A,X]-E[\mathbb{1}(W=w_i)\mid Z{=}z_0,1-A,X]$, we have

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\xi^{w_i}_{z_j,\theta}(1-A,X)\\
=&E\Big[\Big(\frac{\mathbb{1}(Z=z_j)}{f(Z=z_j\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big)\Big(\mathbb{1}(W=w_i)-E[\mathbb{1}(W=w_i)\mid Z,1-A,X]\Big)\\
&S(W,Z,1-A,X)\Big|1-A,X\Big].
\end{aligned}\tag{39}$$

Generalizing to column vector $\mathbf{\Gamma}_W$, we have

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{W}}_{z_j,\theta}(1-A,X)_{k\times 1}\\
=&E\Big[\Big(\frac{\mathbb{1}(Z=z_j)}{f(Z=z_j\mid 1-A,X)}-\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid 1-A,X)}\Big)\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z,1-A,X]\Big)_{k\times 1}\\
&S(W,Z,1-A,X)\Big|1-A,X\Big].
\end{aligned}\tag{40}
$$

Generalizing to row vector $\mathbf{Z}$, we have

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{w_i}_{\boldsymbol{Z},\theta}(1-A,X)_{1\times k}\\
=&E\Big[\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\Big(\mathbb{1}(W=w_i)-E[\mathbb{1}(W=w_i)\mid Z,1-A,X]\Big)S(W,Z,1-A,X)\Big|1-A,X\Big].
\end{aligned}\tag{41}
$$

Generalizing to a matrix, we have

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z},\theta}(1-A,X)_{k\times k}\\
=&E\Big[\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z,1-A,X]\Big)_{k\times 1}\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}S(W,Z,1-A,X)\Big|1-A,X\Big].
\end{aligned}\tag{42}
$$

From the above, we finally have

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}_\theta(1-A,X)_{1\times k}=\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z},\theta}(1-A,X)^{\intercal}_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z},\theta}(1-A,X)^{-1}_{k\times k}\\
=&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z},\theta}(1-A,X)^{\intercal}_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}_{k\times k}+\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z}}(1-A,X)^{\intercal}_{1\times k}\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z},\theta}(1-A,X)^{-1}_{k\times k}\\
\\
=&E\Big[\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\Big(Y-E[Y\mid Z,1-A,X]\Big)S(Y,Z,1-A,X)\Big|1-A,X\Big]\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}-\\
&\boldsymbol{R}(1-A,X)_{1\times k}\cdot E\Big[\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z,1-A,X]\Big)_{k\times 1}\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}S(W,Z,1-A,X)\Big]\Big|1-A,X\Big]\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}\\
=&E\Big[\Big(Y-E[Y\mid Z,1-A,X]\Big)\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}S(Y,Z,1-A,X)\\
&-\boldsymbol{R}(1-A,X)_{1\times k}\cdot\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z,1-A,X]\Big)_{k\times 1}\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}S(W,Z,1-A,X)\Big|1-A,X\Big]\\
=&E\Big\{\Big[Y-E[Y\mid Z,1-A,X]-\boldsymbol{R}(1-A,X)_{1\times k}\cdot\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z,1-A,X]\Big)_{k\times 1}\Big]\cdot\\
&\boldsymbol{\Pi}(Z\mid A,X)^{\intercal}_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(1-A,X)^{-1}S(Y,W,Z,1-A,X)\Big|1-A,X\Big\}
\end{aligned}\tag{43}
$$

Recall that $\boldsymbol{\Gamma}_Z=\{\mathbb{1}(Z=z_1),\mathbb{1}(Z=z_2),\ldots,\mathbb{1}(Z=z_k)\}^{\intercal}$ denote a $k\times 1$ vector generalizing the binary Z, with $\boldsymbol{\Gamma}_{Zi}=\mathbb{1}(Z=z_i)$. We note that

$$
E[Y\mid Z,1-A,X]\mathbb{1}_{1\times k}=\boldsymbol{\xi}^{\boldsymbol{Y}}_{\boldsymbol{Z}}(1-A,X)^{\intercal}_{1\times k}\mathrm{diag}[\boldsymbol{\Gamma}_Z]_{k\times k}+E[Y\mid Z=z_0,1-A,X]\mathbb{1}_{1\times k},
$$

and

$$E[\mathbf{\Gamma}_W \mid Z, 1-A, X]_{k\times 1}\mathbb{1}_{1\times k} = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1-A, X)_{k\times k}\text{diag}[\mathbf{\Gamma}_Z]_{k\times k} + E[\mathbf{\Gamma}_W \mid Z = z_0, 1-A, X]_{k\times 1}\mathbb{1}_{1\times k}.$$

Therefore,

$$\begin{aligned}
&\Big(Y - E[Y \mid Z, 1-A, X] - \boldsymbol{R}(1-A, X)_{1\times k} \cdot \Big(\mathbf{\Gamma}_W - E[\mathbf{\Gamma}_W \mid Z, 1-A, X]\Big)_{k\times 1}\Big)\mathbb{1}_{1\times k} \\
=&\Big(Y - \boldsymbol{R}(1-A, X)\mathbf{\Gamma}_W\Big)\mathbb{1}_{1\times k} - \Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(1-A, X)^{\intercal}_{1\times k}\text{diag}[\mathbf{\Gamma}_Z]_{k\times k} + E[Y \mid Z = z_0, 1-A, X]\mathbb{1}_{1\times k}\Big) \\
&+ \boldsymbol{R}(1-A, X)_{1\times k}\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1-A, X)_{k\times k}\text{diag}[\mathbf{\Gamma}_Z]_{k\times k} + E[\mathbf{\Gamma}_W \mid Z = z_0, 1-A, X]_{k\times 1}\mathbb{1}_{1\times k}\Big) \\
=&\Big[Y - E[Y \mid Z = z_0, 1-A, X] - \boldsymbol{R}(1-A, X)_{1\times k}\Big(\mathbf{\Gamma}_W - E[\mathbf{\Gamma}_W \mid Z = z_0, 1-A, X]\Big)_{k\times 1}\Big]\mathbb{1}_{1\times k}.
\end{aligned}$$

Thus, Eq. (43) can be simplified to

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}_\theta(1-A, X)_{1\times k} \\
=&E\Big\{\Big[Y - E[Y \mid Z = z_0, 1-A, X] - \boldsymbol{R}(1-A, X)_{1\times k} \cdot \Big(\mathbf{\Gamma}_W - E[\mathbf{\Gamma}_W \mid Z = z_0, 1-A, X]\Big)_{k\times 1}\Big] \cdot \\
&\mathbf{\Pi}(Z \mid A, X)^{\intercal}_{1\times k}\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1-A, X)\Big)^{-1}_{k\times k} S(Y, W, Z, 1-A, X)\Big| 1-A, X\Big\}
\end{aligned} \tag{44}$$

Now we consider the pathwise derivative of $E[\boldsymbol{R}(1-A, X)_{1\times k}\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)_{k\times k}]$. Note that

$$\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\boldsymbol{R}_\theta(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A},\boldsymbol{\theta}}^{\boldsymbol{W}}(Z, X)] \\
=&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A},\boldsymbol{\theta}}^{\boldsymbol{W}}(Z, X)] + E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)] + E[\boldsymbol{R}(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)S(A, Z, X)],
\end{aligned} \tag{45}$$

thus we consider $E[\boldsymbol{R}(1-A, X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\delta_{A,\theta}^{W}(Z, X)]$ and $E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)]$ seperately.

