Supplementary for Theorem 3.1

proof for Theorem 3.1

Proof. Overall speaking, the derivation of the variance bound is built upon the classic theory in Bickel et. al. (1993) (Section 3). Denote $f(y(0), y(1), |\boldsymbol{x})$ as the conditional probability density of the potential outcomes $(Y(0), Y(1))$ given $\boldsymbol{x}$, $e(\boldsymbol{x}) := p(T=1|\boldsymbol{x})$, and $f(\boldsymbol{x})$ as the density of $\boldsymbol{x}$. Then the joint density of $(Y(0), Y(1), T, \boldsymbol{X})$ is

$$p(y(0), y(1), t, x) = f(y(0), y(1), |\boldsymbol{x})e(\boldsymbol{x}).$$

Let $f_1(\cdot|\boldsymbol{x}) := \int f(y(0), \cdot)d_{y(0)}$ and $f_0(\cdot|\boldsymbol{x}) := \int f(\cdot, y(1))d_{y(1)}$, then joint density of $(Y, T, \boldsymbol{X})$ is

$$p(y, t, x) = [f_1(y|\boldsymbol{x})e(\boldsymbol{x})]^t[f_0(y|\boldsymbol{x})(1-e(\boldsymbol{x}))]^{1-t}f(\boldsymbol{x}).$$

Consider a parametric model $p(y, t, \boldsymbol{x}|\theta) = [f_1(y|\boldsymbol{x}, \theta)e(\boldsymbol{x}, \theta)]^t[f_0(y|\boldsymbol{x}, \theta)(1-e(\boldsymbol{x}, \theta))]^{1-t}f(\boldsymbol{x}, \theta)$ such that $p(y, t, x|\theta_0) = p(y, t, \boldsymbol{x})$. The corresponding score $s(t, y, \boldsymbol{x}|\theta) = \frac{\partial}{\partial\theta}\log p(y, t, \boldsymbol{x}|\theta)$ is

$$\begin{aligned} s(t, y, x \mid \theta) \equiv t \cdot [\frac{\partial}{\partial\theta}\log f_1(y|\boldsymbol{x}, \theta)] + (1-t) \cdot [\frac{\partial}{\partial\theta}\log f_0(y|\boldsymbol{x}, \theta)] \\ + \frac{t - p(x, \theta)}{e(x, \theta)(1-e(x, \theta))} \cdot [\frac{\partial}{\partial\theta}e(\boldsymbol{x}, \boldsymbol{\theta})] + \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}, \theta). \end{aligned}$$

Then the tangent space of the model $p(y, t, \boldsymbol{x}|\theta)$ is

$$\mathcal{P} := \left\{ t \cdot [\frac{\partial}{\partial\theta}\log f_1(y|\boldsymbol{x}] + (1-t) \cdot [\frac{\partial}{\partial\theta}\log f_0(y|\boldsymbol{x})] + a(\boldsymbol{x})(t - e(\boldsymbol{x})) + \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}) \right\},$$

where $a(\boldsymbol{x})$ is any square-integrable measurable function of $\boldsymbol{x}$. Now we turn back to the formulation of $\tau(\boldsymbol{x})$,

$$\tau_\theta(\boldsymbol{x}) = \int y f_1(y|\boldsymbol{x}, \theta)dy - \int y f_0(y|\boldsymbol{x}, \theta)dy.$$

Let $F_\tau(Y, T, \boldsymbol{X}) = \frac{T}{e(\boldsymbol{X})}(Y - \mu_1(\boldsymbol{X})) - \frac{1-T}{1-e(\boldsymbol{X})}(Y - \mu_0(\boldsymbol{X})))$, where $\mu_t(\boldsymbol{X}) := \mathbb{E}[Y(t)|\boldsymbol{X}]$. Then we may validate that

$$\frac{\partial}{\partial\theta}\tau(\theta_0) = \mathbb{E}[F_\tau(Y, T, \boldsymbol{X})s(T, Y, X|\theta_0)].$$

