9.1 Proof of Theorem 4.1

The proof directly follows the following two lemmas. We denote the set of probabilistic functions of $\mathbf{X}$ into an arbitrary target space as $\mathcal{F}(\mathbf{X})$, and as $\mathcal{S}(Y)$ the set of sufficient statistics for Y. Since $\mathbf{h}(\cdot)$ is a function of combination , we have $I(Y; \mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}) = I(Y; \mathbf{X})$

Lemma 9.1. *Let $\mathbf{Z}$ be a probabilistic function of $\mathbf{X}$ and $\mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y})$, where $\mathbf{h}$ is function of combination. Then $\mathbf{Z}$ is a sufficient causes for Y if and only if*

$$I(Y; \mathbf{Z}) = \max_{\mathbf{Z}' \in \mathcal{F}(\mathbf{X})} I(Y, \mathbf{Z}')$$

Proof. Lemma is an extension of Lemma 12 in [33]. The differences lie on that $\mathbf{X}$ is consist of $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$ and we focus on the sufficient causes defined in Definition 2. Firstly for the sufficient condition, for every $\mathbf{Z}'$ which is a probabilistic function of $\mathbf{X}$, we have Markov Chain $Y - \mathbf{X} - \mathbf{Z}'$, so from data processing inequality [9] we have $I(Y; \mathbf{X}) \geq I(Y; \mathbf{Z}')$. Therefore we have $I(Y; \mathbf{Z}) = \max_{\mathbf{Z}' \in \mathcal{F}(\mathbf{X})} I(Y, \mathbf{Z}')$. We also have Markov Chain $Y - \mathbf{Z} - \mathbf{X}$, for the data processing inequality we have $I(Y; \mathbf{X}) \leq I(Y; \mathbf{Z})$. Thus $I(Y; \mathbf{Z}) = \max_{\mathbf{Z}' \in \mathcal{F}(\mathbf{X})} I(Y, \mathbf{Z}')$ is held.

Then for necessary condition, assume that we have a Markov Chain $Y - \mathbf{X} - \mathbf{Z}$. According to data processing inequality, the $I(Y; \mathbf{Z}) = I(Y; \mathbf{X})$ holds if and only if $I(Y; \mathbf{X}|\mathbf{Z}) = 0$. Since $\mathbf{X}$ is consist of $\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}$, the $I(Y; \mathbf{pa_Y}, \mathbf{nd_Y}|\mathbf{Z}) = 0$. in other word, Y and $\mathbf{pa_Y}, \mathbf{nd_Y}$ are conditionally independent by $\mathbf{Z}$, hence $\mathbf{Z}$ is a sufficient causes satisfied Definition 2.

□

Lemma 9.2. *Let $\mathbf{Z}$ be sufficient statistics of Y and $\mathbf{X} = \mathbf{h}(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y})$, where $\mathbf{h}$ is function of combination. Then $\mathbf{Z}$ is minimal sufficient causes for $\mathbf{Y}$ if and only if*

$$I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) = \min_{\mathbf{Z}' \in \mathcal{S}(Y)} I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}') \tag{19}$$

Proof. Firstly, for the sufficient condition, let $\mathbf{Z}$ be a minimal sufficient causes, and $\mathbf{Z}'$ be some sufficient causes. Because there is a function $\mathbf{Z} = f(\mathbf{Z}')$ from Definition 3, it has Markov Chain $(\mathbf{nd_Y}, \mathbf{pa_Y}) - Y - \mathbf{Z}' - \mathbf{Z}$, and we get $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) \leq I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}')$. So that $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) = \min_{Z' \in \mathcal{S}(Y)} I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}')$ holds.