First, by Eq. (37), and using similar argument as the derivation of Eq. (28) in Appendix D, we have

$$\begin{aligned}
&E[\boldsymbol{R}(1-A, X)\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{\delta}_{\boldsymbol{A},\theta}^{\boldsymbol{W}}(Z, X)] \\
=&E\Big[e_{\boldsymbol{R}}(Z, X)_{1\times k}(\mathbf{\Gamma}_W - \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)A - E[\mathbf{\Gamma}_W \mid A=0, Z, X])_{k\times 1}\frac{2A-1}{f(A \mid Z, X)}S(W, A, Z, X)\Big].
\end{aligned} \tag{46}$$

where $e_{\boldsymbol{R}}(z, x) = E[\boldsymbol{R}(1-A, X) \mid Z=z, X=x]$.

Second, we consider $E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}_\theta(1-A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)]$. By Eq. (44) and using the argument as the derivation of Eq. (32), we have

$$
\begin{aligned}
&E[\frac{\partial}{\partial\theta}\Big|_{\theta=0}\boldsymbol{R}_\theta(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)] \\
=&E\Big\{\Big[Y-E[Y\mid Z=z_0,A,X]-\boldsymbol{R}(A,X)\cdot\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z=z_0,A,X]\Big)\Big] \\
&\boldsymbol{\Pi}(Z\mid A,X)^\intercal\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A)\Big)^{-1}\Big(E[\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big)S(Y,W,A,Z,X)\Big\}.
\end{aligned} \tag{47}
$$

Combining Eq. (45), (46), and (47) we have

$$
\begin{aligned}
&\frac{\partial}{\partial\theta}\Big|_{\theta=0}E_\theta[\boldsymbol{\delta}_{\boldsymbol{A},\boldsymbol{\theta}}^{\boldsymbol{W}}(Z,X)\cdot\boldsymbol{R}_\theta(1-A,X)] \\
=&E\Big\{\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X) \\
+&E[\boldsymbol{R}(1-A,X)\mid Z,X](\boldsymbol{\Gamma}_W-\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)A-E[\boldsymbol{\Gamma}_W\mid A=0,Z,X])\frac{2A-1}{f(A\mid Z,X)} \\
+&\Big[Y-E[Y\mid Z=z_0,A,X]-\boldsymbol{R}(A,X)\cdot\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z=z_0,A,X]\Big)\Big] \\
&\boldsymbol{\Pi}(Z\mid A,X)^\intercal\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A)\Big)^{-1}\Big(E[\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big)S(Y,W,A,Z,X)\Big\}.
\end{aligned}
$$

Therefore, the influence function for $\Delta=E[\delta_A^Y(Z,X)]-E[\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)]$ is given by

$$
\begin{aligned}
&\mathrm{IF}_\Delta(Y,W,A,Z,X) \\
=&\frac{2A-1}{f(A\mid Z,X)}\Big(Y-\delta_A^Y(Z,X)A-E[Y\mid A=0,Z,X]\Big) \\
-&E[\boldsymbol{R}(1-A,X)\mid Z,X]\cdot\Big(\boldsymbol{\Gamma}_W-\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)A-E[\boldsymbol{\Gamma}_W\mid A=0,Z,X]\Big)\frac{2A-1}{f(A\mid Z,X)} \\
-&\Big[Y-E[Y\mid Z=z_0,A,X]-\boldsymbol{R}(A,X)\Big(\boldsymbol{\Gamma}_W-E[\boldsymbol{\Gamma}_W\mid Z=z_0,A,X]\Big)\Big]\cdot \\
&\boldsymbol{\Pi}(Z\mid A,X)^\intercal\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A)\Big)^{-1}\Big(E[\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)\mid 1-A,X]\frac{f(1-A\mid X)}{f(A\mid X)}\Big) \\
+&\delta_A^Y(Z,X)-\boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)-\Delta.
\end{aligned}
$$

□

G Proof of Theorem C.2

Proof. Under the regularity conditions given in Theorem 3.2 of Newey and McFadden (1994), the estimated nuisance parameters

$$\hat{\theta} = \{(\hat{\alpha}^{A,Z}_{\text{mle}})^{\intercal}, (\hat{\beta}^{Y}_{\text{mle}})^{\intercal}, (\hat{\beta}^{W0}_{\text{mle}})^{\intercal}, (\hat{\beta}^{WZ}_{\text{dr}})^{\intercal}, (\hat{\beta}^{WA}_{\text{dr}})^{\intercal}, (\hat{\beta}^{R}_{\text{dr}})^{\intercal}\}^{\intercal}$$

from solving the moment function $\mathbb{P}_n\{U_\theta(O;\theta)\} = 0$ are asymptotically normal and converge at $o(n^{-1/2})$ rate to some fixed values

$$\theta^* = \{(\alpha^{A,Z}_*)^{\intercal}, (\beta^{Y}_*)^{\intercal}, (\beta^{W0}_*)^{\intercal}, (\beta^{WZ}_*)^{\intercal}, (\beta^{WA}_*)^{\intercal}, (\beta^{R}_*)^{\intercal}\}^{\intercal}$$

satisfying $E[U_\theta(O;\theta^*)] = 0$ regardless of whether the corresponding nuisance models are correctly specified. Accordingly, we let $\text{D}^*_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}$, $\text{D}^*_{\boldsymbol{R}(1-A,X)}$, $\boldsymbol{R}^*(1-A,X)$, $E^*[Y \mid Z = z_0, A, X]$, $E^*[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X]$, $\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)$, $\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X)$, $E^*[\boldsymbol{\Gamma}_W \mid Z = z_0, A, X]$, $f^*(A \mid Z,X)$, $f^*(Z \mid A,X)$, and $f^*(A \mid X)$ denote the probability limit of the estimated nuisance models. Similarly, we let Δ^*_{mr}, $\Delta^*_{\text{confounded,dr}}$, and $\Delta^*_{\text{bias,mr}}$ denote the limit of the estimated parameters of interest. In addition, recall that $\boldsymbol{\Pi}(Z \mid A,X) = \{\mathbb{1}(Z = z_1)/f(Z = z_1 \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X), \mathbb{1}(Z = z_2)/f(Z = z_2 \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X), \ldots, \mathbb{1}(Z = z_k)/f(Z = z_k \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X)\}^{\intercal}$ denote a $k \times 1$ vector generalizing $(2Z-1)/f(Z \mid A,X)$ in the binary case, with $\boldsymbol{\Pi}(Z \mid A,X)_j = \mathbb{1}(Z = z_j)/f(Z = z_j \mid A,X) - \mathbb{1}(Z = z_0)/f(Z = z_0 \mid A,X)$. With slight abuse of notation, we let $\boldsymbol{\Pi}^*(Z \mid A,X)$ denote the limit of $\boldsymbol{\Pi}(Z \mid A,X)$.