Then based on the result in Bickel et. al. (1993), the variance bound of $\tau(\boldsymbol{x})$ is the expected squares of the projection $F_\tau(Y, T, \boldsymbol{X})$ in the tangent space $\mathcal{P}$, which is equal to $\mathbb{E}[\frac{\sigma_1^2(\boldsymbol{X})}{e(\boldsymbol{X})} + \frac{\sigma_0^2(\boldsymbol{X})}{1-e(\boldsymbol{X})}]$. □

Proof. Overall speaking, the derivation of the variance bound is built upon the classic theory in Bickel et. al. (1993) (Section 3). Denote $f(y(0), y(1), |\boldsymbol{x})$ as the conditional probability density of the potential outcomes $(Y(0), Y(1))$ given $\boldsymbol{x}$, $e(\boldsymbol{x}) := p(T=1|\boldsymbol{x})$, and $f(\boldsymbol{x})$ as the density of $\boldsymbol{x}$. Then the joint density of $(Y(0), Y(1), T, \boldsymbol{X})$ is

$$p(y(0), y(1), t, x) = f(y(0), y(1), |\boldsymbol{x})e(\boldsymbol{x}).$$

Let $f_1(\cdot|\boldsymbol{x}) := \int f(y(0), \cdot)d_{y(0)}$ and $f_0(\cdot|\boldsymbol{x}) := \int f(\cdot, y(1))d_{y(1)}$, then joint density of $(Y, T, \boldsymbol{X})$ is

$$p(y, t, x) = [f_1(y|\boldsymbol{x})e(\boldsymbol{x})]^t[f_0(y|\boldsymbol{x})(1-e(\boldsymbol{x}))]^{1-t}f(\boldsymbol{x}).$$

Consider a parametric model $p(y, t, \boldsymbol{x}|\theta) = [f_1(y|\boldsymbol{x}, \theta)e(\boldsymbol{x}, \theta)]^t[f_0(y|\boldsymbol{x}, \theta)(1-e(\boldsymbol{x}, \theta))]^{1-t}f(\boldsymbol{x}, \theta)$ such that $p(y, t, x|\theta_0) = p(y, t, \boldsymbol{x})$. The corresponding score $s(t, y, \boldsymbol{x}|\theta) = \frac{\partial}{\partial\theta}\log p(y, t, \boldsymbol{x}|\theta)$ is

$$\begin{aligned} s(t, y, x \mid \theta) \equiv t \cdot [\frac{\partial}{\partial\theta}\log f_1(y|\boldsymbol{x}, \theta)] + (1-t) \cdot [\frac{\partial}{\partial\theta}\log f_0(y|\boldsymbol{x}, \theta)] \\ + \frac{t - p(x, \theta)}{e(x, \theta)(1-e(x, \theta))} \cdot [\frac{\partial}{\partial\theta}e(\boldsymbol{x}, \boldsymbol{\theta})] + \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}, \theta). \end{aligned}$$

Then the tangent space of the model $p(y, t, \boldsymbol{x}|\theta)$ is

$$\mathcal{P} := \left\{ t \cdot [\frac{\partial}{\partial\theta}\log f_1(y|\boldsymbol{x}] + (1-t) \cdot [\frac{\partial}{\partial\theta}\log f_0(y|\boldsymbol{x})] + a(\boldsymbol{x})(t - e(\boldsymbol{x})) + \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}) \right\},$$

where $a(\boldsymbol{x})$ is any square-integrable measurable function of $\boldsymbol{x}$. Now we turn back to the formulation of $\tau(\boldsymbol{x})$,

$$\tau_\theta(\boldsymbol{x}) = \int y f_1(y|\boldsymbol{x},\theta)dy - \int y f_0(y|\boldsymbol{x},\theta)dy.$$

Let $F_\tau(Y,T,\boldsymbol{X}) = \frac{T}{e(\boldsymbol{X})}(Y-\mu_1(\boldsymbol{X})) - \frac{1-T}{1-e(\boldsymbol{X})}(Y-\mu_0(\boldsymbol{X})))$, where $\mu_t(\boldsymbol{X}) := \mathbb{E}[Y(t)|\boldsymbol{X}]$. Then we may validate that

$$\frac{\partial}{\partial\theta}\tau(\theta_0) = \mathbb{E}[F_\tau(Y,T,\boldsymbol{X})s(T,Y,X|\theta_0)].$$

