Estimating individual treatment effects under unobserved confounding using binary instruments Appendix

Anonymous Author(s)
Affiliation
Address
email

Contents

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

A Proofs

We start by deriving an auxiliary Lemma. That is, we derive an explicit expression for the Stage 2 oracle pseudo outcome regression $\mathbb{E}[\hat{Y}_0 \mid X = x]$ of MRIV.

Lemma 4.

$$
\begin{aligned}
&\mathbb{E}[\hat{Y}_0 \mid X = x] \\
=&\frac{\pi(x)}{\hat{\delta}_A(x)\hat{\pi}(x)} \left(\mu_1^Y(x) - \mu_1^A(x)\,\hat{\tau}_{\text{init}}(x)\right) + \frac{(1-\pi(x))}{\hat{\delta}_A(x)(1-\hat{\pi}(x))} \left(\mu_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \mu_0^Y(x)\right) \\
&+ \frac{\hat{\mu}_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \hat{\mu}_0^Y(x)}{\hat{\delta}_A(x)} \left(\frac{\pi(x)}{\hat{\pi}(x)} - \frac{1-\pi(x)}{1-\hat{\pi}(x)}\right) + \hat{\tau}_{\text{init}}(x)
\end{aligned}
\tag{1}
$$

Proof.

$$
\mathbb{E}[\hat{Y}_0 \mid X = x] \tag{2}
$$

$$
\begin{aligned}
=&\pi(x)\mathbb{E}\left[\frac{Y - A\,\hat{\tau}_{\text{init}}(X) - \hat{\mu}_0^Y(X) + \hat{\mu}_0^A(X)\,\hat{\tau}_{\text{init}}(X)}{\hat{\delta}_A(X)\,\hat{\pi}(X)} \,\middle|\, X = x, Z = 1\right] \\
&+ (1-\pi(x))\mathbb{E}\left[\frac{Y - A\,\hat{\tau}_{\text{init}}(X) - \hat{\mu}_0^Y(X) + \hat{\mu}_0^A(X)\,\hat{\tau}_{\text{init}}(X)}{\hat{\delta}_A(X)\,(1-\hat{\pi}(X))} \,\middle|\, X = x, Z = 0\right] + \hat{\tau}_{\text{init}}(x)
\end{aligned}
\tag{3}
$$

$$
\begin{aligned}
=&\frac{\pi(x)}{\hat{\delta}_A(x)\,\hat{\pi}(x)} \left(\mu_1^Y(x) - \mu_1^A(x)\,\hat{\tau}_{\text{init}}(x) - \hat{\mu}_0^Y(x) + \hat{\mu}_0^A(x)\,\hat{\tau}_{\text{init}}(x)\right) \\
&+ \frac{1-\pi(x)}{\hat{\delta}_A(x)\,(1-\hat{\pi}(x))} \left(\mu_0^Y(x) - \mu_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \hat{\mu}_0^Y(x) + \hat{\mu}_0^A(x)\,\hat{\tau}_{\text{init}}(x)\right) + \hat{\tau}_{\text{init}}(x)
\end{aligned}
\tag{4}
$$

Rearranging the terms yields the desired result. □

A.1 Proof of Theorem 1 (multiple robustness property)

We use Lemma 4 to show that under each of the three conditions it follows that $\mathbb{E}[\hat{Y}_0 \mid X = x] = \tau(x)$.

1.

$$
\mathbb{E}[\hat{Y}_0 \mid X = x] \tag{5}
$$

$$
\begin{aligned}
=&\frac{\pi(x)}{\delta_A(x)\,\hat{\pi}(x)} \left(\mu_1^Y(x) - \mu_1^A(x)\,\tau(x) + \mu_0^A(x)\,\tau(x) - \mu_0^Y(x)\right) \\
&+ \frac{(1-\pi(x))}{\delta_A(x)\,(1-\hat{\pi}(x))} \left(\mu_0^A(x)\,\tau(x) - \mu_0^Y(x) - \mu_0^A(x)\,\tau(x) + \mu_0^Y(x)\right) + \tau(x)
\end{aligned}
\tag{6}
$$

$$
=\frac{\pi(x)}{\delta_A(x)\,\hat{\pi}(x)} \left(\delta_Y(x) - \delta_Y(x)\right) + \tau(x) = \tau(x). \tag{7}
$$

2.

$$
\mathbb{E}[\hat{Y}_0 \mid X = x] = \frac{\left(\mu_1^Y(x) - \mu_1^A(x)\,\hat{\tau}_{\text{init}}(x)\right)}{\delta_A(x)} + \frac{\left(\mu_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \mu_0^Y(x)\right)}{\delta_A(x)} + \hat{\tau}_{\text{init}}(x) \tag{8}
$$

$$
= \frac{\delta_Y(x) - \hat{\tau}_{\text{init}}(x)\,\delta_A(x)}{\delta_A(x)} + \hat{\tau}_{\text{init}}(x) = \tau(x). \tag{9}
$$

3.

$$
\mathbb{E}[\hat{Y}_0 \mid X = x] = \frac{\left(\mu_1^Y(x) - \mu_1^A(x)\,\tau(x)\right)}{\hat{\delta}_A(x)} + \frac{\left(\mu_0^A(x)\,\tau(x) - \mu_0^Y(x)\right)}{\hat{\delta}_A(x)} + \tau(x) \tag{10}
$$

$$
= \frac{\delta_Y(x)}{\hat{\delta}_A(x)} - \tau(x)\frac{\delta_A(x)}{\hat{\delta}_A(x)} + \tau(x) = \tau(x) \tag{11}
$$

A.2 Proof of Theorem 2 (Convergence rate of MRIV)

To prove Theorem 2, we need an additional assumption on the second stage regression estimator $\hat{\mathbb{E}}_n$. We refer to Kennedy [8] (Theorem 1) for a detailed discussion on this assumption.

Assumption 5 (From Theorem 1 of Kennedy [8])**.** The following two statements hold:

1. $\hat{\mathbb{E}}_n[W + c \mid X = x] = \hat{\mathbb{E}}_n[W \mid X = x] + c$ for any random W and constant c
2. If $\mathbb{E}[W \mid X = x]$ = $E[V \mid X = x]$ then

$$\mathbb{E}\left[\left(\hat{\mathbb{E}}_n[W \mid X = x] - \mathbb{E}[W \mid X = x]\right)^2\right] \asymp \mathbb{E}\left[\left(\hat{\mathbb{E}}_n[V \mid X = x] - \mathbb{E}[V \mid X = x]\right)^2\right]. \tag{12}$$

Proof of Theorem 2. Using Assumption 5, we can apply Theorem 1 of Kennedy [8] and obtain

$$\mathbb{E}\left[\left(\hat{\tau}_{\text{init}}(x) - \tau(x)\right)^2\right] \lesssim \mathcal{R}(x) + \mathbb{E}\left[\hat{r}(x)^2\right], \tag{13}$$

where $\mathcal{R}(x) = \mathbb{E}\left[\left(\widetilde{\tau}_{MR}(x) - \tau(x)\right)^2\right]$ is the oracle risk of the second stage regression and $r(x) = \mathbb{E}[\hat{Y}_0 \mid X = x] - \tau(x)$. We can apply Lemma 4 to obtain