We start with showing that $\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X) = \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)$ and $\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X) = \boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)$ under $\mathcal{M}_2 \cup \mathcal{M}_3$, and $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$ under $\mathcal{M}_1 \cup \mathcal{M}_3$. First, we show that $\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X) = \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)$ and $\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X) = \boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)$ under $\mathcal{M}_2 \cup \mathcal{M}_3$. It suffice to show that $\delta^{w_i^*}_A(Z,X) = \delta^{w_i}_A(Z,X)$ and $\xi^{w_i^*}_Z(A,X) = \xi^{w_i}_Z(A,X)$ for all i. Or equivalently $(\beta^{W_iA}_*, \beta^{W_iZ}_*) = (\beta^{W_iA}, \beta^{W_iZ})$ where $(\beta^{W_iA}, \beta^{W_iZ})$ is the subset of (β^{WA}, β^{WZ}) corresponding to the i-th level of W. Recall that $\hat{\beta}^{W_iA}_{\text{dr}}$ and $\hat{\beta}^{W_iZ}_{\text{dr}}$ solves $\mathbb{P}_n\Big\{U_{\beta^{W_iA},\beta^{W_iZ}}(\hat{\beta}^{W_iA}_{\text{dr}}, \hat{\beta}^{W_iZ}_{\text{dr}})\Big\} = 0$ with $\lim_{n\to\infty} \mathbb{P}_n\Big\{U_{\beta^{W_iA},\beta^{W_iZ}}(\hat{\beta}^{W_iA}_{\text{dr}}, \hat{\beta}^{W_iZ}_{\text{dr}})\Big\} = E[U_{\beta^{W_iA},\beta^{W_iZ}}(\beta^{W_iA}_*, \beta^{W_iZ}_*)]$. Now consider $E[U_{\beta^{W_iA},\beta^{W_iZ}}(\beta^{W_iA}_*, \beta^{W_iZ}_*)]\Big|_{\beta^{W_iA}_*=\beta^{W_iA},\beta^{W_iZ}_*=\beta^{W_iZ}}$ under $\mathcal{M}_2$ where $\xi^{W_i}_Z(A,X;\beta^{W_iZ})$ and $\delta^{W_i}_A(Z,X;\beta^{W_iA})$ are correctly specified, i.e. $\xi^{W_i}_Z(A,X) = \xi^{W_i}_Z(A,X;\beta^{W_iZ})$ and $\delta^{W_i}_A(Z,X) =$

$\delta_A^{W_i}(Z, X; \beta^{W_iA})$, we have

$$
\begin{aligned}
&E[U_{\beta^{W_iA},\beta^{W_iZ}}(\beta^{W_iA}, \beta^{W_iZ})] \\
=&E\{\Big[g_0(A,Z,X) - E^*[g_0(A,Z,X) \mid X]\Big]\Big[\boldsymbol{\Gamma}_{Wi} - E[\boldsymbol{\Gamma}_{Wi} \mid A, Z, X; \beta_*^{W0}, \beta^{W_iZ}, \beta^{W_iA}]\Big]\} \\
=&E\{\Big[g_0(A,Z,X) - E^*[g_0(A,Z,X) \mid X]\Big]\Big[E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] - E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X; \beta_*^{W0}]+ \\
&[\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A=0,X) - \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A=0,X;\beta^{W_iZ})]\boldsymbol{\Gamma}_Z + [\delta_A^{W_i}(Z=z_0,X) - \delta_A^{W_i}(Z=z_0,X;\beta^{W_iA})]A+ \\
&[\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X) - \boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X;\beta^{WAZ})]A\boldsymbol{\Gamma}_Z\Big]\} \\
=&E\{\Big[g_0(A,Z,X) - E[g_0(A,Z,X) \mid X]\Big]\Big[E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] - E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X; \beta_*^{W0}]\Big]\} \\
=&0
\end{aligned}
$$

because $E\Big[\{g_0(A,Z,X) - E[g_0(A,Z,X) \mid X]\}h(X)\Big] = 0$ for any function h. Thus, under $\mathcal{M}_2$ where $\xi_Z^{W_i}(A, X; \beta^{W_iZ})$ and $\delta_A^{W_i}(Z, X; \beta^{W_iA})$ are correctly specified, $\mathbb{P}_n\Big\{U_{\beta^{W_iA},\beta^{W_iZ}}(\hat{\beta}_{\text{dr}}^{W_iA}, \hat{\beta}_{\text{dr}}^{W_iZ})\Big|_{\hat{\beta}_{\text{dr}}^{W_iA}=\beta^{W_iA}, \hat{\beta}_{\text{dr}}^{W_iZ}=\beta^{W_iZ}}\Big\}$ converges to zero, i.e. $(\beta^{W_iA}, \beta^{W_iZ})$ is a solution to the probability limit of $\mathbb{P}_n\Big\{U_{\beta^{W_iA},\beta^{W_iZ}}(\hat{\beta}_{\text{dr}}^{W_iA}, \hat{\beta}_{\text{dr}}^{W_iZ})\Big\} = 0$. Thus $\beta_*^{W_iA} = \beta^{W_iA}$ and $\beta_*^{W_iZ} = \beta^{W_iZ}$, and $\delta_A^{w_i^*}(Z, X) = \delta_A^{w_i}(Z, X)$ and $\xi_Z^{w_i^*}(A, X) = \xi_Z^{w_i}(A, X)$ for all i. Therefore $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W^*}}(Z, X) = \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)$ and $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W^*}}(A, X) = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A, X)$ under $\mathcal{M}_2$.