Then based on the result in Bickel et. al. (1993), the variance bound of $\tau(\boldsymbol{x})$ is the expected squares of the projection $F_\tau(Y,T,\boldsymbol{X})$ in the tangent space $\mathcal{P}$, which is equal to $\mathbb{E}[\frac{\sigma_1^2(\boldsymbol{X})}{e(\boldsymbol{X})} + \frac{\sigma_0^2(\boldsymbol{X})}{1-e(\boldsymbol{X})}]$. $\square$

Extension of Theorem 3.1

- when T multi-class categorical: $\mathbb{V} = \sum_{t\in\mathcal{T}} \mathbb{E}\left[\frac{\sigma_t^2(\boldsymbol{X})}{e_t(\boldsymbol{X})}\right]$, where $\sigma_t^2(\boldsymbol{X}) = \mathrm{Var}[Y(t)|\boldsymbol{X}]$ and $e_t(\boldsymbol{X}) = \mathbb{P}(T=t|\boldsymbol{X})$.
- when T continuous: $\mathbb{V} = \int_{t\in\mathcal{T}} \mathbb{E}\left[\frac{\sigma_t^2(\boldsymbol{X})}{e_t(\boldsymbol{X})}\right] dt$, where $\sigma_t^2(\boldsymbol{X}) = \mathrm{Var}[Y(t)|\boldsymbol{X}]$ and $e_t(\boldsymbol{X})$ is the conditional density of T given $\boldsymbol{X}$.

Supplementary for Proposition 3.1

Definition. 3.1 *Define Instrumental variables (I), Confounders (C), Adjustment variables (A) as*

$\boldsymbol{I} = \{X_i|$ *there exists an unblocked path from X_i to T and $X_i \notin PA(Y)$ and X_i is not a collider*$\}$;
$\boldsymbol{C} = \{X_i|$ *there exists an unblocked path from X_i to T and $X_i \in PA(Y)$*$\}$;
$\boldsymbol{A} = \{X_i|$ *there exists an unblocked path from X_i to Y, and no unblocked paths from X_i to T*$\}$,
where $PA(Y)$ denotes the set of parent nodes of Y.

Proposition. 3.1 *Let $\boldsymbol{I}, \boldsymbol{C}, \boldsymbol{A}$ be the variables set in Definition 3.1. Then (i) $\boldsymbol{C}$ blocks all the back-door paths from T to Y; (ii) $P(Y|\boldsymbol{X}, do(t)) = P(Y|\boldsymbol{C}, \boldsymbol{A}, do(t))$*

Proof. (i) All the back-door paths from T to Y are in the form $T \leftarrow \cdots Y$. Then we have two sub-cases according to the nearest edge to Y,
1) If the path is in the form $T \leftarrow \cdots \leftarrow Y$: since T is earlier than Y, there exist no directed paths from Y to T, so there exists at least one collider on this path. Let X_i denote the one nearest to Y, then the path is in the form $T \leftarrow \cdots \rightarrow X_i \leftarrow \cdots \leftarrow Y$. Therefore, this path is blocked by empty set.
2) If the path is in the form $T \leftarrow \cdots \rightarrow Y$, let X_i be the one nearest to Y in the form $T \leftarrow \cdots X_i \rightarrow Y$. If the first segment $T \leftarrow \cdots X_i$ is an unblocked path, then $X_i \in \boldsymbol{C}$ and hence the path is blocked by $\boldsymbol{C}$. Otherwise, if the first segment $T \leftarrow \cdots X_i$ is a blocked path (blocked by empty set), then the whole path $T \leftarrow \cdots X_i \rightarrow Y$ is also blocked by empty set.
In summary, any back-door path from T to Y is either blocked by $\boldsymbol{C}$ or empty set. Thus $\boldsymbol{C}$ blocks all the back-door from T to Y. (ii) Let $G_{\overline{T}}$ be the causal graph by removing all the edges into T, then it suffices to show that all the paths from $X \in \boldsymbol{X}$ and Y are blocked by $\boldsymbol{C} \cup \boldsymbol{A}$. Suppose π is an unblocked path between X and Y. First, note that π would not be a directed path from Y to X since X is a pre-treatment variable. Second, $PA(Y)$ is on the path π, otherwise π is a blocked path with collider(s). Finally, note that $PA(Y)$ is either in $\boldsymbol{C}$ (if $PA(Y)$ has unblocked path to T) or $\boldsymbol{A}$ (if $PA(Y)$ has no unblocked paths to T), we may conclude that the path is blocked by $\boldsymbol{C} \cup \boldsymbol{A}$