For the necessary condition, assume that $\mathbf{Z}$ is not minimal, then there exist another sufficient statistics $\mathbf{V}$ allows $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) > I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V})$ and let $\mathbf{V} : \mathcal{X} \to \mathcal{Z}$ is a function of $\mathbf{X}$ such that $\forall \mathbf{x}, \quad \mathbf{V}(\mathbf{x}) \in \{\mathbf{z} \mid \mathbf{z} \sim \mathbf{Z}(\mathbf{x})\}$. Inspired by Fisher–Neyman factorization theorem [10], we can factorize $p(\mathbf{x})$ as below

$$\forall \mathbf{x}, y \quad p(\mathbf{x} \mid y) = l^3_{\mathbf{Z}}(\mathbf{dc_Y}|\mathbf{pa_Y}) l^1_{\mathbf{Z}}(\mathbf{pa_Y}, \mathbf{nd_Y}) l^2_{\mathbf{Z}}(\mathbf{Z}(\mathbf{x}), y) \tag{20}$$

In above equation since $\mathbf{dc_Y}$ is decided by $\mathbf{pa_Y}$ and Y. We can drop the $\mathbf{dc_Y}$ by defining $l^3_{\mathbf{Z}}(\mathbf{dc_Y}|\mathbf{pa_Y}) \triangleq l^3_{\mathbf{V}}(\mathbf{dc_Y}|\mathbf{pa_Y})$ and we can rewrite the sufficient causes condition as below

$$\forall \mathbf{pa_Y}, \mathbf{nd_Y}, y \quad p(\mathbf{pa_Y}, \mathbf{nd_Y} \mid y) = l^1_{\mathbf{Z}}(\mathbf{pa_Y}, \mathbf{nd_Y}) l^2_{\mathbf{Z}}(\mathbf{Z}(\mathbf{x}), y) \tag{21}$$

We define a equivalence relation $\sim$ by

$$\mathbf{z_1} \sim \mathbf{z_2} \Longleftrightarrow \frac{l^2_{\mathbf{Z}}(\mathbf{z_1}, y)}{l^2_{\mathbf{Z}}(\mathbf{z_2}, y)} \text{is a constant function of } Y \tag{22}$$

There exists a sufficient cause $\mathbf{Z}'$ such that $\mathbf{Z}$ is not a function of $\mathbf{Z}'$. The following process proof that $\mathbf{V}$ is also sufficient cause of Y:

$$l^1_{\mathbf{V}}(\mathbf{pa_Y}, \mathbf{nd_Y}) \triangleq l^1_{\mathbf{Z}}(\mathbf{pa_Y}, \mathbf{nd_Y}) \frac{l^2_{\mathbf{Z}}(\mathbf{Z}(\mathbf{x}), y)}{l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x}), y)}$$
$$l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x}), y) \triangleq l^2_{\mathbf{V}}(\mathbf{V}(\mathbf{x}), y)$$

Then

$$\begin{aligned} p(\mathbf{pa_Y}, \mathbf{nd_Y} \mid y) &= l^1_{\mathbf{Z}}(\mathbf{pa_Y}, \mathbf{nd_Y}) l^2_{\mathbf{Z}}(\mathbf{Z}(\mathbf{x}), y) \\ &= l^1_{\mathbf{Z}}(\mathbf{pa_Y}, \mathbf{nd_Y}) \frac{l^2_{\mathbf{Z}}(\mathbf{Z}(\mathbf{x}), y)}{l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x}), y)} l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x}), y) \\ &= l^1_{\mathbf{V}}(\mathbf{pa_Y}, \mathbf{nd_Y}) l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x}), y) \end{aligned}$$

Since above equation holds, $\mathbf{V}$ have factorization formulation of sufficient statistics, $\mathbf{V}$ is also a sufficient statistic. Let $\mathbf{x_1}, \mathbf{x_2}$ such that $\mathbf{Z}'(\mathbf{x_1}) = \mathbf{Z}'(\mathbf{x_2})$, then $l^2_{\mathbf{Z}'}(\mathbf{Z}'(\mathbf{x_1}), y) = l^2_{\mathbf{Z}'}(\mathbf{Z}'(\mathbf{x_2}), y)$