Similar arguments apply to the scenario under $\mathcal{M}_3$. Under $\mathcal{M}_3$ where working models $\boldsymbol{R}(A, X; \beta^R)$, $E[Y \mid Z = z_0, A, X; \beta^Y]$, $\xi_Z^{W_i}(A, X; \beta^{W_iZ})$, $\delta_A^{W_i}(Z, X; \beta^{W_iA})$, and $E[\boldsymbol{\Gamma}_{Wi} \mid A = 0, Z = z_0, X; \beta^{W_i}]$ are correctly specified, we have $\beta_*^{W0} = \beta^{W0}$ and thus $E^*[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X] = E[\boldsymbol{\Gamma}_{Wi} \mid A = 0, Z = z_0, X]$. We again consider

$$
\begin{aligned}
&E[U_{\beta^{W_iA},\beta^{W_iZ}}(\beta^{W_iA}, \beta^{W_iZ})] \\
=&E\{\Big[g_0(A,Z,X) - E^*[g_0(A,Z,X) \mid X]\Big]\Big[\boldsymbol{\Gamma}_{Wi} - E[\boldsymbol{\Gamma}_{Wi} \mid A, Z, X; \beta_*^{W0}, \beta^{W_iZ}, \beta^{W_iA}]\Big]\} \\
=&E\{\Big[g_0(A,Z,X) - E^*[g_0(A,Z,X) \mid X]\Big]\Big[E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] - E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X; \beta_*^{W0}]+ \\
&[\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A=0,X) - \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{w_i}}(A=0,X;\beta^{W_iZ})]\boldsymbol{\Gamma}_Z + [\delta_A^{W_i}(Z=z_0,X) - \delta_A^{W_i}(Z=z_0,X;\beta^{W_iA})]A+ \\
&[\boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X) - \boldsymbol{\eta}_{\boldsymbol{AZ}}^{\boldsymbol{w_i}}(X;\beta^{WAZ})]A\boldsymbol{\Gamma}_Z\Big]\} \\
=&E\{\Big[g_0(A,Z,X) - E^*[g_0(A,Z,X) \mid X]\Big]\Big[E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] - E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X; \beta_*^{W0}]\Big]\} \\
=&0
\end{aligned}
$$

because $E[\boldsymbol{\Gamma}_{Wi} \mid A = 0, Z = z_0, X; \beta_*^{W0}] = E[\boldsymbol{\Gamma}_{Wi} \mid A = 0, Z = z_0, X]$. Therefore $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W^*}}(Z, X) =$

$\boldsymbol{\delta_A^W}(Z,X)$ and $\boldsymbol{\xi_Z^{W*}}(A,X) = \boldsymbol{\xi_Z^W}(A,X)$ under $\mathcal{M}_3$. In addition, we have that

$$\begin{aligned} E^*[\boldsymbol{\Gamma}_{Wi} \mid A,Z,X] =& E^*[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] + \delta_A^{w_i^*}(Z=z_0,X)A + \boldsymbol{\xi_Z^{w_i^*}}(A,X)\boldsymbol{\Gamma}_Z + \boldsymbol{\eta_{AZ}^{w_i^*}}(X)A\,\boldsymbol{\Gamma}_Z \\ =& E[\boldsymbol{\Gamma}_{Wi} \mid A=0, Z=z_0, X] + \delta_A^{w_i}(Z=z_0,X)A + \boldsymbol{\xi_Z^{w_i}}(A,X)\boldsymbol{\Gamma}_Z + \boldsymbol{\eta_{AZ}^{w_i}}(X)A\,\boldsymbol{\Gamma}_Z \\ =& E[\boldsymbol{\Gamma}_{Wi} \mid A,Z,X], \end{aligned} \tag{48}$$

i.e., $E^*[\boldsymbol{\Gamma}_W \mid A,Z,X] = E[\boldsymbol{\Gamma}_W \mid A,Z,X]$.

Second, we show that $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$ under $\mathcal{M}_1 \cup \mathcal{M}_3$. Under $\mathcal{M}_1$ where working models $f(A,Z \mid X;\alpha^{A,Z})$ and $\boldsymbol{R}(A,X;\beta^R)$ are correctly specified, we have $\alpha_*^{A,Z} = \alpha^{A,Z}$, $f^*(A,Z \mid X) = f(A,Z \mid X)$, and thus $E^*[g_1(A,Z,X) \mid A,X] = E[g_1(A,Z,X) \mid A,X]$ for any function $g_1(A,Z,X)$. Recall that $\hat{\beta}_{\text{dr}}^R$ solves $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = 0$ with $\lim_{n\to\infty}\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = E[U_{\beta^R}(\beta_*^R)]$. Now consider $E[U_{\beta^R}(\beta_*^R)]\Big|_{\beta_*^R=\beta^R}$ under $\mathcal{M}_1$ where $\boldsymbol{R}(A,X;\beta^R)$ is correctly specified, i.e. $\boldsymbol{R}(A,X) = \boldsymbol{R}(A,X;\beta^R)$, we have

$$\begin{aligned} &E[U_{\beta^R}(\beta^R)] = E\Big\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A,X]\Big]\Big[Y - E^*[Y \mid Z=z_0,A,X] - \\ &\boldsymbol{R}(A,X;\beta^R)(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X])\Big]\Big\} \\ =&E\Big\{\Big[g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\Big]\Big[\{\boldsymbol{R}(A,X) - \boldsymbol{R}(A,X;\beta^R)\}\boldsymbol{\xi_Z^W}(A,X)Z + \\ &\{E[Y \mid Z=z_0,A,X] - E^*[Y \mid Z=z_0,A,X]\} + \{E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\}\boldsymbol{R}(A,X;\beta^R)\Big]\Big\} \\ =&E\Big\{\Big[g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\Big]\Big[\{E[Y \mid Z=z_0,A,X] - E^*[Y \mid Z=z_0,A,X]\} + \\ &\{E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\}\boldsymbol{R}(A,X;\beta^R)\Big]\Big\} \\ =&0 \end{aligned}$$

because $E\Big[\{g_1(A,Z,X) - E[g_1(A,Z,X) \mid A,X]\}h(A,X)\Big] = 0$ for any function h. Thus, under $\mathcal{M}_1$ where $\boldsymbol{R}(A,X;\beta^R)$ is correctly specified, $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big|_{\hat{\beta}_{\text{dr}}^R=\beta^R}\Big\}$ converges to zero, i.e. β^R is a solution to the probability limit of $\mathbb{P}_n\Big\{U_{\beta^R}(\hat{\beta}_{\text{dr}}^R)\Big\} = 0$. Thus $\beta_*^R = \beta^R$ and $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$ under $\mathcal{M}_1$.