$$\begin{aligned}
\hat{r}(x) &= \frac{\pi(x)}{\hat{\delta}_A(x)\,\hat{\pi}(x)}\left(\mu_1^Y(x) - \mu_1^A(x)\,\hat{\tau}_{\text{init}}(x)\right) + \frac{(1-\pi(x))}{\hat{\delta}_A(x)\,(1-\hat{\pi}(x))}\left(\mu_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \mu_0^Y(x)\right) \\
&\quad + \frac{\hat{\mu}_0^A(x)\,\hat{\tau}_{\text{init}}(x) - \hat{\mu}_0^Y(x)}{\hat{\delta}_A(x)}\left(\frac{\pi(x)}{\hat{\pi}(x)} - \frac{1-\pi(x)}{1-\hat{\pi}(x)}\right) + \hat{\tau}_{\text{init}}(x) - \tau(x) \qquad (14) \\
&= \left(\frac{\mu_1^Y(x) - \mu_0^Y(x)}{\hat{\delta}_A(x)}\right)\frac{\pi(x)}{\hat{\pi}(x)} + \frac{\mu_0^Y(x) - \hat{\mu}_0^Y(x)}{\hat{\delta}_A(x)}\left(\frac{\pi(x)}{\hat{\pi}(x)} - \frac{1-\pi(x)}{1-\hat{\pi}(x)}\right) + \left(\hat{\tau}_{\text{init}}(x) - \tau(x)\right) \\
&\quad + \left(\frac{\left(\mu_0^A(x) - \mu_1^A(x)\right)\hat{\tau}_{\text{init}}(x)}{\hat{\delta}_A(x)}\right)\frac{\pi(x)}{\hat{\pi}(x)} + \frac{\left(\hat{\mu}_0^D(x) - \mu_0^D(x)\right)\hat{\tau}_{\text{init}}(x)}{\hat{\delta}_A(x)}\left(\frac{\pi(x)}{\hat{\pi}(x)} - \frac{1-\pi(x)}{1-\hat{\pi}(x)}\right) \\
&\qquad (15) \\
&= \frac{\delta_Y(x)\,\pi(x)}{\hat{\delta}_A(x)\,\hat{\pi}(x)} + \frac{\left(\mu_0^Y(x) - \hat{\mu}_0^Y(x)\right)\left(\pi(x) - \hat{\pi}(x)\right)}{\hat{\delta}_A(x)\,\hat{\pi}(x)\,(1-\hat{\pi}(x))} + \left(\hat{\tau}_{\text{init}}(x) - \tau(x)\right) \\
&\quad - \frac{\delta_A(x)\,\pi(x)\,\hat{\tau}_{\text{init}}(x)}{\hat{\delta}_A(x)\,\hat{\pi}(x)} + \frac{\left(\hat{\mu}_0^A(x) - \mu_0^A(x)\right)\hat{\tau}_{\text{init}}(x)\left(\pi(x) - \hat{\pi}(x)\right)}{\hat{\delta}_A(x)\,\hat{\pi}(x)\,(1-\hat{\pi}(x))} \qquad (16) \\
&= \frac{\left(\pi(x) - \hat{\pi}(x)\right)}{\hat{\delta}_A(x)\,\hat{\pi}(x)\,(1-\hat{\pi}(x))}\left[\left(\mu_0^Y(x) - \hat{\mu}_0^Y(x)\right) + \left(\hat{\mu}_0^A(x) - \mu_0^A(x)\right)\hat{\tau}_{\text{init}}(x)\right] \\
&\quad + \left(\hat{\tau}_{\text{init}}(x) - \tau(x)\right) + \frac{\pi(x)\delta_A(x)}{\hat{\pi}(x)\hat{\delta}_A(x)}\left(\tau(x) - \hat{\tau}_{\text{init}}(x)\right) \qquad (17) \\
&= \frac{\left(\pi(x) - \hat{\pi}(x)\right)}{\hat{\delta}_A(x)\,\hat{\pi}(x)\,(1-\hat{\pi}(x))}\left[\left(\mu_0^Y(x) - \hat{\mu}_0^Y(x)\right) + \left(\hat{\mu}_0^A(x) - \mu_0^A(x)\right)\hat{\tau}_{\text{init}}(x)\right] \\
&\quad + \left(\tau(x) - \hat{\tau}_{\text{init}}(x)\right)\left(\delta_A(x) - \hat{\delta}_A(x)\right)\pi(x) + \left(\tau(x) - \hat{\tau}_{\text{init}}(x)\right)\left(\pi(x) - \hat{\pi}(x)\right)\hat{\delta}_A(x). \\
&\qquad (18)
\end{aligned}$$

Applying the inequality $(a+b)^2 \leq 2(a^2+b^2)$ together with Assumption 4 and the fact that $\pi(x) \leq 1$ yields

$$\begin{aligned}
\hat{r}(x)^2 &\leq \frac{4}{\epsilon^4\rho^2}\left(\pi(x) - \hat{\pi}(x)\right)^2\left[\left(\mu_0^Y(x) - \hat{\mu}_0^Y(x)\right)^2 + \left(\hat{\mu}_0^A(x) - \mu_0^A(x)\right)^2 K^2\right] \\
&\quad + 4\left(\tau(x) - \hat{\tau}_{\text{init}}(x)\right)^2\left(\delta_A(x) - \hat{\delta}_A(x)\right)^2 + 4\left(\tau(x) - \hat{\tau}_{\text{init}}(x)\right)^2\left(\pi(x) - \hat{\pi}(x)\right)^2. \qquad (19)
\end{aligned}$$

By setting $\widetilde{K} = \max\{K, 1\}$, we obtain

$$\hat{r}(x)^2 \leq \frac{4\widetilde{K}^2}{\epsilon^4 \rho^2} \Big((\pi(x) - \hat{\pi}(x))^2 \Big[\big(\mu_0^Y(x) - \hat{\mu}_0^Y(x)\big)^2 + \big(\hat{\mu}_0^A(x) - \mu_0^A(x)\big)^2 + (\hat{\tau}_{\text{init}}(x) - \tau(x))^2 \Big] \\ + \, (\tau(x) - \hat{\tau}_{\text{init}}(x))^2 \Big(\delta_A(x) - \hat{\delta}_A(x)\Big)^2 \Big). \tag{20}$$

Applying expectations on both sides yields

$$\mathbb{E}\left[(\hat{\tau}_{\text{init}}(x) - \tau(x))^2\right] \tag{21}$$

$$\lesssim \mathcal{R}(x) + \mathbb{E}\left[(\hat{\tau}_{\text{init}}(x) - \tau(x))^2\right] \left(\mathbb{E}\left[\Big(\hat{\delta}_A(x) - \delta_A(x)\Big)^2\right] + \mathbb{E}\left[(\hat{\pi}(x) - \pi(x))^2\right] \right) \\ + \mathbb{E}\left[(\hat{\pi}(x) - \pi(x))^2\right] \left(\mathbb{E}\left[\big(\hat{\mu}_0^Y(x) - \mu_0^Y(x)\big)^2\right] + \mathbb{E}\left[\big(\hat{\mu}_0^A(x) - \mu_0^A(x)\big)^2\right] \right), \tag{22}$$

because $(\hat{\pi}(x), \hat{\delta}_A(x)) \perp\!\!\!\perp (\hat{\mu}_0^Y(x), \hat{\mu}_0^A(x), \hat{\tau}_{\text{init}}(x))$ due to sample splitting. The claim follows now by applying Assumption 3. □

A.3 Proof of Theorem 3 (Convergence rate of the Wald estimator)

Proof. We define $\widetilde{C} = \max\{C, 1\}$ and obtain the upper bound

$$(\hat{\tau}_W(x) - \tau(x))^2 \tag{23}$$

$$= \left(\frac{(\hat{\mu}_1^Y(x) - \mu_1^Y(x))\,\delta_A(x) + (\mu_0^Y(x) - \hat{\mu}_0^Y(x))\,\delta_A(x) + (\delta_A(x) - \hat{\delta}_A(x))\,\delta_Y(x)}{\delta_A(x)\,\hat{\delta}_A(x)} \right)^2 \tag{24}$$

$$\leq \frac{4\widetilde{C}^2}{\rho^2 \widetilde{\rho}^2} \left[(\hat{\mu}_1^Y(x) - \mu_1^Y(x))^2 + (\hat{\mu}_0^Y(x) - \mu_0^Y(x))^2 + (\delta_A(x) - \hat{\delta}_A(x))^2 \right] \tag{25}$$

$$\leq \frac{8\widetilde{C}^2}{\rho^2 \widetilde{\rho}^2} \left[(\hat{\mu}_1^Y(x) - \mu_1^Y(x))^2 + (\hat{\mu}_0^Y(x) - \mu_0^Y(x))^2 + (\hat{\mu}_1^A(x) - \mu_1^A(x))^2 \right. \\ \left. + (\hat{\mu}_0^A(x) - \mu_0^A(x))^2 \right], \tag{26}$$

where we used the inequality $(a + b)^2 \leq 2(a^2 + b^2)$ several times. Taking expectations and applying the smoothness assumptions yields the result. □

B Theoretical analysis under sparsity assumptions

In Sec. 4.2, we analyzed MRIV theoretically by imposing smoothness assumptions on the underlying data generating process. In particular, we derived a multiple robust convergence rate and showed that MRIV outperforms the Wald estimator if the oracle ITE is smoother than its components. In this section, we derive similar results by relying on a different set of assumptions. Instead of using smoothness, we make assumptions on the level of sparsity of the ITE components. This assumption is often imposed in high-dimensional settings ($n < p$) and is in line with previous literature on analyzing ITE estimators [4, 8].