$$\begin{aligned} \frac{l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x_1}), y)}{l^2_{\mathbf{Z}}(\mathbf{V}(\mathbf{x_2}), y))} &= \frac{p(\mathbf{x_1} \mid y) l^1_{\mathbf{Z}}(\mathbf{x_2})}{p(\mathbf{x_2} \mid y) l^1_{\mathbf{Z}}(\mathbf{x_1})} \\ &= \frac{l^1_{\mathbf{Z}'}(\mathbf{x_1}) l^2_{\mathbf{Z}'}(\mathbf{Z}'(\mathbf{x_1}), y) l^1_{\mathbf{Z}}(\mathbf{x_2})}{l^1_{\mathbf{Z}}(\mathbf{x_1}) l^2_{\mathbf{Z}'}(\mathbf{Z}'(\mathbf{x_2}), y) l^1_{\mathbf{Z}'}(\mathbf{x_2})} \\ &= \frac{l^1_{\mathbf{Z}'}(\mathbf{x_1}) l^1_{\mathbf{Z}}(\mathbf{x_2})}{l^1_{\mathbf{Z}}(\mathbf{x_1}) l^1_{\mathbf{Z}'}(\mathbf{x_2})} \end{aligned}$$

From the above equation we can get $\mathbf{Z}(\mathbf{x_1}) = \mathbf{Z}(\mathbf{x_2})$, then we have $\mathbf{V}(\mathbf{x_1}) = \mathbf{V}(\mathbf{x_2})$ because $\mathbf{V}$ is function of $\mathbf{Z}'$. Since $\mathbf{Z}$ is sufficient cause of Y, and $\mathbf{pa_Y}, \mathbf{nd_Y} \perp Y|\mathbf{Z}$ in Definition 1 holds. There exists Markov Chains $\mathbf{X} - \mathbf{Z} - \mathbf{V}$ and $(\mathbf{pa_Y}, \mathbf{nd_Y}) - \mathbf{Z} - \mathbf{V}$. From data processing inequality, $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) \geq I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V})$. The term $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z})$ can be decomposed as below

$$\begin{aligned} I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) =& I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V}) + I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}|\mathbf{V}) \\ \geq& I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V}) + I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}|\mathbf{Z}', \mathbf{V}) \\ =& I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V}) + I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}|\mathbf{Z}') \end{aligned} \tag{23}$$

Since $\mathbf{Z}'$ is not the function of $\mathbf{Z}$, thus $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}|\mathbf{Z}') > 0$, therefore we have $I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{Z}) > I(\mathbf{nd_Y}, \mathbf{pa_Y}; \mathbf{V})$. Thus Eq. 19 does not hold if $\mathbf{Z}$ is not minimal. The proof completes.

□

9.2 Proof of Proposition 4.3

Proof. Under the assumption that $\mathbf{Z}$ block the path between $\mathbf{X}$ and $\mathbf{dc_Y}$, $\mathbf{X}$ and $\mathbf{dc_Y}$ are conditional independent by variable $\mathbf{Z}$. $\mathbf{X} = h(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{dc_Y}) = h(\mathbf{pa_Y}, \mathbf{nd_Y}, \mathbf{Z}) = h(\mathbf{pa_Y}, \mathbf{nd_Y})$. Since all generative function of factors are invertible, we can replace $(\mathbf{nd_Y}, \mathbf{dc_Y})$ in Markov Chain shown in the proof of Theorem 4.1 by variable $\mathbf{X}$. Therefore, $p(y|\mathbf{z}, \mathbf{pa_y}, \mathbf{nd_y}) = p(y|\mathbf{z})$ is held if and only if $p(y|\mathbf{z}, \mathbf{x}) = p(y|\mathbf{z})$ holds. Thus, under the the assumption that $\mathbf{Z}$ block the path between $\mathbf{X}$ and $\mathbf{dc_Y}$ and h is a linear invertible function, the optimization processes defined in Proposition 4.3 and Theorem 4.1 are equivalence. □