Similar arguments apply to the scenario under $\mathcal{M}_3$. Under $\mathcal{M}_3$ where working models $\boldsymbol{R}(A,X;\beta^R)$, $E[Y \mid Z = z_0,A,X;\beta^Y]$ and $E[\boldsymbol{\Gamma}_W \mid A,Z,X;\beta^{W_i}]$ are correctly specified, we have $E^*[Y \mid Z = z_0,A,X] = E[Y \mid Z=z_0,A,X]$ and by (48) we have $E^*[\boldsymbol{\Gamma}_W \mid A,Z,X] = E[\boldsymbol{\Gamma}_W \mid A,Z,X]$. We again consider

$$
\begin{aligned}
&E[U_{\beta^R}(\beta^R)] = E\Big\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A,X]\Big]\Big[Y - E^*[Y \mid Z=z_0,A,X] - \\
&\boldsymbol{R}(A,X;\beta^R)(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X])\Big]\Big\} \\
=&E\Big\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A,X]\Big]\Big[\{\boldsymbol{R}(A,X) - \boldsymbol{R}(A,X;\beta^R)\}\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)Z + \\
&\{E[Y \mid Z=z_0,A,X] - E^*[Y \mid Z=z_0,A,X]\} + \{E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\}\boldsymbol{R}(A,X;\beta^R)\Big]\Big\} \\
=&E\Big\{\Big[g_1(A,Z,X) - E^*[g_1(A,Z,X) \mid A,X]\Big]\Big[\{E[Y \mid Z=z_0,A,X] - E^*[Y \mid Z=z_0,A,X]\} + \\
&\{E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\}\boldsymbol{R}(A,X;\beta^R)\Big]\Big\} \\
=&0
\end{aligned}
$$

because $E^*[Y \mid Z=z_0,A,X] = E[Y \mid Z=z_0,A,X]$ and $E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] = E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]$. Therefore $\beta_*^R = \beta^R$ and $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$ under $\mathcal{M}_3$.

We now show that $E[\Delta^*_{\text{mr}}] = \Delta$ under $\mathcal{M}_{\text{union}}$. To this end, we consider $E[\Delta^*_{\text{confounded}}]$, $E[\text{D}^*_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)}]$, and $E[\text{D}^*_{\boldsymbol{R}(1-A,X)}]$ respectively. Under $\mathcal{M}_1$ where working models $f(A,Z \mid X;\alpha^{A,Z})$ and $R(A,X;\beta^R)$ are correctly specified, we have $f^*(A,Z \mid X) = f(A,Z \mid X)$ and $R^*(A,X) = R(A,X)$. First we consider

$$
E[\Delta^*_{\text{confounded}}] = E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E^*[Y \mid A,Z,X]\Big) + E^*[Y \mid A{=}1,Z,X] - E^*[Y \mid A{=}0,Z,X]\Big].
$$

When $f^*(A \mid Z,X) = f(A \mid Z,X)$, by Eq. (34) we have

$$
\begin{aligned}
E[\Delta^*_{\text{confounded}}] =&E\Big[\frac{2A-1}{f(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E^*[Y \mid A,Z,X]\Big) + E^*[Y \mid A{=}1,Z,X] - E^*[Y \mid A{=}0,Z,X]\Big] \\
=&E[\delta_A^Y(Z,X) - \delta_A^{Y^*}(Z,X) + \delta_A^{Y^*}(Z,X)] = \Delta_{\text{confounded}}.
\end{aligned}
$$

Second, consider

$$
\begin{aligned}
E[\text{D}^*{}_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)}] =& E\Big[\Big(\sum_A \boldsymbol{R}^*(1{-}A,X)f^*(A \mid Z,X)\Big)\Big(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid A,Z,X]\Big)\frac{2A-1}{f^*(A \mid Z,X)}\Big] \\
=&E\Big[\Big(\sum_A \boldsymbol{R}^*(1{-}A,X)f^*(A \mid Z,X)\Big)\Big(E[\boldsymbol{\Gamma}_W \mid A,Z,X] - E^*[\boldsymbol{\Gamma}_W \mid A,Z,X]\Big)\frac{2A-1}{f^*(A \mid Z,X)}\Big]
\end{aligned}
$$

When $f^*(A \mid Z,X) = f(A \mid Z,X)$, $\sum_A \boldsymbol{R}^*(1{-}A,X)f^*(A \mid Z,X) = E[\boldsymbol{R}^*(1{-}A,X) \mid Z,X]$. By Eq. (34) we have

$$
E[\text{D}^*_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)}] = E\Big[E[\boldsymbol{R}^*(1{-}A,X) \mid Z,X]\Big(\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X) - \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)\Big)\Big].
$$

Because we also have $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$,

$$\begin{aligned}
&E[\mathrm{D}^*_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)} + \boldsymbol{R}^*(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)] \\
=&E\Big[E[\boldsymbol{R}(1-A,X) \mid Z,X]\Big(\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X) - \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)\Big) + \boldsymbol{R}(1-A,X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)\Big] = \Delta_{\text{bias}}.
\end{aligned}$$

Third, we consider $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]$. Because

$$E[Y \mid A,Z,X]\mathbb{1}_{1\times k} = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{Y}}(A,X)^{\intercal}_{1\times k}\text{diag}[\boldsymbol{\Gamma}_Z]_{k\times k} + E[Y \mid Z=z_0,A,X]\mathbb{1}_{1\times k},$$

and

$$E[\boldsymbol{\Gamma}_W \mid A,Z,X]_{k\times 1}\mathbb{1}_{1\times k} = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)_{k\times k}\text{diag}[\boldsymbol{\Gamma}_Z]_{k\times k} + E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]_{k\times 1}\mathbb{1}_{1\times k},$$

we have

$$\begin{aligned}
&E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] \\
=&E\Big\{\Big[Y - E^*[Y \mid Z=z_0,A,X] - \boldsymbol{R}^*(A,X)\Big(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\Big)\Big] \\
&\Big(\boldsymbol{\Pi}^*(Z \mid A,X)\Big)\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}^*}(A,X)^{-1}\Big(\sum_Z \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)f^*(Z \mid 1-A,X)\cdot\frac{f^*(1-A \mid X)}{f^*(A \mid X)}\Big)\Big\} \\
=&E\Big\{\Big[\{\boldsymbol{R}(A,X) - \boldsymbol{R}^*(A,X)\}_{1\times k}\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A,X)_{k\times k}\text{diag}[\boldsymbol{\Gamma}_Z]_{k\times k} + \{E[Y \mid Z=z_0,A,X] - E^*[Y \mid Z=z_0,A,X]\}\mathbb{1}_{1\times k} \\
&+ \boldsymbol{R}^*(A,X)_{1\times k}\{E[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X] - E^*[\boldsymbol{\Gamma}_W \mid Z=z_0,A,X]\}_{k\times 1}\mathbb{1}_{1\times k}\Big] \\
&\text{diag}[\boldsymbol{\Pi}^*(Z \mid A,X)]_{k\times k}\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}^*}(A,X)^{-1}\Big(\sum_Z \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z,X)f^*(Z \mid 1-A,X)\Big)\frac{f^*(1-A \mid X)}{f^*(A \mid X)}\Big\}.
\end{aligned} \tag{49}$$

When $\boldsymbol{\Pi}^*(Z \mid A,X) = \boldsymbol{\Pi}(Z \mid A,X)$, by similar argument as Eq. (34), we have for any function $h(A,Z,X)$

$$E[h(A,Z,X)\mathbb{1}_{1\times k}\text{diag}[\boldsymbol{\Pi}(Z \mid A,X)]_{k\times k}] = E[\{h(A,z_1,X)-h(A,z_0,X),\ldots,h(A,z_k,X)-h(A,z_0,X)\}]_{1\times k}. \tag{50}$$