In the following, we say a function $f(x)$ is k-sparse, if it is linear in $x \in \mathbb{R}^p$ and it only depends on $k < \min\{n, p\}$ predictors. [22] showed, that in this case the minimax rate of $f(x)$ is given by $\frac{k \log(p)}{n}$. The linearity assumption can be relaxed to an additive structural assumption, which we omit here for simplicity. In the following, we replace the smoothness conditions in Assumption 3 with sparsity conditions.

Assumption 6 (Sparsity)**.** We assume that (1) the nuisance components $\mu_i^Y(\cdot)$ are α-sparse, $\mu_i^A(\cdot)$ and $\delta_A(\cdot)$ are β-sparse, and $\pi(\cdot)$ is δ-sparse; (2) all nuisance components are estimated with their respective minimax rate of $\frac{k \log(p)}{n}$, where $k \in \{\alpha, \beta, \delta\}$; and (3) the oracle ITE $\tau(\cdot)$ is γ-sparse and the initial ITE estimator $\hat{\tau}_{\text{init}}$ converges with rate $r_\tau(n)$.

We restate now our result from Theorem 3 for MRIV using the sparsity assumption.

Theorem 5 (MRIV upper bound under sparsity)**.** *We consider the same setting as in Theorem 2 under the sparsity assumption 6. If the second-stage estimator $\hat{\mathbb{E}}_n$ yields the minimax rate $\frac{\gamma \log(p)}{n}$ and satisfies Assumption 5, the oracle risk is upper bounded by*

$$\mathbb{E}\left[(\hat{\tau}_{\text{MRIV}}(x) - \tau(x))^2\right] \lesssim \frac{\gamma \log(p)}{n} + r_\tau(n) \frac{(\beta + \delta) \log(p)}{n} + \frac{(\alpha + \beta)\delta \log^2(p)}{n^2}.$$

Proof. Follows immediately from the proof of Theorem 2, i.e., from Eq.(21) by applying Ass- 6. □

Again, we obtain a multiple robust convergence rate for MRIV in the sense that MRIV achieves a fast rate even if the initial estimator or several nuisance estimators converge slowly. More precisely, for a fast convergence rate of $\hat{\tau}_{\text{MRIV}}(x)$, it is sufficient if either: (1) $r_\tau(n)$ decreases fast and δ is small; (2) $r_\tau(n)$ decreases fast and α and β are small; or (3) all α, β, and δ are small.

We derive now the corresponding rate for the Wald estimator.

Theorem 6 (Wald oracle upper bound)**.** *Given estimators $\hat{\mu}_i^Y(x)$ and $\hat{\mu}_i^A(x)$. Let $\hat{\delta}_A(x) = \hat{\mu}_1^A(x) - \hat{\mu}_0^A(x)$ satisfy Assumption 4. Then, under Assumption 6 the oracle risk of the Wald estimator $\hat{\tau}_W(x)$ is bounded by*

$$\mathbb{E}\left[(\hat{\tau}_{\text{W}}(x) - \tau(x))^2\right] \lesssim \frac{(\alpha + \beta) \log(p)}{n} \tag{27}$$

Proof. Follows immediately from the proof of Theorem 3, i.e., from Eq.(23) by applying Ass- 6. □

If $\alpha = \beta = \delta$, we obtain the rates

$$\mathbb{E}\left[(\hat{\tau}_{\text{MRIV}}(x) - \tau(x))^2\right] \lesssim \frac{\gamma \log(p)}{n} + \frac{\alpha^2 \log^2(p)}{n^2} \quad \text{and} \quad \mathbb{E}\left[(\hat{\tau}_{\text{W}}(x) - \tau(x))^2\right] \lesssim \frac{\alpha \log(p)}{n}, \tag{28}$$

which means that $\hat{\tau}_{\text{MRIV}}(x)$ outperforms $\hat{\tau}_{\text{W}}(x)$ for $\gamma < \alpha$, i.e., if the oracle ITE is more sparse than its components.

C Simulated data

In the following, we describe how we simulate synthetic data for the experiments in Sec. 5.1 from the main paper. As mentioned therein, we simulate the ITE components from Gaussian processes using the prior induced by the Matern kernel [12]

$$K_{\ell,\nu}(x_i,x_j) = \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{\ell}\|x_i-x_j\|_2\right)^{\nu} K_{\nu}\left(\frac{\sqrt{2\nu}}{\ell}\|x_i-x_j\|_2\right), \tag{29}$$

where $\Gamma(\cdot)$ is the Gamma function and $K_\nu(\cdot)$ is the modified Bessel function of second kind. Here, ℓ is the length scale of the kernel and ν controls the smoothness of the sampled functions.

We set $\ell = 1$ and sample functions $\delta_Y \sim \mathcal{GP}(0, K_{\ell,\gamma})$, $\mu_0^Y \sim \mathcal{GP}(0, K_{\ell,\alpha})$, $f_1 \sim \mathcal{GP}(0, K_{\ell,\beta})$, $f_0 \sim \mathcal{GP}(0, K_{\ell,\beta})$ and $g \sim \mathcal{GP}(0, K_{\ell,\beta})$. Then, we define $\mu_1^Y = \delta_Y + \mu_0^Y$, $\mu_1^A = 0.3 \cdot \sigma \circ f_1 + 0.7$, $\mu_0^A = 0.3 \cdot \sigma \circ f_0$, $\delta_A = \mu_1^A - \mu_0^A$, $\mu_0^Y = c_0 \delta_A$, and $\pi = \sigma \circ g$. Finally, we set the oracle ITE to

$$\tau = \frac{\mu_1^Y - \mu_0^Y}{\mu_1^A - \mu_0^A} = \frac{\delta_Y}{\delta_A}. \tag{30}$$

Note that we can create a setup where the ITE τ is smoother than its components by using a small α/β ratio. An example is shown in Fig. 1.

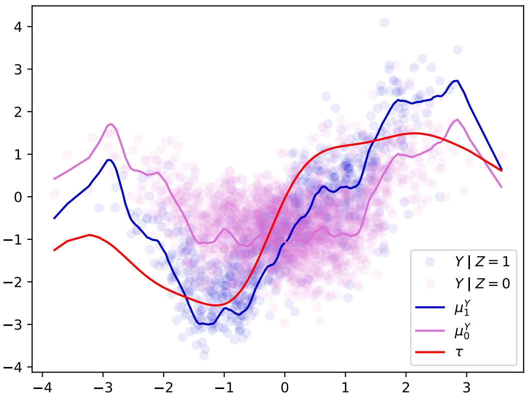

Figure 1: Gaussian process simulation for $\alpha = 1.5$ and $\beta = 50$.