9.3 Proof of Theorem 6.1

The proof follows [33] Theorem 3. The sketch of proof contains two steps: (i) we decompose the original objective $|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})|$ into two parts. (ii) for each part, we deduce the deterministic finite sample bound by concentration of measure arguments on L2 norms of random vector. Let $H(X)$ denote the entropy of X, we have

$$|I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq |H(Y|\mathbf{Z}) - \hat{H}(Y|\mathbf{Z})| + |H(Y) - \hat{H}(Y)| \tag{24}$$

Let $\zeta(x)$ denote a continuous, monotonically increasing and concave function.

$$\zeta(x) = \begin{cases} 0 & x = 0 \\ x\log(1/x) & 0 < x \leq 1/\mathrm{e} \\ 1/\mathrm{e} & x > 1/\mathrm{e} \end{cases} \tag{25}$$

for the term $|H(Y|Z) - \hat{H}(Y|Z)|$

$$\begin{aligned} |H(Y|\mathbf{Z}) - \hat{H}(Y|\mathbf{Z})| &= \left|\sum_{\mathbf{z}} (p(\mathbf{z})H(Y \mid \mathbf{z}) - \hat{p}(\mathbf{z})\hat{H}(Y \mid \mathbf{z}))\right| \\ &\leq \left|\sum_{\mathbf{z}} p(\mathbf{z})(H(Y \mid \mathbf{z}) - \hat{H}(Y \mid \mathbf{z}))\right| + \left|\sum_{\mathbf{z}} (p(\mathbf{z}) - \hat{p}(\mathbf{z}))\hat{H}(Y \mid \mathbf{z})\right| \end{aligned} \tag{26}$$

For the first summand in this bound, we introduce variable $\boldsymbol{\epsilon}$ to help decompose $p(y|\mathbf{z})$, where $\boldsymbol{\epsilon}$ is independent with the parents $\mathbf{pa_y}$ (i.e. $\boldsymbol{\epsilon} \perp \mathbf{pa_y}$)

$$\begin{aligned} &\left|\sum_{\mathbf{z}} p(\mathbf{z})(H(Y \mid \mathbf{z}) - \hat{H}(Y \mid \mathbf{z}))\right| \\ &\leq \left|\sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} (\hat{p}(y \mid \mathbf{z}) \log(\hat{p}(y \mid \mathbf{z})) - p(y \mid \mathbf{z}) \log(p(y \mid \mathbf{z})))\right| \\ &\leq \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} \zeta(|\hat{p}(y \mid \mathbf{z}) - p(y \mid \mathbf{z})|) \\ &= \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} \zeta\left(\left|\sum_{\boldsymbol{\epsilon}} p(\boldsymbol{\epsilon} \mid \mathbf{z})(\hat{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - p(y \mid \mathbf{z}, \boldsymbol{\epsilon}))\right|\right) \\ &= \sum_{\mathbf{z}} p(\mathbf{z}) \sum_{y} \zeta(\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}) \end{aligned} \tag{27}$$

where $\frac{1}{m}V(x)$ denote the variance of vector x. For the second summand in Eq.26.

$$\left|\sum_{\mathbf{z}} (p(\mathbf{z}) - \hat{p}(z))\hat{H}(Y \mid z)\right| \leq \|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\| \cdot \sqrt{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))} \tag{28}$$

For the summand $|H(Y) - \hat{H}(Y)|$:

$$\begin{aligned} |H(Y) - \hat{H}(Y)| &= \left| \sum_y p(y)\log(p(y)) - \hat{p}(y)\log(\hat{p}(y)) \right| \\ &\leq \sum_y \zeta(|p(y) - \hat{p}(y)|) \\ &= \sum_y \zeta\left(\left| \sum_{\mathbf{z}} \sum_{\boldsymbol{\epsilon}} p(\boldsymbol{\epsilon} \mid \mathbf{z})(p(\mathbf{z})p(y|\boldsymbol{\epsilon}) - \hat{p}(\mathbf{z})p(y|\boldsymbol{\epsilon})) \right| \right) \\ &\leq \sum_y \zeta(\|\mathbf{p}(\mathbf{z})p(y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z})p(y|\boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}) \end{aligned} \tag{29}$$