$\square$

Supplementary for Theorem 3.2

Theorem. 3.2 *The $\{\boldsymbol{I}, \boldsymbol{C}, \boldsymbol{A}\}$ are identifiable from the joint distribution $\mathbb{P}(\boldsymbol{X}, T, Y)$ as follows*

- $X_i \in \boldsymbol{A} \Leftrightarrow \left\{X_i | X_i \perp T \text{ and } X_i \not\perp Y\right\}$
- $X_i \in \boldsymbol{I} \Leftrightarrow \left\{X_i | X_i \notin \boldsymbol{A}, X_i \not\perp T, \text{ and there exists a subset } \boldsymbol{X}' \subset \boldsymbol{X} \ \ s.t.\ \ X_i \perp Y | \boldsymbol{X}' \cup \{T\}\right\}$
- $X_i \in \boldsymbol{C} \Leftrightarrow \left\{X_i | X_i \notin \boldsymbol{A} \text{ and } X_i \notin \boldsymbol{I} \text{ and } X_i \not\perp T \text{ and } X_i \not\perp Y\right\}$

Further, the confounders $\boldsymbol{C}$ may serve as the variables set $\boldsymbol{X}'$, i.e., *$X_i \perp Y | \boldsymbol{C} \cup \{T\}$ for $X_i \in \boldsymbol{I}$.*

Proof. 1) As for $X_i \in \boldsymbol{A}$, according to Definition 3.1 and the d-separation criterion, $X_i \perp T$ and $X_i \not\perp Y$. Now we show $\boldsymbol{A}$ is not empty, i.e., X_i that has an unblocked path to Y may have no unblocked paths to T. The path between X_i and Y is in the form $X_i \cdots \to Y$ or $X_i \cdots \leftarrow Y$. Note that X_i is a pre-treatment variable, the latter one has at least one collider otherwise there exists a directed path from Y to X_i. So we may only consider the form $X_i \cdots \to Y$, note that there is a directed path $T \to Y$, the path $T \to Y \leftarrow \cdots X_i$ is an unblocked path as Y is a collider.
2) Let π denote the path between $X_i \in \boldsymbol{I}$ and Y.
(a) If T is on path π, we have two sub-cases depending on whether T is a collider. (a.1) If T is a collider such that $\pi = I \cdots \to T \leftarrow \cdots Y$, then the second segment $T \leftarrow \cdots Y$ is either $T \leftarrow \cdots \to Y$ or $T \leftarrow \cdots \leftarrow Y$. For the former one, let $X^{'} \in \text{PA}(Y)$ be the covariate closet to Y. Note that X_i is not a collider and has an unblocked path to T, thus $X^{'}$ also has an unblocked path to T, and hence $X^{'} \in \boldsymbol{C}$ and the path is blocked by $\boldsymbol{C}$. For the latter one, $T \leftarrow \cdots \leftarrow Y$ must be a blocked path (because T is prior to Y, there would be a collider in this case). (a.2) If T is not a collider such that $\pi = I \cdots \to T \to \cdots T$, then the path is blocked by T. To summarize, for a path π with T, the path is blocked by $T \cup \boldsymbol{C}$.
(b) If π does not pass T and is an unblocked path without collider(s), then π must be in the form $X_i \cdots \to Y$ since X_i is prior to Y. Denote X_j as the parent node of Y on this path, note X_j has an unblocked path to X_i and X_i has an unblocked path to T, we conclude that $X_j \in \text{PA}(Y)$ has an unblocked path to T. Thus we have $X_j \in \boldsymbol{C}$, and π is blocked by $\boldsymbol{C}$.
Overall, based on (a) and (b), any path π between $X_i \in \boldsymbol{I}$ and Y is blocked by $\boldsymbol{C} \cup T$.
3) The equivalent condition for $\boldsymbol{C}$ is readily from the definition. Since $X_i \in \boldsymbol{C}$ has unblocked paths to both T and Y, we have $X_i \not\perp T$ and $X_i \not\perp Y$.