In the following, we describe how we generate data the (X, Z, A, Y) using the ITE components $\mu_i^Y(x)$, $\mu_i^A(x)$, and $\pi(x)$. We begin by sampling n observed confounder $X \sim \mathcal{N}(0,1)$, unobserved confounders $U \sim \mathcal{N}\left(0, 0.2^2\right)$, and instruments $Z \sim \text{Bernoulli}(\pi(X))$. Then, we obtain treatments via

$$A = Z\,\mathbb{1}\{U + \epsilon_A > \alpha_1(X)\} + (1-Z)\,\mathbb{1}\{U + \epsilon_A > \alpha_0(X)\} \tag{31}$$

with indicator function $\mathbb{1}$, noise $\epsilon_A \sim \mathcal{N}\left(0, 0.1^2\right)$, and $\alpha_i(X) = \Phi^{-1}\left(1 - \mu_i^A(X)\right)\sqrt{0.1^2 + 0.2^2}$, where Φ^{-1} denotes the quantile function of the standard normal distribution. Finally, we generate the outcomes via

$$\begin{aligned} Y = {} & A\left(\frac{(\mu_1^A(X) - 1)\mu_0^Y(X) - \mu_0^A(X)\mu_1^Y(X) + \mu_1^Y(X)}{\delta_A(X)}\right) & (32) \\ & + (1-A)\left(\frac{\mu_1^A(X)\mu_0^Y(X) - \mu_0^A(X)\mu_1^Y(X)}{\delta_A(X)}\right) + \alpha_U U + \epsilon_Y, & (33) \end{aligned}$$

where $\epsilon_Y \sim \mathcal{N}\left(0, 0.3^2\right)$ is noise and $\alpha_U > 0$ is a parameter indicating the level of unobserved confounding. This choice of A and Y in Eq. (31) and Eq. (32), respectively, implies that $\tau(x)$ is indeed the ITE, i. e., it holds that $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$.

Lemma 7. *Let (X, Z, A, Y) be sampled from the the previously described procedure. Then, it holds that*

$$\mu_i^A(x) = \mathbb{E}[A \mid Z = i, X = x] \quad \textit{and} \quad \mu_i^Y(x) = \mathbb{E}[Y \mid Z = i, X = x]. \tag{34}$$

Proof. The first claim follows from

$$\mathbb{E}[A \mid Z = i, X = x] = \mathbb{P}\left(U + \epsilon_A > \alpha_i(x)\right) = 1 - \Phi(\Phi^{-1}(1 - \mu_i^A(x))) = \mu_i^A(x), \tag{35}$$

because $U + \epsilon_A \sim \mathcal{N}(0, \sqrt{0.1^2 + 0.2^2})$. The second claim follows from

$$\begin{aligned}\mathbb{E}[Y \mid Z = i, X = x] &= \mu_i^A(x) \left(\frac{(\mu_1^A(x) - 1)\mu_0^Y(x) - \mu_0^A(x)\mu_1^Y(x) + \mu_1^Y(x)}{\delta_A(x)} \right) &(36)\\ &\quad + (1 - \mu_i^A(x)) \left(\frac{\mu_1^A(x)\mu_0^Y(x) - \mu_0^A(x)\mu_1^Y(x)}{\delta_A(x)} \right) &(37)\\ &= \frac{\mu_i^Y(x)\delta_A(x)}{\delta_A(x)} = \mu_i^Y(x). &(38)\end{aligned}$$

□

D Oregon health insurance experiment

The so-called *Oregon health insurance experiment*[1] (OHIE) [6] was an important RCT with non-compliance. It was intentionally conducted as large-scale effort among public health to assess the effect of health insurance on several outcomes such as health or economic status. In 2008, a lottery draw offered low-income, uninsured adults in Oregon participation in a Medicaid program, providing health insurance. Individuals whose names were drawn could decide to sign up for the program.

In our analysis, the lottery assignment is the instrument Z, the decision to sign up for the Medicaid program is the treatment A, and an overall health score is the outcome Y. The outcome was obtained after a period of 12 months during in-person interviews. We use the following covariates X: age, gender, language, the number of emergency visits before the experiment, and the number of people the individual signed up with. The latter is used to control for peer effects, and it is important to include this variable in our analysis as it is the only variable influencing the propensity score (see below). We extract $\sim$ 10,000 observations from the OHIE data and plot the histograms of all variables in Fig. 2. We can clearly observe the presence of non-compliance within the data, because the ratio of treated / untreated individuals is much lower than the corresponding ratio for the treatment assignment.

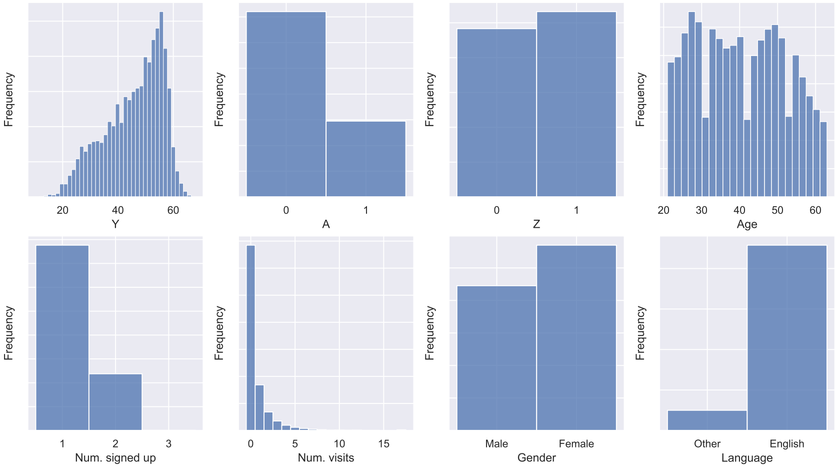

Figure 2: Histograms of each variable in our sample from OHIE.

The data collection in the OHIE was done follows: After excluding individuals below the age of 19, above the age of 64, and individuals with residence outside of Oregon, 74,922 individuals were considered for the lottery. Among those, 29,834 were selected randomly and were offered participation in the program. However, the probability of selection depended on the number of household members on the waiting list: for instance, an individual who signed up with another person was twice as likely to be selected. From the 74,922 individuals, 57,528 signed up alone, 17,236 signed up with another person, and 158 signed up with two more people on the waiting list. Thus, the probability of being selected conditional on the number of household members on the waiting list follows the multivariate version of Wallenius' noncentral hypergeometric distribution [2].

Propensity score: We computed the propensity score as follows. To account for the Wallenius' noncentral hypergeometric distribution, we use the R package *BiasedUrn* to calculate the propensity score $\pi(x) = \mathbb{P}(Z = 1 \mid X = x)$. We obtained

$$\pi(x) = \begin{cases} 0.345, & \text{if individual } x \text{ signed up alone,} \\ 0.571, & \text{if individual } x \text{ signed up with one more person,} \\ 0.719, & \text{if individual } x \text{ signed up with two more people.} \end{cases} \tag{39}$$

During the training of both MRIV and DRIV, we use the calculated values from Eq. (39) for the propensity score.

[1] Data available here: https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment

E Details for baseline methods

In this section, we give a brief overview on the baselines which we used in our experiments. We implemented: (1) ITE methods for unconfoundedness [8, 13]; (2) general IV methods, i.e., IV methods developed for IV settings with multiple or continuous instruments and treatments [1, 7, 14, 15, 20, 21]; and (3) two instantiations of the Wald estimator for the binary IV setting [16].

E.1 ITE methods for unconfoundedness

Many ITE methods assume *unconfoundedness*, i.e., that all confounders are observed in the data. Formally, the unconfoundedness assumption can be expressed in the potential outcomes framework as

$$Y(1), Y(0) \perp\!\!\!\perp A \mid X. \tag{40}$$

Under unconfoundedness, the ITE is identified as

$$\tau(x) = \mu_1(x) - \mu_0(x) \quad \text{with} \quad \mu_i(x) = \mathbb{E}[Y \mid A = i, X = x]. \tag{41}$$

Methods that assume unconfoundedness proceed by estimating $\mu_i(x) = \mathbb{E}[Y \mid A = i, X = x]$ from Eq. (41). However, if unobserved confounders U exist, it follows that

$$\tau(x) = \mathbb{E}[Y \mid A = 1, X = x, U] - \mathbb{E}[Y \mid A = 0, X = x, U] \neq \mu_1(x) - \mu_0(x), \tag{42}$$

which means that estimators that assume unconfoundedness are generally biased. Nevertheless, we include two baselines that assume unconfoundedness into our experiments: TARNet [13] and the DR-learner [8].