Combining above bounds:

$$\begin{aligned} |I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq & \sum_y \zeta(\|\mathbf{p}(\mathbf{z}, y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z}, y|\boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}) \\ & + \sum_{\mathbf{z}} p(\mathbf{z}) \sum_y \zeta(\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \epsilon) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\| \sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid z))}) \\ & + \|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\| \cdot \sqrt{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))} \end{aligned} \tag{30}$$

Let $\boldsymbol{\rho}$ be a distribution vector of arbitrary cardinality, and let $\hat{\boldsymbol{\rho}}$ be an empirical estimation of $\boldsymbol{\rho}$ based on a sample of size m. Then the error $\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\|$ will be bounded with a probability of at least $1 - \delta$

$$\|\boldsymbol{\rho} - \hat{\boldsymbol{\rho}}\| \leq \frac{2 + \sqrt{2\log(1/\delta)}}{\sqrt{m}} \tag{31}$$

Following the proof of Theorem 3 in [33], to make sure the bounds hold over $|\mathcal{Y}| + 2|$ quantities, we replace δ in Eq.31 by $\delta/(|\mathcal{Y}| + 2|$, than substitute $\|\mathbf{p}(\mathbf{z}, y|\boldsymbol{\epsilon}) - \hat{\mathbf{p}}(\mathbf{z}, y|\boldsymbol{\epsilon})\|$ $\|\hat{\mathbf{p}}(y \mid \mathbf{z}, \boldsymbol{\epsilon}) - \mathbf{p}(y \mid \mathbf{z}, \boldsymbol{\epsilon})\|$, $\|\mathbf{p}(\mathbf{z}) - \hat{\mathbf{p}}(\mathbf{z})\|$, by Eq.31.

$$\begin{aligned} |I(Y;\mathbf{Z}) - \hat{I}(Y;\mathbf{Z})| \leq & (2 + \sqrt{2\log((|\mathcal{Y}| + 2)/\delta)}) \sqrt{\frac{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))}{m}} \\ & + 2|\mathcal{Y}| h\left(2 + \sqrt{2\log((|\mathcal{Y}| + 2)/\delta)} \sqrt{\frac{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}{m}} \right) \end{aligned} \tag{32}$$

There exist a constant C, where $2 + \sqrt{2\log((|\mathcal{Y}| + 2)/\delta)} \leq \sqrt{C\log((|\mathcal{Y}|)/\delta)}$. From the fact that variance of any random variable bounded in $[0, 1]$ is at most $1/4$, we analyze the bound under two different cases:

In general case ($\mathbf{z} = \phi(\mathbf{x})$ is arbitrary representation of $\mathbf{x}$),

$$V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z})) \leq \frac{|\mathcal{Z}|}{4} \tag{33}$$

let m denote the number of sample, we get a lower bound of m, which is also known as sample complexity.

$$m \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\mathcal{Z}|\mathrm{e}^2 \tag{34}$$

In ideal case($\mathbf{z}$ is sufficient cause of $\mathbf{x}$) in that case $\mathbf{z}$ is independent with the exogenous noise $\boldsymbol{\epsilon}$, $\mathbf{z} \perp \boldsymbol{\epsilon}$:

$$V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z})) \leq \beta \tag{35}$$

$$m \geq \frac{C}{4}\log(|\mathcal{Y}|/\delta)|\beta|\mathrm{e}^2 \tag{36}$$

$$\sqrt{\frac{C\log(|\mathcal{Y}|/\delta)V(\mathbf{p}(\epsilon \mid \mathbf{z}))}{m}} \leq \sqrt{\frac{C\log(|\mathcal{Y}|/\delta)|\mathcal{Z}|}{4m}} \leq 1/\mathrm{e} \tag{37}$$

