Moderately-Balanced Representation Learning for Orthogonal Estimation on Average Treatment Effect

Anonymous Authors

1 Experiments

The final full results for 1000 IHDP and 100 Twins experiments are reported in Table 1. The mark * indicates that the baseline models did not report relevant results.

As a supplementary note, the outcome takes a binary value for Twins experiments. So the factual outcome loss in Eqn (6) will be

$$\mathcal{L}_{fo} = -\frac{1}{N}\sum_{m=1}^{N}[y_m \log f(\Phi(\mathbf{z}_m)) + (1-y_m)\log(1-f(\Phi(\mathbf{z}_m)))].$$

2 Assumptions

Assumption 1 (SUTVA)**.** *The potential outcomes for any individual are not affected by the treatment assignment of other individuals.*

Assumption 2 (Strong Ignorability)**.** *Given the covariates* $\mathbf{Z}$*, the potential outcomes are independent of the treatment assignment* D*:* $(Y(0), Y(1)) \perp\!\!\!\perp D \mid \mathbf{Z}$.

Assumption 3 (Overlap)**.** *The probability of treatment assignment for any unit is positive:* $0 < Pr(D = d \mid \mathbf{Z} = \mathbf{z}) < 1, \forall\, d \in \{0,1\}$ *and* $\mathbf{z} \in \mathcal{Z}$.

Assumption 4 (Consistency)**.** *The potential outcome for treatment* d *of each unit is equal to the observed factual outcome if the actual treatment is* d*:* $(Y(d) = Y^F) \mid D = d, \forall\, d \in \{0,1\}$.

3 Proofs

We skip the proofs of Proposition 1 since it can be seen in [Chernozhukov *et al.*, 2018]. In the following, we prove Proposition 2 and Property 1.

3.1 Proof of Proposition 2

Proof. The score functions stated in the Eqn. (2) and Eqn. (3) in the main paper are

$$\begin{aligned}
&\psi_1(W,\theta^i,\rho) = \theta^i - g(i,\mathbf{Z}) \\
&- (Y - g(i,\mathbf{Z}))\frac{iD + (1-i)(1-D)}{im(\mathbf{Z}) + (1-i)(1-m(\mathbf{Z}))}; \\
&\psi_2(W,\theta^i,\rho) = \theta^i - g(i,\mathbf{Z}) \\
&- (Y(i) - g(i,\mathbf{Z}))\frac{((D-m(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])^2}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}.
\end{aligned}$$

We then check if the orthogonal condition (Definition 1) holds for $\psi_1(W,\theta^i,\rho)$.

$$\begin{aligned}
&\partial_g \psi_1(W,\theta^i,\rho) = -1 + \frac{iD + (1-i)(1-D)}{im(\mathbf{Z}) + (1-i)(1-m(\mathbf{Z}))}; \\
&\partial_m \psi_1(W,\theta^i,\rho) = (Y - g(i,\mathbf{Z}))\frac{D}{m(\mathbf{Z})^2},\ if\ i = 1; \\
&\partial_m \psi_1(W,\theta^i,\rho) = -(Y - g(i,\mathbf{Z}))\frac{1-D}{(1-m(\mathbf{Z}))^2},\ if\ i = 0.
\end{aligned}$$

If $i = 1$ and under the noise conditions $\mathbb{E}[\nu \mid \mathbf{Z}] = 0$ and $\mathbb{E}[\xi \mid D, \mathbf{Z}] = 0$, then we have