TARNet [13]: TARNet [13] is a neural network that estimates the ITE components $\mu_i(x)$ from Eq. 41 by learning a shared representation $\Phi(x)$ and two potential outcome heads $h_i(\Phi(x))$. We train TARNet by minimizing the loss

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} L\left(h_{a_i}(\Phi(x_i, \theta_\Phi), \theta_{h_i}), y_i\right), \tag{43}$$

where $\theta = (\theta_{h_1}, \theta_{h_0}, \theta_\Phi)$ denotes the model parameters and L denotes squared loss if Y is continuous or binary cross entropy loss if Y is binary.

Note regarding balanced representations: In [13], the authors propose to add an additional regularization term inspired from domain adaptation literature, which forces TARNet to learn a balanced representation $\Phi(x)$, i.e., that minimizes the distance the treatment and control group in the feature space. They showed that this approach leads to minimization of a generalization bound on the ITE estimation error if the representation is invertible.

In our experiments, we refrained from learning balanced representations because minimizing the regularized loss from [13] does not necessarily result in an invertible representation and thus may even harm the estimation performance. For a detailed discussion, we refer to [4]. Furthermore, by leaving out the regularization, we ensure comparability between the different baselines. If balanced representations are desired, the balanced representation approach could also be extended to MRIV-Net, as we also build MRIV-Net on learning shared representations.

DR-learner [8]: The DR-learner [8] is a meta learner that takes arbitrary estimators of the ITE componenets μ_i and the propensity score $\pi(x) = \mathbb{P}(A = 1 \mid X = x)$ as input and performs a pseudo outcome regression by using the pseudo outcome

$$\hat{Y}_0 = \left(\frac{A}{\hat{\pi}(X)} - \frac{1-A}{1-\hat{\pi}(X)}\right) Y + \left(1 - \frac{A}{\hat{\pi}(X)}\right) \hat{\mu}_1(X) - \left(1 - \frac{1-A}{1-\hat{\pi}(X)}\right) \hat{\mu}_0(X). \tag{44}$$

In our experiments, we use TARNet as base method to provide initial estimators $\hat{\mu}_i(X)$. We further learn propensity score estimates $\hat{\pi}(X)$ by adding a seperate representation to TARNet as done in [13].

E.2 General IV methods

2SLS [20]: 2SLS [20] is a linear two-stage approach. First, the treatments A are regressed on the instruments Z and fitted values $\hat{A}$ are obtained. In the second stage, the outcome Y is regressed on $\hat{A}$. We implement 2SLS using the scikit-learn package.

KIV [14]: Kernel IV [14] generalizes 2SLS to nonlinear settings. KIV assumes that the data is generated by

$$Y = f(A) + U, \tag{45}$$

where U is an additive unobserved confounder and f is some unknown (potentially nonlinear) structural function. KIV then models the structural function via

$$f(a) = \mu^t \psi(a) \quad \text{and} \quad \mathbb{E}[\psi(A) \mid Z = z] = V\phi(z), \tag{46}$$

where ψ andϕ are feature maps. Here, kernel ridge regressions instead of linear regressions are used in both stages to estimate μ and V.

Following [14] we use the exponential kernel [12] and set the length scale to the median inter-point distance. KIV does not provide a direct way to incorporate the observed confounders X. Hence, we augment both the instrument and the treatment with X, which is consistent with previous work [1, 21]. We also use two different samples for each stage as recommended in [14].

DFIV [21]: DFIV [21] is a similar approach KIV in generalizing 2SLS to nonlinear setting by assuming Eq. (45) and Eq. (46). However, instead of using kernel methods, DFIV models the features maps ψ_{θ_A} and ϕ_{θ_Z} as neural networks with parameters θ_A and θ_Z, respectively. DFIV is trained by iteratively updating the parameters θ_A and θ_Z. The authors also provide a training algorithm that incorporates observed confounders X, which we implemented for our experiments. During training, we used two different datasets for each of the two stages as described in in the paper.

DeepIV [7]: DeepIV [7] also assumes additive unobserved confounding as in Eq. (45), but leverages the identification result [10]

$$\mathbb{E}[Y \mid X = x, Z = z] = \int h(a, x) \, \mathrm{d}F(a \mid x, z), \tag{47}$$

where $h(a, x) = f(a, x) + \mathbb{E}[U \mid X = x]$ is the target counterfactual prediction function. DeepIV estimates $F(a \mid x, z)$, i.e., the conditional distribution function of the treatment A given observed covariates X and instruments Z, by using neural networks. Because we consider only binary treatments, we simply implement a (tunable) feed-forward neural network with sigmoid activation function. Then, DeepIV proceeds by learning a second stage neural network to solve the inverse problem defined by Eq. (47).

DeepGMM [1]: DeepGMM [1] adopts neural networks for IV estimation inspired by the (optimally weighted) Generalized Method of Moments. The DeepGMM estimator is defined as the solution of the following minimax game:

$$\hat{\theta} \in \operatorname*{arg\,min}_{\theta \in \Theta} \sup_{\tau \in \mathrm{T}} \frac{1}{n} \sum_{i=1}^{n} f(z_i, \tau)(y_i - g(a_i, \theta)) - \frac{1}{4n} \sum_{i=1}^{n} f^2(z_i, \tau)(y_i - g(a_i, \widetilde{\theta}))^2, \tag{48}$$

where $f(z_i, \cdot)$ and $g(a_i, \cdot)$ are parameterized by neural networks. As recommended in [1], we solve this optimization via adversarial training with the Optimistic Adam optimizer [5], where we set the parameter $\widetilde{\theta}$ to the previous value of θ.

DMLIV [15]: DMLIV [15] assumes that the data is generated via

$$Y = \tau(X)A + f(X) + U, \tag{49}$$

where τ is the ITE f some function of the observed covariates. First, DMLIV estimates the functions $q(X) = \mathbb{E}[Y \mid X]$, $h(Z, X) = \mathbb{E}[A \mid Z, X]$, and $p(X) = \mathbb{E}[A \mid X]$. Then, the ITE is learned by minimizing the loss

$$\mathcal{L}(\theta) = \sum_{i=1} (y_i - \hat{q}(x_i) - \hat{\tau}(x_i, \theta)(\hat{h}(z_i, x_i) - \hat{p}(x_i))^2, \tag{50}$$

where $\hat{\tau}(X, \cdot)$ is some model for $\tau(X)$. In our experiments, we use (tunable) feed-forward neural networks for all estimators.

DRIV [15]: DRIV [15] is a meta learner, originally proposed in combination with DMLIV. It requires initial estimators for $q(X)$, $p(X)$, $\pi(X) = \mathbb{E}[Z \mid X = x]$, and $f(X) = \mathbb{E}[A \cdot Z \mid X = x]$ as well as an initial ITE estimatior $\hat{\tau}_{\text{init}}(X)$ (e.g., from DMLIV). The ITE is then estimated by a pseudo regression on the following doubly robust pseudo outcome:

$$\hat{Y}_{\text{DR}} = \hat{\tau}_{\text{init}}(X) + \frac{(Y - \hat{q}(X) - \hat{\tau}_{\text{init}}(X)(A - \hat{p}(X))Z - \hat{\pi}(X))}{\hat{f}(X) - \hat{p}(X)\hat{r}(X)}. \tag{51}$$

We implement all regressions using (tunable) feed-forward neural networks.

Comparison between DRIV vs. MRIV: There are two key differences between our paper and [15]: (i) Our MRIV is multiply robust, while DRIV is only doubly robust. (ii) We derive a multiple robust convergence rate, while the rate in [15] is not robust with respect to the nuisance rates.