Then, from the fact that ([33]):

$$\begin{aligned} h\left(\sqrt{\frac{v}{m}}\right) &= \left(\sqrt{\frac{v}{m}}\log\left(\sqrt{\frac{m}{v}}\right)\right) \\ &\leq \frac{\sqrt{v}\log(\sqrt{m}) + 1/\mathrm{e}}{\sqrt{m}}, \end{aligned} \tag{38}$$

We can get the upper bound of second summand in Eq.43 as follows

$$\begin{aligned}&\sum_{y} h\left(\sqrt{C\log((|\mathcal{Y}|)/\delta)}\sqrt{\frac{V(\mathbf{p}(\epsilon \mid z))}{m}}\right)\\ \leq &\frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}}\end{aligned} \tag{39}$$

In general case:

$$Eq.39 \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}} \tag{40}$$

In ideal case:

$$Eq.39 \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\log(m)\left(|\mathcal{Y}|\sqrt{\beta}\right)+\frac{2}{\mathrm{e}}|\mathcal{Y}|}{2\sqrt{m}} \tag{41}$$

For the first summand in Eq.43, we follow the fact ([33] Theorem 3) that:

$$V(\mathbf{H}(Y \mid \mathbf{z})) \leq \frac{|Z|\log^2(|\mathcal{Y}|)}{4} \tag{42}$$

Finally we accomplish the proof of Theorem 6.1.

9.4 Extension of Theorem 6.1 for distribution shift

Proof. The risk under target domain is defined as $|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})|$, the proof is start by the following equation shown in the proof of Theorem 6.1.

$$\begin{aligned}|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})| \leq &(2+\sqrt{2\log((|\mathcal{Y}|+2)/\delta)})\sqrt{\frac{V(\hat{\mathbf{H}}_{\mathcal{T}}(Y \mid \mathbf{z}))}{m}}\\ &+2|\mathcal{Y}|h\left(2+\sqrt{2\log((|\mathcal{Y}|+2)/\delta)}\sqrt{\frac{V(\mathbf{p}(\boldsymbol{\epsilon} \mid \mathbf{z}))}{m}}\right)\end{aligned} \tag{43}$$

We will then bound the term $V(\hat{\mathbf{H}}_{\mathcal{T}}(Y \mid \mathbf{z}))$ by the variance of entropy on source data. From the definition of function V, we have

$$\begin{aligned}\sqrt{V(\hat{\mathbf{H}}(Y \mid \mathbf{z}))} &\leq \sqrt{\sum_{y}(\hat{H}(Y \mid \mathbf{z}) - \hat{H}(Y))^2} + \sqrt{\sum_{y}\left(\hat{H}(Y) - \frac{1}{|\mathcal{Z}|}\sum_{z'}\hat{H}\left(Y \mid \mathbf{z}'\right)\right)^2}\\ &\leq \left(1+\frac{1}{\sqrt{|\mathcal{Z}|}}\right)\left|\sum_{z'}\left(\hat{H}(Y) - \hat{H}\left(Y \mid \mathbf{z}'\right)\right)\right|\\ &= \left(1+\frac{1}{\sqrt{|\mathcal{Z}|}}\right)\frac{1}{\min_{\mathbf{z}} p(\mathbf{z})}\left(\hat{H}(Y) - \sum_{\mathbf{z}} p(y)\hat{H}(Y \mid \mathbf{z})\right)\\ &= \left(1+\frac{1}{\sqrt{|\mathcal{Z}|}}\right)\frac{1}{\min_{\mathbf{z}} p(\mathbf{z})}\hat{I}(\mathbf{Z};Y)\end{aligned} \tag{44}$$

Supposing that only source data $\mathcal{S}(\mathbf{X}, Y)$ is available, for the term $\hat{I}_{\mathcal{T}}(\mathbf{Z};Y)$ evaluated on target dataset, we change the measure by important sampling and Jensen's inequality. The way helps bound $\hat{I}_{\mathcal{T}}(\mathbf{Z};Y)$ by the evaluation on source domain. Denoting $D_{KL}(P||Q)$ by the Kullback-Leibler divergence between distribution P and Q. $\mathcal{S}(\mathbf{z}, y)$ and $\mathcal{T}(\mathbf{z}, y)$ are the distribution of $p(\mathbf{z}, Y)$ on source domain and target domain separately.