$$\begin{aligned}
&\mathbb{E}\left[\partial_g \psi_1(W,\theta^1,\rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^1=\theta^1_0} \\
&= -1 + \mathbb{E}\left[\frac{D}{m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 + \mathbb{E}\left[\frac{D - m_0(\mathbf{Z}) + m_0(\mathbf{Z})}{m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 + \mathbb{E}\left[\frac{\nu + m_0(\mathbf{Z})}{m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 + \frac{\mathbb{E}[\nu \mid \mathbf{Z}]}{m_0(\mathbf{Z})} + 1 \\
&= 0.
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}\left[\partial_m \psi_1(W,\theta^1,\rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^1=\theta^1_0} \\
&= \mathbb{E}\left[(Y - g_0(i,\mathbf{Z}))\frac{D}{m_0(\mathbf{Z})^2} \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(Y - g_0(i,\mathbf{Z}))\frac{D}{m_0(\mathbf{Z})^2} \mid D, \mathbf{Z}\right] \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\xi\frac{D}{m_0(\mathbf{Z})^2} \mid D, \mathbf{Z}\right] \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\mathbb{E}[\xi \mid D, \mathbf{Z}]\frac{D}{m_0(\mathbf{Z})^2} \mid \mathbf{Z}\right] \\
&= 0.
\end{aligned}$$

Table 1: The full results of 1000 IHDP and 100 Twins experiments.

Method	IHDP In-sample $\sqrt{\epsilon_{PEHE}}$	IHDP In-sample ϵ_{ATE}	IHDP Out-of-sample $\sqrt{\epsilon_{PEHE}}$	IHDP Out-of-sample ϵ_{ATE}	Twins In-sample AUC	Twins In-sample ϵ_{ATE}	Twins Out-of-sample AUC	Twins Out-of-sample ϵ_{ATE}
OLS/LR$_1$	$5.8 \pm .3$	$.73 \pm .04$	$5.8 \pm .3$	$.94 \pm .06$	$.660 \pm .005$	$.004 \pm .003$	$.500 \pm .028$	$.007 \pm .006$
OLS/LR$_2$	$2.4 \pm .1$	$.14 \pm .01$	$2.5 \pm .1$	$.31 \pm .02$	$.660 \pm .004$	$.004 \pm .003$	$.500 \pm .016$	$.007 \pm .006$
k-NN	$2.1 \pm .1$	$.14 \pm .01$	$4.1 \pm .2$	$.79 \pm .05$	$.609 \pm .010$	$\mathbf{.003 \pm .002}$	$.492 \pm .012$	$\mathbf{.005 \pm .004}$
BART	$2.1 \pm .1$	$.23 \pm .01$	$2.3 \pm .1$	$.34 \pm .02$	$.506 \pm .014$	$.121 \pm .024$	$.500 \pm .011$	$.127 \pm .024$
CF	$3.8 \pm .2$	$.18 \pm .01$	$3.8 \pm .2$	$.40 \pm .03$	*	$.029 \pm .004$	*	$.034 \pm .008$
CEVAE	$2.7 \pm .1$	$.34 \pm .01$	$2.6 \pm .1$	$.46 \pm .02$	$.845 \pm .003$	$.022 \pm .002$	$.841 \pm .004$	$.032 \pm .003$
SITE	$.69 \pm .0$	$.22 \pm .01$	$.75 \pm .0$	$.24 \pm .01$	$.862 \pm .002$	$.016 \pm .001$	$.853 \pm .006$	$.020 \pm .002$
GANITE	$1.9 \pm .4$	$.43 \pm .05$	$2.4 \pm .4$	$.49 \pm .05$	*	$.006 \pm .002$	*	$.009 \pm .008$
BLR	$5.8 \pm .3$	$.72 \pm .04$	$5.8 \pm .3$	$.93 \pm .05$	$.611 \pm .009$	$.006 \pm .004$	$.510 \pm .018$	$.033 \pm .009$
BNN	$2.2 \pm .1$	$.37 \pm .03$	$2.1 \pm .1$	$.42 \pm .03$	$.690 \pm .008$	$.006 \pm .003$	$.676 \pm .008$	$.020 \pm .007$
TARNet	$.88 \pm .0$	$.26 \pm .01$	$.95 \pm .0$	$.28 \pm .01$	$.849 \pm .002$	$.011 \pm .002$	$.840 \pm .006$	$.015 \pm .002$
CFR-WASS	$.71 \pm .0$	$.25 \pm .01$	$.76 \pm .0$	$.27 \pm .01$	$.850 \pm .002$	$.011 \pm .002$	$.842 \pm .005$	$.028 \pm .003$
Dragonnet	$1.3 \pm .4$	$.14 \pm .01$	$1.3 \pm .5$	$.20 \pm .05$	*	$.006 \pm .005$	*	$.006 \pm .005$
MBRL	$\mathbf{.522 \pm .007}$	$.121 \pm .005$	$\mathbf{.565 \pm .008}$	$\mathbf{.133 \pm .005}$	$\mathbf{.879 \pm .000}$	$\mathbf{.003 \pm .000}$	$\mathbf{.874 \pm .001}$	$.007 \pm .001$
MBRL+θ_1^i	$\mathbf{.522 \pm .007}$	$\mathbf{.102 \pm .004}$	$\mathbf{.565 \pm .008}$	$.166 \pm .007$	$\mathbf{.879 \pm .000}$	$\mathbf{.003 \pm .000}$	$\mathbf{.874 \pm .001}$	$.008 \pm .000$
MBRL+θ_2^i	$\mathbf{.522 \pm .007}$	$.114 \pm .005$	$\mathbf{.565 \pm .008}$	$.204 \pm .008$	$\mathbf{.879 \pm .000}$	$\mathbf{.003 \pm .000}$	$\mathbf{.874 \pm .001}$	$.006 \pm .001$