Ad (i): Both MRIV and DRIV perform a pseudo-outcome regression on the efficient influence function (EIF) of the ATE. The key difference: DRIV uses the doubly robust parametrization of the EIF from [11], whereas we use the multiply robust parametrization of the EIF from [17] [2]. Hence, our MRIV frameworks extends DRIV in a non-trivial way to achieve multiple robustness (rather than doubly robustness). Thus, our estimator is consistent in the union of *three* different model specifications rather than *two*.[3]

Ad (ii): Here, we compare the convergence rates from DRIV and our MRIV and, thereby, show the strengths of our MRIV. To this end, let us assume that the pseudo regression function is γ-smooth and that we use the same second-stage estimator $\hat{\mathrm{E}}_n$ with minimax rate $n^{-\frac{2\gamma}{2\gamma+p}}$ for both DRIV and MRIV. If the nuisance parameters $q(X)$, $p(X)$, $f(X)$, and $\pi(X)$ are α-smooth and further are estimated with minimax rate $n^{\frac{-2\alpha}{2\alpha+p}}$, Corollary 4 from [15] states that DRIV converges with rate

$$\mathbb{E}\left[\left(\hat{\tau}_{\mathrm{DRIV}}(x) - \tau(x)\right)^2\right] \lesssim n^{\frac{-2\gamma}{2\gamma+p}} + n^{\frac{-4\alpha}{2\alpha+p}}.$$

In contrast, MRIV assumes estimation of the nuisance parameters $\mu_0^Y(x)$ with rate $n^{\frac{-2\alpha}{2\alpha+p}}$, $\mu_0^A(x)$ and $\delta_A(x)$ with rate $n^{\frac{-2\beta}{2\beta+p}}$, and $\pi(x)$ with rate $n^{\frac{-2\delta}{2\delta+p}}$. If the initial estimator $\hat{\tau}_{\mathrm{init}}(x)$ converges with rate $r_\tau(n)$, our Theorem 2 yields the rate

$$\mathbb{E}\left[\left(\hat{\tau}_{\mathrm{MRIV}}(x) - \tau(x)\right)^2\right] \lesssim n^{\frac{-2\gamma}{2\gamma+p}} + r_\tau(n)\left(n^{\frac{-2\beta}{2\beta+p}} + n^{\frac{-2\delta}{2\delta+p}}\right) + n^{-2\left(\frac{\alpha}{2\alpha+p}+\frac{\delta}{2\delta+p}\right)} + n^{-2\left(\frac{\beta}{2\beta+p}+\frac{\delta}{2\delta+p}\right)}.$$

If all nuisance parameters converge with the same minimax rate of $n^{\frac{-2\alpha}{2\alpha+p}}$, the rates of DRIV and our MRIV coincide. However, different to DRIV, our rate is additionally multiple robust in spirit of Theorem 1. This presents a crucial strength of our MRIV over DRIV: For example, if δ is small (slow convergence of $\hat{\pi}(x)$), our MRIV still with fast rate as long as α and β are large (i.e., if the other nuisance parameters are sufficiently smooth).

E.3 Wald estimator

Finally, we consider the Wald estimator [16] for the binary IV setting. More precisely, we estimate the ITE components $\mu_i^Y(x)$ and $\mu_i^A(x)$ seperately and plug them into

$$\tau(x) = \frac{\hat{\mu}_1^Y(x) - \hat{\mu}_0^Y(x)}{\hat{\mu}_1^A(x) - \hat{\mu}_0^A(x)}. \tag{52}$$

We consider two versions of the Wald estimator:

Linear: We use linear regressions to estimate the $\mu_i^Y(x)$ and logistic regressions to estimate the $\mu_i^A(x)$.

BART: We use Bayesian additive regression trees [3] trees to estimate the $\mu_i^Y(x)$ and random forest classifier to estimate the $\mu_i^A(x)$.

[2] For a detailed discussion on multiple robustness and the importance of the EIF parametrization, we refer to [18], Section 4.5.

[3] On a related note, a similar, important contribution of developing multiply robust method was recently made for the average treatment effect. Here, the estimator of [11] was extended by the estimator of [17] to allow for multi robustness. Yet, this different from our work in that it focuses on the average treatment effect, while we study the individual treatment effect in our paper.

F Visualization of predicted ITEs

We plot the predicted ITEs for the different baselines and MRIV-Net in Fig. 3 (for $n = 3000$). As expected, the linear methods (2SLS and linear Wald) are not flexible enough to provide accurate ITE estimates. We also observe that the curve of MRIV-Net without MRIV is quite wiggly, i.e., the estimator has a relatively large variance. This variance is reduced when the full MRIV-Net is applied. As a result, curve is much smoother. This is reasonable because MRIV does not estimate the ITE components individually, but estimates the ITE directly via the Stage 2 pseudo outcome regression. Overall, this confirms the superiority of our proposed framework.

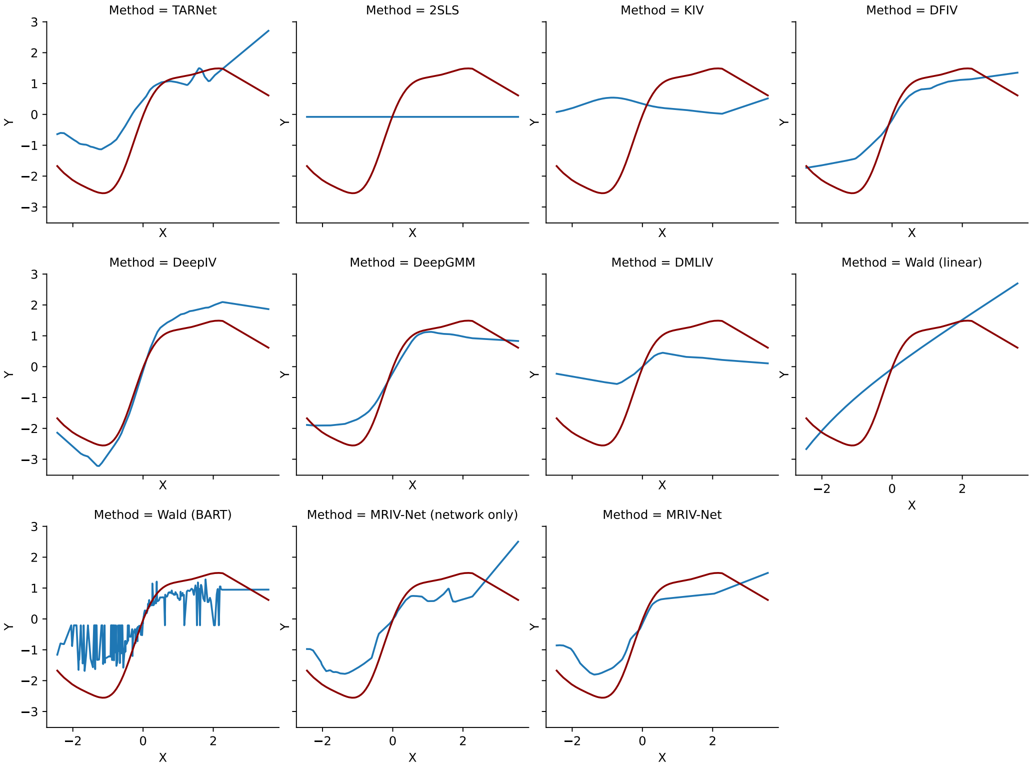

Figure 3: Predicted ITEs (blue) and oracle ITE (red) for different baselines.

G Implementation details and hyperparameter tuning

Implementation details for deep learning models: To make the performance of the deep learning models comparable, we implemented all feed-forward neural networks (including MRIV-Net) as follows: We use two hidden layers with RELU activation functions. We also incorporated a dropout layer for each hidden layer. We trained all models with the Adam optimizer [9] using 100 epochs. Exceptions are only DFIV and DeepGMM, where we used 200 epochs for training, accounting for slower convergence of the respective (adversarial) training algorithms. For DeepGMM, we further used Optimistic Adam [5] as in the original paper.

Training times: We report the approximate times needed to train the deep learning models on our simulated data with $n = 5000$ in Table 1. For training, we used an AMD Ryzen Pro 7 CPU. Compared to DMLIV and DRIV, the training of MRIV-Net is faster because only a single neural network is trained.

Table 1: Training times for deep learning models (in seconds).