$\square$

Supplementary for Proposition 3.2

Proposition. *Denote $l(\cdot,\cdot)$ as the cross-entropy loss (for categorical) or l_2 loss (for numerical). Let $\hat{h}_{A\to T}(\cdot) := \arg\min_h l(h(A(X)), T)$ for given $A(\cdot)$, $\hat{h}_{C\cup T\to Y}(\cdot) := \arg\min_h \mathcal{L}(h(\ C(X)\cup T), Y)$, $\hat{h}_{I\cup C\cup T\to Y}(\cdot) := \arg\min_h l(h(C(X) \cup I(X) \cup T), Y)$ for given $C(\cdot)$ and $I(\cdot)$. Then*

(i) let $L_A := l(\hat{h}_{A\to T}(A(x)), T)$, then L_A is maximized when $A(X) \perp T$;
(ii) let $L_{I,C} := l_d\big(\hat{h}_{C\cup T\to Y}(C(X)\cup T), \hat{h}_{I\cup C\cup T\to Y}(I(X)\cup C(X) \cup T)\big)$, where $l_d()$ denote the KL divergence (categorical Y) or l_2 loss (numerical Y), then $L_{I,C}$ is minimized when $I(X) \perp Y | \{T, C(X)\}$.

Proof. Firstly, suppose that T is binary and $l(\cdot,\cdot)$ denotes the cross-entropy loss, let

$$\begin{aligned}
\mathcal{L}_A^h &= -\sum_i \{T_i \log h_i + (1-T_i)\log(1-h_i)\} \\
&= \sum_{A(x)\sim p(A(x)|T=1)} \log h_i + \sum_{A(x)\sim p(A(x)|T=0)} \log(1-h_i)
\end{aligned}$$

For each $X = x$, by setting the derivative $\frac{\partial}{\partial h_i}\mathcal{L}_A^h = 0$, we have

$$\hat{h}_{A\to T}(A(x)) = \frac{p(A(x)|t=1)}{p(A(x)|t=1) + p(A(x)|t=0)}.$$

Substituting $\hat{h}_{A\to T}(A(x))$ into L_A, we have

$$\begin{aligned}
L_A &= -\Big\{ \sum_{A(x)\sim p(A(x)|T=1)} \log \frac{p(A(x)|T=1)}{p(A(x)|T=1) + p(A(x)|T=0)} + \\
&\qquad \sum_{A(x)\sim p(A(x)|T=0)} \log \frac{p(A(x)|T=0)}{p(A(x)|T=1) + p(A(x)|T=0)} \Big\} \\
&= \log 4 - D_{KL}\big(p(A(x)|T=1) || \frac{p(A(x)|T=1) + p(A(x)|T=0)}{2}\big) \\
&\qquad - D_{KL}\big(p(A(x)|T=0) || \frac{p(A(x)|T=1) + p(A(x)|T=0)}{2}\big) \\
&\le \log 4.
\end{aligned}$$

Meanwhile, note that $L_A = \log 4$ when $p(A(x)|T=1) = p(A(x)|T=0)$. We may conclude that L_A is maximized when $A(x) \perp T$.

Secondly, when T is numeric, $l(\cdot,\cdot)$ denotes the l_2 loss, and

$$\hat{h}_{A\to T}(A(x)) = \mathbb{E}[T|A(x)].$$

Substituting $\hat{h}_{A\to T}(A(x))$ into L_A, we have

$$L_A = \mathbb{E}[T - \mathbb{E}[T|A(x)]]^2 = Var\big[\mathbb{E}(T|A(X))\big]$$

Note that

$$Var(T) = Var\big[\mathbb{E}(T|A(X))\big] + \mathbb{E}\big[Var(T|A(X))\big],$$

we have $L_A <= Var(T)$.

Meanwhile, note that when $T \perp A(X)$, we have $\mathbb{E}(T|A(X)) = \mathbb{E}(T)$, and hence

$$L_A = \mathbb{E}[T - \mathbb{E}[T]]^2 = Var(T).$$

Thus, L_A is maximized when $T \perp A(X)$.