Thus, Eq. (49) can be simplified to

$$
\begin{aligned}
&E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] \\
=&E\Big\{[\boldsymbol{R}(A,X)-\boldsymbol{R}^*(A,X)]_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)_{k\times k}\text{diag}\left[\boldsymbol{\Pi}(Z\mid A,X)+\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid A,X)}\right]_{k\times k} \qquad (51)\\
&\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X)^{-1}\Big(\sum_Z\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)f(Z\mid 1-A,X)\Big)\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big\},
\end{aligned}
$$

where

$$
\begin{aligned}
\text{diag}\left[\boldsymbol{\Pi}(Z\mid A,X)+\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid A,X)}\right]&=\text{diag}[\boldsymbol{\Gamma}_Z]\text{diag}[\boldsymbol{\Pi}(Z\mid A,X)]\\
=\text{diag}\{\frac{\mathbb{1}(Z=z_1)}{f(Z=z_1\mid A,X)},&\frac{\mathbb{1}(Z=z_2)}{f(Z=z_1\mid A,X)},\dots,\frac{\mathbb{1}(Z=z_k)}{f(Z=z_k\mid A,X)}\}.
\end{aligned}
$$

Because $E[\text{diag}\left[\boldsymbol{\Pi}(Z\mid A,X)+\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0|A,X)}\right]_{k\times k}\mid A,X]=\mathrm{I}_{k\times k}$, we can simplify Eq. (51) as follows

$$
\begin{aligned}
&E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] \\
=&E\Big\{[\boldsymbol{R}(A,X)-\boldsymbol{R}^*(A,X)]_{1\times k}\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)_{k\times k}\text{diag}\left[\boldsymbol{\Pi}(Z\mid A,X)+\frac{\mathbb{1}(Z=z_0)}{f(Z=z_0\mid A,X)}\right]_{k\times k}\\
&\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X)^{-1}\Big(\sum_Z\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)f(Z\mid 1-A,X)\Big)\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big\}\\
=&E\Big\{[\boldsymbol{R}(A,X)-\boldsymbol{R}^*(A,X)]\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X)^{-1}\Big(\sum_Z\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)f(Z\mid 1-A,X)\Big)\frac{f^*(1-A\mid X)}{f^*(A\mid X)}\Big\},
\end{aligned}
\qquad (52)
$$

We can see that when $\boldsymbol{R}^*(A,X)=\boldsymbol{R}(A,X)$ we have $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]=0$.

In summary, under $\mathcal{M}_1$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}]=&E[\Delta^*_{\text{confounded}}]-\{E[\mathrm{D}^*_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}+\boldsymbol{R}^*(1-A,X)\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)]+E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]\}\\
=&\Delta_{\text{confounded}}-\{\Delta_{\text{bias}}+0\}=\Delta
\end{aligned}
$$

Under $\mathcal{M}_2$ where $f(A,Z\mid X;\alpha^{A,Z})$, $\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X;\beta^{WZ})$ and $\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X;\beta^{WA})$ are correctly specified, we have $f^*(A,Z\mid X)=f(A,Z\mid X)$, $\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)=\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)$ and $\boldsymbol{\xi}^{\boldsymbol{W}^*}_{\boldsymbol{Z}}(A,X)=\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)$. First we consider

$$
E[\Delta^*_{\text{confounded}}]=E\Big[\frac{2A-1}{f^*(A\mid Z,X)}\Big(E[Y\mid A,Z,X]-E^*[Y\mid A,Z,X]\Big)+E^*[Y\mid A=1,Z,X]-E^*[Y\mid A=0,Z,X]\Big].
$$

When $f^*(A \mid Z, X) = f(A \mid Z, X)$, by Eq. (34) we have

$$E[\Delta^*_{\text{confounded}}] = E\Big[\frac{2A-1}{f(A \mid Z, X)}\Big(E[Y \mid A, Z, X] - E^*[Y \mid A, Z, X]\Big) + E^*[Y \mid A{=}1, Z, X] - E^*[Y \mid A{=}0, Z, X]\Big]$$
$$= E[\delta_A^Y(Z, X) - \delta_A^{Y^*}(Z, X) + \delta_A^{Y^*}(Z, X)] = \Delta_{\text{confounded}}.$$

Second, consider

$$E[\mathrm{D}^*{}_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)}] = E\Big[\Big(\sum_A \boldsymbol{R}^*(1{-}A, X) f^*(A \mid Z, X)\Big)\Big(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid A, Z, X]\Big)\frac{2A-1}{f^*(A \mid Z, X)}\Big]$$
$$= E\Big[\Big(\sum_A \boldsymbol{R}^*(1{-}A, X) f^*(A \mid Z, X)\Big)\Big(E[\boldsymbol{\Gamma}_W \mid A, Z, X] - E^*[\boldsymbol{\Gamma}_W \mid A, Z, X]\Big)\frac{2A-1}{f^*(A \mid Z, X)}\Big]$$

When $f^*(A \mid Z, X) = f(A \mid Z, X)$, by Eq. (34) we have

$$E[\mathrm{D}^*_{\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z,X)}] = E\Big[E[\boldsymbol{R}^*(1{-}A, X) \mid Z, X]\Big(\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X) - \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X)\Big)\Big] = 0$$

because $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X) = \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X)$.

Third, we consider $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]$. As discussed above, when $\boldsymbol{\Pi}^*(Z \mid A, X) = \boldsymbol{\Pi}(Z \mid A, X)$, we have Eq. (52) hold. Note that when the model for $f(A \mid X)$ is correctly specified, i.e., $f^*(A \mid X) = f(A \mid X)$, in Appendix H we show that for any function $h(Y, W, A, Z, X)$,

$$E[h(Y, W, A, Z, X)\frac{f(1-A \mid X)}{f(A \mid X)}] = E[h(Y, W, Z, 1{-}A, X)].$$

Let

$$h(Y, W, A, Z, X) = [\boldsymbol{R}(A, X) - \boldsymbol{R}^*(A, X)]\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A, X)\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}^*}(A, X)^{-1}\Big(\sum_Z \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X) f(Z \mid 1 - A, X)\Big),$$

then Eq. (52) is equivalent to

$$E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]$$
$$= E\Big\{[\boldsymbol{R}(1{-}A, X) - \boldsymbol{R}^*(1{-}A, X)]\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(1{-}A, X)\Big(\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}^*}(1{-}A, X)\Big)^{-1}\Big(\sum_Z \boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X) f(Z \mid 1 - (1{-}A), X)\Big)\Big\}.$$

In this case, because we also have that $\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)$ and $\boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}^*}(A, X) = \boldsymbol{\xi}_{\boldsymbol{Z}}^{\boldsymbol{W}}(A, X)$

$$E[\mathrm{D}^*_{R(1-A,X)} + \boldsymbol{R}^*(1{-}A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}^*}(Z, X)]$$
$$= E\Big\{[\boldsymbol{R}(1{-}A, X) - \boldsymbol{R}^*(1{-}A, X)]E[\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X) \mid A, X] + \boldsymbol{R}^*(1{-}A, X)\boldsymbol{\delta}_{\boldsymbol{A}}^{\boldsymbol{W}}(Z, X)\Big\} = \Delta_{\text{bias}}.$$