If $i = 0$ and under the noise conditions $\mathbb{E}[\nu \mid \mathbf{Z}] = 0$ and $\mathbb{E}[\xi \mid D, \mathbf{Z}] = 0$, then we have

$$
\begin{aligned}
&\mathbb{E}\left[\partial_g \psi_1(W, \theta^0, \rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^0=\theta^0_0} \\
&= -1 + \mathbb{E}\left[\frac{1-D}{1-m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 + \mathbb{E}\left[\frac{1-D-m_0(\mathbf{Z})+m_0(\mathbf{Z})}{1-m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 + \mathbb{E}\left[\frac{-\nu+1-m_0(\mathbf{Z})}{1-m_0(\mathbf{Z})} \mid \mathbf{Z}\right] \\
&= -1 - \frac{\mathbb{E}[\nu \mid \mathbf{Z}]}{1-m_0(\mathbf{Z})} + 1 \\
&= 0.
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{E}\left[\partial_m \psi_1(W, \theta^0, \rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^0=\theta^0_0} \\
&= \mathbb{E}\left[-(Y - g_0(i,\mathbf{Z}))\frac{1-D}{(1-m_0(\mathbf{Z}))^2} \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[-(Y - g_0(i,\mathbf{Z}))\frac{1-D}{(1-m_0(\mathbf{Z}))^2} \mid D, \mathbf{Z}\right] \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[-\xi\frac{1-D}{(1-m_0(\mathbf{Z}))^2} \mid D, \mathbf{Z}\right] \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[-\mathbb{E}[\xi \mid D, \mathbf{Z}]\frac{1-D}{(1-m_0(\mathbf{Z}))^2} \mid \mathbf{Z}\right] \\
&= 0.
\end{aligned}
$$

We then check if the orthogonal condition (Definition 1) holds for $\psi_2(W, \theta^i, \rho)$.

$$
\partial_m \psi_2(W, \theta^i, \rho) = (Y(i) - g(i,\mathbf{Z}))\frac{2((D-m(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}
$$

$$
\partial_g \psi_2(W, \theta^i, \rho) = -1 + \frac{((D-m(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])^2}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}
$$