TARNet	TARNet + DR	DFIV	DeepIV	DeepGMM	DMLIV	DMLIV + DRIV	MRIV-Net
$\sim$10.62	$\sim$28.57	$\sim$164.98	$\sim$30.21	$\sim$17.31	$\sim$74.98	$\sim$91.12	$\sim$32.20

Hyperparameter tuning: We performed hyperparameter tuning for all deep learning models (including MRIV-Net), KIV, and the BART Wald estimator on all datasets. For all methods except KIV and DFIV, we split the data into a training set (80%) and a validation set (20%). We then performed 40 random grid search iterations and chose the set of parameters that minimized the respective training loss on the validation set. In particular, the tuning procedure was the same for all baselines, which ensures that the performance gain of MRIV-Net is due to the method itself and not due to larger flexibility. Exceptions are only KIV and DFIV, for which we implemented the customized hyperparameter tuning algorithms proposed in [14] and [21] to ensure consistency with prior literature. For the meta learners (DR-learner, DRIV, and MRIV), we first performed hyperparameter tuning for the base methods and nuisance models, before tuning the pseudo outcome regression neural network by using the input from the tuned models. The tuning ranges for the hyperparameter are shown in Table 2. These include both the hyperparameter rangers shared across all neural networks and the model-specific hyperparameters. For reproducibility purposes, we publish the selected hyperparameters in our GitHub project as *.yaml* files.[4]

Table 2: Hyperparameter tuning ranges.

MODEL	HYPERPARAMETER	TUNING RANGE
Feed-forward neural networks (Shared parameter ranges for all deep learning baselines)	Hidden layer size(es)	p, $5p$, $10p$, $20p$, $30p$ (simulated data) p, $3p$, $5p$, $8p$, $10p$ (OHIE)
	Learning rate	0.0001, 0.0005, 0.001, 0.005, 0.01
	Batch size	64, 128, 256
	Dropout probability	0, 0.1, 0.2, 0.3
KIV	λ (Ridge penalty first stage)	5, 6, 7, 8, 9, 10, 12
	ξ (Ridge penalty second stage)	5, 6, 7, 8, 9, 10, 12
DFIV	λ_1 (Ridge penalty first stage)	0.0001, 0.001, 0.01, 0.1 (simulated data) 0.01, 0.05, 0.1 (OHIE)
	λ_2 (Ridge penalty second stage)	0.0001, 0.001, 0.01, 0.1 (simulated data) 0.01, 0.05, 0.1 (OHIE)
DeepGMM	λ_f (learning rate multiplier)	0.5, 1, 1.5, 2, 5
Wald (BART)	Number of trees (BART)	20, 30, 40, 50
	Number of trees (Random forest classifier)	20, 30, 40, 50

p = network input size

Hyperparameter robustness checks: We also investigate the robustness of MRIV-Net with respect to hyperparameter choice. To to this, we fix the optimal hyperparameter constellation for our simulated data for $n = 3000$ and perturb the hidden layer sizes, learning rate, dropout probability, and batch size.

[4] Codes are in the supplementary materials. Codes are also available at https://anonymous.4open.science/r/MRIV-Net-0AC4 (Upon acceptance, we replace the link and point to a public GitHub repository).

The results are shown in Fig. 4. We observe that the RMSE only changes marginally when perturbing the different hyperparameters, indicating that our method is to a certain degree robust against hyperparameter misspecification. Furthermore, our results indicate that the performance improvement of MRIV-Net over the baselines observed in our experiments is not due to hyperparameter tuning, but to our method itself.

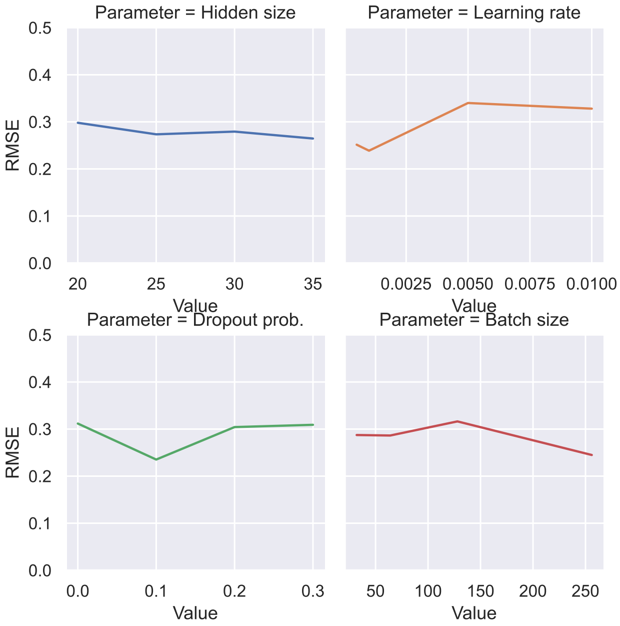

Figure 4: Robustness checks for different hyperparameters of MRIV-Net.

H Results for semi-synthetic data

In the main paper, we evaluated MRIV-Net both on synthetic and real-world data. Here, we provide additional results by constructing a semi-synthetic dataset on the basis of OHIE. It is common practice in causal inference literature to use semi-synthetic data for evaluation, because it combines advantages of both synthetic and real-world data. On the one hand, the real-world data part ensures that the data distribution is realistic and matches those in practice. On the other hand, the counterfactual ground-truth is still available, which makes it possible to measure the performance of ITE methods.

We construct our semi-synthetic data as follows: First, we extract the covariates $X \in \mathbb{R}^5$ and instruments $Z \in \{0, 1\}$ of our OHIE dataset from Sec. D. Then, we construct the treatment components $\mu_i^A(x)$ via

$$\mu_1^A(X) = 0.3 \cdot \sigma(X_1) + 0.7 \quad \text{and } \mu_0^A(X) = 0.3 \cdot \sigma(X_1), \tag{53}$$

where X_1 is the (standardized) age and $\sigma(\cdot)$ is the sigmoid function. The outcome components are constructed via

$$\mu_1^Y(X) = 0.5X_1^2 + \sum_{i=2}^{5} X_i^2 \quad \text{and } \mu_0^Y(X) = -0.5X_1^2 + \sum_{i=2}^{5} X_i^2. \tag{54}$$

We then sample treatments A and outcomes Y as in Eq. (31) and Eq. (32). Lemma 7 ensures that $\mu_i^Y(X) = \mathbb{E}[Y \mid Z = i, X]$ and $\mu_i^A(X) = \mathbb{E}[A \mid Z = i, X]$.

Given the above, the oracle ITE becomes

$$\tau(X) = \frac{X_1^2}{0.7}. \tag{55}$$

Note that $\tau(X)$ is sparse in the sense that it only depends on age, while the outcome components depend on all five covariates. Following our theoretical analysis in Sec. B, MRIV-Net should thus outperform methods that aim at estimating the components directly. This is confirmed in Table 3, where we show the results for all baselines and MRIV-Net on the semi-synthetic data. Indeed, we observe that MRIV-Net outperforms all other baselines, confirming both the superiority of our method as well as our theoretical results under sparsity assumptions from Sec. B.

Table 3: Results for semi-synthetic data.