In summary, under $\mathcal{M}_2$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}] =& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}] + E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)} + \boldsymbol{R}^*(1-A,X)\boldsymbol{\delta}^{\boldsymbol{W}*}_{\boldsymbol{A}}(Z,X)]\} \\
=& \Delta_{\text{confounded}} - \{0 + \Delta_{\text{bias}}\} = \Delta
\end{aligned}
$$

Under $\mathcal{M}_3$ where $\boldsymbol{R}(A,X;\beta^R)$, $E[Y \mid Z = z_0, A, X; \beta^Y]$, $\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X;\beta^{WZ})$, $\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X;\beta^{WA})$, and $E[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X; \beta^W]$ are correctly specified, we have $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$, $E^*[Y \mid Z = z_0, A, X] = E[Y \mid Z = z_0, A, X]$, $\boldsymbol{\xi}^{\boldsymbol{W}*}_{\boldsymbol{Z}}(A,X) = \boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)$, $\boldsymbol{\delta}^{\boldsymbol{W}*}_{\boldsymbol{A}}(Z,X) = \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)$, and $E^*[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X] = E[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X]$. First we consider

$$
E[\Delta^*_{\text{confounded}}] = E\Big[\frac{2A-1}{f^*(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E^*[Y \mid A,Z,X]\Big) + E^*[Y \mid A{=}1,Z,X] - E^*[Y \mid A{=}0,Z,X]\Big].
$$

Note that

$$
\begin{aligned}
E^*[Y \mid Z,A,X] =& E^*[Y \mid Z = z_0, A, X] + \boldsymbol{R}^*(A,X)\boldsymbol{\xi}^{\boldsymbol{W}*}_{\boldsymbol{Z}}(A=0,X)\boldsymbol{\Gamma}_Z \\
=& E[Y \mid Z = z_0, A, X] + \boldsymbol{R}(A,X)\boldsymbol{\xi}^{\boldsymbol{W}}_{\boldsymbol{Z}}(A,X)\boldsymbol{\Gamma}_Z = E[Y \mid Z,A,X],
\end{aligned}
$$

therefore we have

$$
\begin{aligned}
E[\Delta^*_{\text{confounded}}] =& E\Big[\frac{2A-1}{f(A \mid Z,X)}\Big(E[Y \mid A,Z,X] - E[Y \mid A,Z,X]\Big) + E[Y \mid A{=}1,Z,X] - E[Y \mid A{=}0,Z,X]\Big] \\
=& E\{E[Y \mid A{=}1,Z,X] - E[Y \mid A{=}0,Z,X]\} = \Delta_{\text{confounded}}.
\end{aligned}
$$

Second, consider

$$
\begin{aligned}
E[\mathrm{D}^*{}_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}] =& E\Big[\Big(\sum_A \boldsymbol{R}^*(1-A,X)f^*(A \mid Z,X)\Big)\Big(\boldsymbol{\Gamma}_W - E^*[\boldsymbol{\Gamma}_W \mid A,Z,X]\Big)\frac{2A-1}{f^*(A \mid Z,X)}\Big] \\
=& E\Big[\Big(\sum_A \boldsymbol{R}^*(1-A,X)f^*(A \mid Z,X)\Big)\Big(E[\boldsymbol{\Gamma}_W \mid A,Z,X] - E^*[\boldsymbol{\Gamma}_W \mid A,Z,X]\Big)\frac{2A-1}{f^*(A \mid Z,X)}\Big] \\
=& 0
\end{aligned}
$$

because $E^*[\boldsymbol{\Gamma}_W \mid A,Z,X] = E[\boldsymbol{\Gamma}_W \mid A,Z,X]$ by (48).

Third, consider we consider $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}]$. Because $\boldsymbol{R}^*(A,X) = \boldsymbol{R}(A,X)$, $E^*[Y \mid Z = z_0, A, X] = E[Y \mid Z = z_0, A, X]$, $\boldsymbol{\delta}^{\boldsymbol{W}*}_{\boldsymbol{A}}(Z = z_0, X) = \boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z = z_0, X)$, and $E^*[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X] = E[\boldsymbol{\Gamma}_W \mid A = 0, Z = z_0, X]$, it is straightforward to see from Eq. (49) that $E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] = 0$.

In summary, under $\mathcal{M}_3$, we have

$$
\begin{aligned}
E[\Delta^*_{\text{mr}}] =& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}] + E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] + E[\boldsymbol{R}^*(1-A,X)\boldsymbol{\delta}^{\boldsymbol{W}^*}_{\boldsymbol{A}}(Z,X)]\} \\
=& E[\Delta^*_{\text{confounded}}] - \{E[\mathrm{D}^*_{\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)}] + E[\mathrm{D}^*_{\boldsymbol{R}(1-A,X)}] + E[\boldsymbol{R}(1-A,X)\boldsymbol{\delta}^{\boldsymbol{W}}_{\boldsymbol{A}}(Z,X)]\} \\
=& \Delta_{\text{confounded}} - \{0 + 0 + \Delta_{\text{bias}}\} = \Delta
\end{aligned}
$$

In summary, $E[\Delta^*_{\text{mr}}] = \Delta$ under $\mathcal{M}_{\text{union}} = \mathcal{M}_1 \cup \mathcal{M}_2 \cup \mathcal{M}_3$.

The rest of the arguments are the same as Appendix E. $\square$

H Change from A to $1-A$

In this section, we show that

$$E[h(Y,W,A,Z,X)\frac{f(1-A\mid X)}{f(A\mid X)}]=E[h(Y,W,Z,1-A,X)].$$

Proof. Consider

$$\begin{aligned}E[h(Y,W,A,Z,X)\frac{f(1-A\mid X)}{f(A\mid X)}]&=\int E[h(Y,W,A,Z,X)\mid A=a,X=x]\frac{f(1-a\mid x)}{f(a\mid x)}f(a,x)dadx\\&=\int E[h(Y,W,A,Z,X)\mid A=a,X=x]f(1-A,X)dadx.\end{aligned}$$

Because A is binary, we have

$$\begin{aligned}&\int E[h(Y,W,A,Z,X)\mid A=a,X=x]f(1-A,X)dadx\\&=\int E[h(Y,W,A,Z,X)\mid A=1,X=x]P(A=0,x)+E[h(Y,W,A,Z,X)\mid A=0,X=x]P(A=1,x)dx\\&=\int E[h(Y,W,Z,1-A)\mid 1-A=0,X=x]P(A=0,x)+E[h(Y,W,Z,1-A)\mid 1-A=1,X=x]P(A=1,x)dx\\&=\int E[h(Y,W,Z,1-A)\mid 1-A=a,X=x]f(a,x)dadx\\&=E[h(Y,W,Z,1-A,X)]\end{aligned}$$