Using the noise condition $\mathbb{E}[\nu \mid \mathbf{Z}] = 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\partial_g \psi_2(W, \theta^i, \rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^i=\theta^i_0} \\
&= -1 + \mathbb{E}\left[\frac{((D-m_0(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])^2}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]} \mid \mathbf{Z}\right] \\
&= -1 + \frac{1}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}\mathbb{E}\left[((D-m_0(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])^2 \mid \mathbf{Z}\right] \\
&= -1 + \frac{1}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}\mathbb{E}[(D-m_0(\mathbf{Z}))^2 + (\mathbb{E}[\nu \mid \mathbf{Z}])^2 \\
&- 2(D-m_0(\mathbf{Z}))\mathbb{E}[\nu \mid \mathbf{Z}] \mid \mathbf{Z}] \\
&= -1 + \frac{1}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}\left[\mathbb{E}\left[\nu^2 \mid \mathbf{Z}\right] + (\mathbb{E}[\nu \mid \mathbf{Z}])^2 - 2(\mathbb{E}[\nu \mid \mathbf{Z}])^2\right] \\
&= -1 + \frac{1}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}\left[\mathbb{E}\left[\nu^2 \mid \mathbf{Z}\right] - (\mathbb{E}[\nu \mid \mathbf{Z}])^2\right] \\
&= -1 + \frac{\mathbb{E}\left[\nu^2 \mid \mathbf{Z}\right]}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]} = 0.
\end{aligned}
$$

By the model setup $Y = g_0(D, \mathbf{Z}) + \xi$, we have the underlying relation for the potential outcome $Y(i)$ that $Y(i) = g_0(i, \mathbf{Z}) + \xi$. Using the noise condition $\mathbb{E}[\xi \mid D, \mathbf{Z}] = 0$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\partial_m \psi_2(W, \theta^i, \rho) \mid \mathbf{Z}\right] |_{(g,m)=(g_0,m_0),\theta^i=\theta^i_0} \\
&= \mathbb{E}\left[\mathbb{E}\left[\xi\frac{2((D-m_0(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]} \mid D, \mathbf{Z}\right] \mid \mathbf{Z}\right] \\
&= \mathbb{E}\left[\frac{2((D-m_0(\mathbf{Z})) - \mathbb{E}[\nu \mid \mathbf{Z}])}{\mathbb{E}[\nu^2 \mid \mathbf{Z}]}\mathbb{E}[\xi \mid D, \mathbf{Z}] \mid \mathbf{Z}\right] \\
&= 0.
\end{aligned}
$$

Therefore, the noise conditions $\mathbb{E}[\xi \mid D, \mathbf{Z}] = 0$ and $\mathbb{E}[\nu \mid \mathbf{Z}] = 0$ are sufficient for the score functions ψ_1 and ψ_2 satisfying the orthogonal condition. $\square$

3.2 Proof of Property 1

Proof. Using the noise condition $\mathbb{E}[\xi \mid D, \mathbf{Z}] = 0$, we have

$$
\begin{aligned}
&\mathbb{E}[(Y - g_0(D, \mathbf{Z}))(D - m_0(\mathbf{Z}))] \\
=&\mathbb{E}[\mathbb{E}[(Y - g_0(D, \mathbf{Z}))(D - m_0(\mathbf{Z})) \mid D, \mathbf{Z}]] \\
=&\mathbb{E}[(D - m_0(\mathbf{Z}))\mathbb{E}[(Y - g_0(D, \mathbf{Z})) \mid D, \mathbf{Z}]] \\
=&\mathbb{E}[(D - m_0(\mathbf{Z}))\mathbb{E}[\xi \mid D, \mathbf{Z}]] \\
=&0.
\end{aligned}
$$

□

References

[Chernozhukov *et al.*, 2018] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

[Huang *et al.*, 2021] Yiyan Huang, Cheuk Hang Leung, Xing Yan, and Qi Wu. Higher-order orthogonal causal learning for treatment effect. *arXiv preprint arXiv:2103.11869*, 2021.