The proof for (ii) follows the similar way. Firstly, when Y is binary, we have

$$\begin{aligned}
\hat{h}_{C\cup T\to Y}(C(x),t) &= \frac{p(C(x),t|Y=1)}{p(C(x),t|Y=1)+p(C(x),t|Y=1)}. \\
\hat{h}_{I\cup C\cup T\to Y}(I(x),C(x),t) &= \frac{p(I(x),C(x),t|Y=1)}{p(I(x),C(x),t|Y=1)+p(I(x),C(x),t|Y=1)}.
\end{aligned}$$

Note that the KL divergence $D_{KL}(\hat{h}_{C\cup T\to Y}(C(x),t)||\hat{h}_{I\cup C\cup T\to Y}(I(x),C(x),t)) \geq 0$, and $\hat{h}_{C\cup T\to Y}(C(x),t) \equiv \hat{h}_{I\cup C\cup T\to Y}(I(x),C(x),t)$ when $p(Y|I(X),C(X),T) = p(Y|C(X),T)$, we have $L_{I,C}$ is minimized when $p(Y|I(X),C(X),T) = p(Y|C(X),T)$, i.e., $Y \perp I(X)|C(X),T$

When Y is numerical, we have

$$\begin{aligned}
\hat{h}_{C\cup T\to Y}(C(x),t) &= \mathbb{E}[Y|C(x),t]. \\
\hat{h}_{I\cup C\cup T\to Y}(I(x),C(x),t) &= \mathbb{E}[Y|I(x),C(x),t].
\end{aligned}$$

Therefore, we have $\hat{h}_{C\cup T\to Y}(C(x),t) = \hat{h}_{I\cup C\cup T\to Y}(I(x),C(x),t)$ when $Y \perp I(X)|C(X),T$, and $L_{I,C} = 0$ in this case. To summarize, $L_{I,C}$ is minimized when $Y \perp I(X)|C(X),T$. $\square$

Supplementary for source code

- `module_DER_extended.py` includes the model for both ADR and DeR-CFR;
- `module_DR.py` includes the model for DR-CFR;
- `train.py` is the script for one-time run by inputting the hyper-parameters in the command line.
- `run.py` is the script for multiple runs by setting a list of parameters in `.json` file. The hyper-parameters setting in our paper can be found in `./configs/params_all.json`.

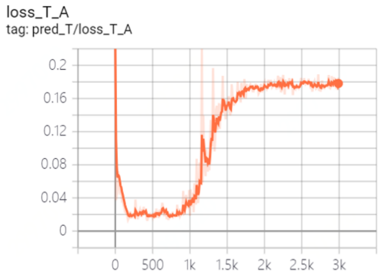
selection of hyper parameters

The hyper-parameters mainly involve the $\{\alpha, \beta, \mu, \lambda\}$ and K (Sec 4.3). .

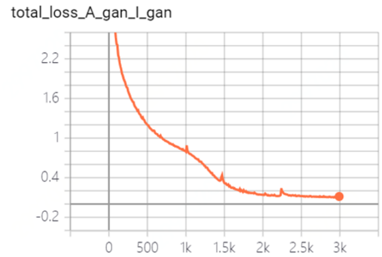
- $\alpha = \beta = 1$ by default when are all cross-entropy loss since the gradients are of the same scale. In the continuous case, we may need to adjust α and β because T and Y may not have a similar scale; The default parameter μ for the orthogonal loss is 10 and we need to adjust by observing the scale of $\mathcal{L}_O$;
- The default parameter of λ is 10^{-3} as the regularization term is commonly much larger;
- As for K, the number of iterations to train the auxiliary predictors h_*'s, we commonly take $K = 1, 2, 3$.

In practice, we suggest to use tensorboard or similar tools to record the details of the loss functions (including each component) and adjust the hyper-parameters to make the parameters convergent and loss become steady.

learning curve



(a) the curve of $\mathcal{L}_A^h$



(b) the curve of $\mathcal{L}$

References

[1] Judea Pearl. *Causality*. Cambridge University Press, 2009.