Therefore

$$E[h(Y,W,A,Z,X)\frac{f(1-A\mid X)}{f(A\mid X)}]=E[h(Y,W,Z,1-A,X)].$$

□

I Sensitivity analysis for observational postlicensure vaccine safety study

The causal model under consideration in the Pentacel vaccine study is Figure 2c of the main text, where both injury/trauma and ringworm in fact satisfy the negative control exposure/outcome assumptions. In this scenario, the roles of W and Z can be switched, and the temporal order of Z and W relative to $A \to Y$ does not matter because they are independent of A and Y. In general, if one were to observe multiple negative control outcomes that are conditionally independent of A and Y, one could in principle split them and use some as negative control exposure and the others as negative control outcome. The vaccine application illustrates such a setting where the study monitored both ringworm and injury/trauma as control outcomes, and we took ringworm as negative control exposure. Analogous approach could be taken if one observes multiple negative control exposures.

We take injury/trauma as negative control outcome mainly because it is consistent with the literature. In the literature of vaccine effectiveness and safety, there is a long-standing history to use injury/trauma as a so-called "control outcome" or "falsification outcome", which is essentially an negative control outcome. However, most studies do not collect negative control exposures. Therefore these traditional studies can only detect but not remove unmeasured confounding bias. In the Pentacel vaccine study, in addition to injury/trauma, we have available data on ringworm as another "control outcome". Therefore we take ringworm as an negative control exposure to further remove confounding bias.

In this section, we perform sensitivity analysis by switching the roles of injury/trauma and ringworm. As presented in Table A, we observed similar conclusions that the difference in risk of fever comparing children who received DTaP-IPV-Hib vaccine relative with children who received other DTap-containing comparator vaccines is not statistically significant, although most point estimates are negative with wider confidence intervals compared to the primary analysis in Table 2. In addition, there was no evidence of unmeasured confounding. Under deviation from the nonparametric model, all methods produced a stable estimate of $\Delta_{\text{confounded}}$, while Δ_{bias} was estimated with larger bias. In general, multiply robust estimation provided protection against model misspecification.

It is also important to note that Figure 2c is consistent with Figure (b) of Kuroki and Pearl (2014), under which Kuroki and Pearl (2014) developed methods to identify the mean of potential outcomes. The first method relies on eigenvalue decomposition to recover $P(W \mid U, X)$. Due to the rare adverse events in our example, such a factor analysis approach may not be stable. Alternatively, further assuming linear structural equation models with normal random noises for relationships among all variables, equation (6) of Kuroki and Pearl (2014) can be used to identify the causal effect, which is essentially an analysis of variance approach. Our nonparametric identification is a generalization of (6) in the sense that we do not require linear structural equation models with normal random noises.

Table A: Adverse effect of DTaP-IPV-Hib vaccine on fever among children.

Scenario	**Method**	**ATE Δ (95\% CI)**	**Proportion Bias (\%)**	p**-val**	$\Delta_{\text{confounded}}$ **(95\% CI)**	Δ_{bias} **(95\% CI)**
	Δ_1	-5.0 (-18.5, 8.6)	-0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
All	Δ_2	-5.0 (-18.5, 8.6)	-0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
models are	Δ_3	-5.0 (-18.5, 8.6)	0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
NP	MLE	-5.0 (-18.5, 8.6)	-0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
	MR	-5.0 (-18.5, 8.6)	-0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
$\mathcal{M}_1$ **is NP**	Δ_2	-3.8 (-15.0, 7.4)	-23.3	0.5	0.5 (-0.5, 1.4)	4.3 (-6.8, 15.4)
$\mathcal{M}_2, \mathcal{M}_3$	Δ_3	0.5 (-2.4, 3.4)	-110.8	0.7	0.5 (-0.4, 1.4)	-0.0 (-2.8, 2.7)
are restricted	MLE	-4.7 (-17.6, 8.2)	-5.7	0.5	0.5 (-0.5, 1.4)	5.2 (-7.6, 17.9)
	MR	-5.0 (-19.1, 9.0)	1.4	0.5	0.5 (-0.5, 1.4)	5.5 (-8.5, 19.5)
$\mathcal{M}_2$ **is NP**	Δ_1	-3.8 (-17.2, 9.6)	-23.2	0.6	0.5 (-0.5, 1.4)	4.3 (-9.1, 17.6)
$\mathcal{M}_1, \mathcal{M}_3$	Δ_3	-3.8 (-17.2, 9.6)	-23.2	0.6	0.5 (-0.4, 1.4)	4.3 (-9.1, 17.6)
are restricted	MR	-7.7 (-33.7, 18.2)	56.1	0.6	0.5 (-0.5, 1.4)	8.2 (-17.6, 34.1)
$\mathcal{M}_3$ **is NP**	Δ_1	-5.1 (-18.9, 8.7)	2.5	0.5	0.5 (-0.5, 1.4)	5.6 (-8.1, 19.2)
$\mathcal{M}_1, \mathcal{M}_2$	Δ_2	-5.1 (-19.5, 9.4)	1.9	0.5	0.5 (-0.5, 1.4)	5.5 (-8.8, 19.9)
are restricted	MR	-5.0 (-18.5, 8.6)	-0.0	0.5	0.5 (-0.5, 1.4)	5.4 (-8.0, 18.8)
	Δ_1	-3.8 (-17.2, 9.6)	-23.2	0.6	0.5 (-0.5, 1.4)	4.3 (-9.1, 17.6)
All	Δ_2	-3.8 (-15.0, 7.4)	-23.3	0.5	0.5 (-0.5, 1.4)	4.3 (-6.8, 15.4)
models are	Δ_3	-2.4 (-16.7, 11.8)	-50.6	0.7	0.5 (-0.5, 1.4)	2.9 (-11.2, 17.1)
restricted	MLE	-4.7 (-17.6, 8.2)	-5.7	0.5	0.5 (-0.5, 1.4)	5.2 (-7.6, 17.9)
	MR	4.3 (-20.4, 29.0)	-186.7	0.7	0.5 (-0.5, 1.4)	-3.8 (-28.5, 20.9)

Note: all point estimates and 95% confidence intervals (CI) are scaled by 10^3. Prop bias (%) is the bias calculated as the proportion of the ATE under the saturated model (NP model) taken as the true value.

References

Gantmakher, F. R. (2000). *The theory of matrices*, volume 131. American Mathematical Soc.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Manski, C. F. (1988). *Analog Estimation Methods in Econometrics: Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Chapman and Hall.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Robins, J. M. and Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer" by P.J. Bickel and J. Kwon. *Statistica Sinica*, 11:920–936.

Wang, L. and Tchetgen Tchetgen, E. (2018). Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):531–550.