Method	$n = 3000$	$n = 5000$	$n = 8000$
(1) STANDARD ITE			
TARNet [13]	1.66 ± 0.11	1.58 ± 0.07	1.57 ± 0.11
TARNet + DR [13, 8]	1.31 ± 0.28	1.22 ± 0.37	1.12 ± 0.15
(2) GENERAL IV			
2SLS [19]	1.34 ± 0.06	1.31 ± 0.03	1.32 ± 0.02
KIV [14]	1.97 ± 0.10	1.92 ± 0.05	1.93 ± 0.05
DFIV [21]	1.67 ± 0.44	1.63 ± 0.47	1.45 ± 0.17
DeepIV [7]	1.24 ± 0.26	0.99 ± 0.22	0.84 ± 0.19
DeepGMM [1]	1.39 ± 0.03	1.37 ± 0.16	1.18 ± 0.16
DMLIV [15]	2.12 ± 0.10	2.09 ± 0.09	2.02 ± 0.11
DMLIV + DRIV [15]	1.22 ± 0.10	1.18 ± 0.19	1.00 ± 0.08
(3) WALD ESTIMATOR [16]			
Linear	1.42 ± 0.24	1.28 ± 0.07	1.32 ± 0.07
BART	1.48 ± 0.24	1.29 ± 0.04	1.06 ± 0.13
MRIV-Net (network only)	1.11 ± 0.15	0.84 ± 0.14	0.95 ± 0.21
MRIV-Net (ours)	$\mathbf{0.71 \pm 0.24}$	$\mathbf{0.75 \pm 0.18}$	$\mathbf{0.78 \pm 0.26}$

Reported: RMSE (mean ± standard deviation). Lower = better (best in bold)

I Results for cross-fitting

Here, we repeat our experiments from the main paper but now make use of *cross-fitting*. Recall that, in Theorem 2, we assume that the nuisance parameter estimation and the pseudo-outcome regression are performed on three independent samples. We now address this through *cross-fitting*. To this end, our aim is to show that our proposed MRIV framework is again superior.

For MRIV, we proceeded as follows: We split the sample $\mathcal{D}$ into three equally sized samples $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$. We then trained $\hat{\tau}_{init}(x)$, $\hat{\mu}_0^Y(x)$, and $\hat{\mu}_0^A(x)$ on $\mathcal{D}_1$, $\hat{\delta}_A(x)$ and $\hat{\pi}(x)$ on $\mathcal{D}_2$, and performed the pseudo-outcome regression on $\mathcal{D}_3$. Then, we repeated the same training procedure two times, but performed the pseudo-outcome regression on $\mathcal{D}_2$ and $\mathcal{D}_1$. Finally, we averaged the resulting three ITE estimators. For DRIV, we implemented the cross-fitting procedure described in [15]. For the DR-learner, we followed [8].

The results are in Table H. Importantly, the results confirm the effectiveness of our proposed MRIV. Overall, we find that our proposed MRIV outperforms DRIV for the vast majority of base methods when performing cross-fitting. Furthermore, MRIV-Net is highly competitive even when comparing it with the cross-fitted estimators. This shows that our heuristic to learn separate representations instead of performing sample splits works in practice. In sum, the results confirm empirically that our MRIV is superior.

Table 4: Results for base methods with different meta-learners (i.e., DRIV, and our MRIV) using cross-fitting and results for MRIV-Net without cross-fitting.

Meta-learners	$n = 3000$		$n = 5000$		$n = 8000$	
Base methods	DRIV	MRIV (ours)	DRIV	MRIV (ours)	DRIV	MRIV (ours)
(1) Standard ITE						
TARNet [13]	**0.30 ± 0.02**	0.36 ± 0.16	0.18 ± 0.06	**0.16 ± 0.03**	0.21 ± 0.08	**0.13 ± 0.04**
TARNet + DR-learner [13, 8]	0.85 ± 0.11		0.66 ± 0.08		0.67 ± 0.12	
(2) General IV						
2SLS [19]	0.42 ± 0.11	**0.33 ± 0.09**	**0.20 ± 0.07**	0.23 ± 0.11	0.24 ± 0.10	**0.14 ± 0.02**
KIV [14]	0.47 ± 0.18	**0.45 ± 0.15**	0.20 ± 0.06	**0.19 ± 0.08**	0.22 ± 0.04	**0.15 ± 0.03**
DFIV [21]	0.35 ± 0.05	**0.28 ± 0.09**	0.22 ± 0.10	**0.18 ± 0.08**	0.24 ± 0.12	**0.16 ± 0.04**
DeepIV [7]	**0.38 ± 0.09**	0.44 ± 0.16	0.20 ± 0.07	**0.19 ± 0.07**	0.20 ± 0.08	**0.12 ± 0.02**
DeepGMM [1]	**0.42 ± 0.09**	**0.42 ± 0.16**	**0.19 ± 0.04**	**0.19 ± 0.07**	0.22 ± 0.06	**0.13 ± 0.02**
DMLIV [15]	**0.44 ± 0.09**	0.46 ± 0.16	0.21 ± 0.04	**0.19 ± 0.07**	0.21 ± 0.05	**0.14 ± 0.02**
(3) Wald estimator [16]						
Linear	0.47 ± 0.23	**0.36 ± 0.12**	0.24 ± 0.05	**0.20 ± 0.08**	0.22 ± 0.05	**0.15 ± 0.02**
BART	0.43 ± 0.12	**0.39 ± 0.12**	0.14 ± 0.05	**0.13 ± 0.05**	0.23 ± 0.08	**0.15 ± 0.02**
MRIV-Net\w network only (ours)	0.35 ± 0.12	**0.26 ± 0.11**	0.19 ± 0.13	**0.15 ± 0.03**	0.18 ± 0.08	**0.13 ± 0.03**

Reported: RMSE (mean ± standard deviation). Lower = better (best in bold)

References

[1] Andrew Bennett, Nathan Kallus, and Tobias Schnabel. “Deep generalized method of moments for instrumental variable analysis”. In: *NeurIPS*. 2019.

[2] Jean Chesson. “A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation”. In: *Journal of Applied Probability* 13.4 (1976), pp. 795–797.

[3] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. “BART: Bayesian additive regression trees”. In: *The Annals of Applied Statistics* 4.1 (2010), pp. 266–298.

[4] Alicia Curth and Mihaela van der Schaar. “Nonparametric estimation of heterogeneous treatment effects: From theory to learning Algorithms”. In: *AISTATS*. 2021.

[5] Constantinos Daskalakis et al. “Training GANs with optimism”. In: *ICLR*. 2018.

[6] Amy Finkelstein et al. “The oregon health insurance experiment: Evidence from the first year”. In: *The Quarterly Journal of Economics* 127.3 (2012), pp. 1057–1106.

[7] Jason Hartford et al. “Deep IV: A flexible approach for counterfactual prediction”. In: *ICML*. 2017.

[8] Edward H. Kennedy. “Optimal doubly robust estimation of heterogeneous causal effects”. In: *arXiv preprint* (2020).

[9] Diederik P. Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *ICLR*. 2015.

[10] Whitney K. Newey and James L. Powell. “Instrumental variable estimation of nonparametric models”. In: *Econometrica* 71.5 (2003), pp. 1565–1578.

[11] Ryo Okui et al. “Doubly robust instrumental variable regression”. In: *Statistica Sinica* 22.1 (2012), pp. 173–205.

[12] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. 3. print. Adaptive computation and machine learning. Cambridge, Mass.: MIT Press, 2008.

[13] Uri Shalit, Fredrik D. Johansson, and David Sontag. “Estimating individual treatment effect: Generalization bounds and algorithms”. In: *ICML*. 2017.

[14] Rahul Singh, Maneesh Sahani, and Arthur Gretton. “Kernel instrumental variable regression”. In: *NeurIPS*. 2019.

[15] Vasilis Syrgkanis et al. “Machine learning estimation of heterogeneous treatment effects with instruments”. In: *NeurIPS*. 2019.

[16] Abraham Wald. “The fitting of straight lines if both variables are subject to error”. In: *Annals of Mathematical Statistics* 11.3 (1940), pp. 284–300.

[17] Linbo Wang and Eric J. Tchetgen Tchetgen. “Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables”. In: *Journal of the Royal Statistical Society: Series B* 80.3 (2018), pp. 531–550.

[18] Yixin Wang and David M. Blei. “The blessings of multiple causes”. In: *Journal of the American Statistical Association* 114.528 (2019), pp. 1574–1596.

[19] Jeffrey M. Wooldridge. *Introductory Econometrics: A modern approach*. Routledge, 2013.

[20] Phillip G. Wright. *The tariff on animal and vegitable oils*. New York: Macmillan, 1928.

[21] Liyuan Xu et al. “Learning deep features in instrumental variable regression”. In: *ICLR*. 2021.

[22] Yun Yang and Surya T. Tokdar. “Minimax-optimal nonparametric regression in high dimensions”. In: *The Annals of Statistics* 43.2 (2015), pp. 652–674.