$$\begin{aligned}\hat{I}_{\mathcal{T}}(\mathbf{Z}, Y) &= \mathbb{E}_{\mathcal{T}(\mathbf{z},y)}\log\frac{\mathbf{p}(\mathbf{z}, y)}{\mathbf{p}(\mathbf{z})\mathbf{p}(y)}\\ &\leq D_{KL}(\mathcal{T}(\mathbf{z}, y)||\mathcal{S}(\mathbf{z}, y)) + \log\mathbb{E}_{\mathcal{S}(\mathbf{z},y)}\frac{\hat{p}(\mathbf{z}, y)}{\hat{p}(\mathbf{z})\hat{p}(y)}\end{aligned} \tag{45}$$

Substituting $\hat{I}_{\mathcal{T}}(\mathbf{Z}, Y)$ into Eq. 44 and Eq. 43. Since $|\mathcal{Z}| > 1$, let $D = \frac{2}{\min_{\mathbf{z}} p(\mathbf{z})}$ and $I_{\mathcal{S}} = \mathbb{E}_{\mathcal{S}(\mathbf{z},y)}\frac{\hat{p}(\mathbf{z},y)}{\hat{p}(\mathbf{z})\hat{p}(y)}$. We can get the bounds under two different cases:

In general case, since since $\mathbf{Z}$ is arbitrary representation of $\mathbf{X}$, we get $D_{KL}(\mathcal{T}(\mathbf{z}, y)||\mathcal{S}(\mathbf{z}, y)) > 0$. We cannot drop the D_{KL} term. Thus we have the bound:

$$|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\mathcal{Z}|}\log(m) + D_{KL}(\mathcal{T}||\mathcal{S}) + DI_{\mathcal{S}}\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{46}$$

In ideal case, since $\mathbf{Z}$ is sufficient cause of Y, we get $D_{KL}(\mathcal{T}(\mathbf{z}, y)||\mathcal{S}(\mathbf{z}, y)) = 0$ from Assumption 6.2.

$$|I_{\mathcal{T}}(Y;\mathbf{Z}) - \hat{I}_{\mathcal{T}}(Y;\mathbf{Z})| \leq \frac{\sqrt{C\log(|\mathcal{Y}|/\delta)}\left(|\mathcal{Y}|\sqrt{|\beta|}\log(m) + DI_{\mathcal{S}}\right) + \frac{2}{e}|\mathcal{Y}|}{\sqrt{m}} \tag{47}$$

□

10 EXPERIMENTAL DETAILS

10.1 Model Architecture and Implementation Details

The hyper-parameters are determined by grid search. Specifically, the learning rate and batch size are tuned in the ranges of $[10^{-4}, 10^{-1}]$ and $[64, 128, 256, 512, 1024]$, respectively. The weighting parameter λ is tuned in $[0.001]$. Perturbation degrees are set to be $\beta = \{0.1, 0.2, 0.1, 0.3\}$ for Coat, Yahoo!R3, PCIC and CPC separately. The representation dimension is empirically set as 64. All the experiments are conducted based on a server with a 16-core CPU, 128g memories and an RTX 5000 GPU. The deep model architecture is shown as follows:

(1)Representation learning method $\phi(\mathbf{x})$: If dataset is Yahoo!R3 or PCIC, in which only user id and item id are the input, we firstly use an embedding layer. The representation function architecture is:

- Concat(Embedding(user id, 32), Embedding(item id, 32))
- Linear(64, 64), ELU()
- Linear(64, representation dim), ELU()

Then for the dataset Coat and CPC, the feature dimension is 29 and 47 separately. It do not use embedding layer at first. The representation function architecture is.

- Linear(64, 64), ELU()
- Linear(64, representation dim), ELU()

(2)Downstream Prediction Model $g(\mathbf{z})$:

- Linear(representation dim, 64), ELU()
- Linear(64, 2)

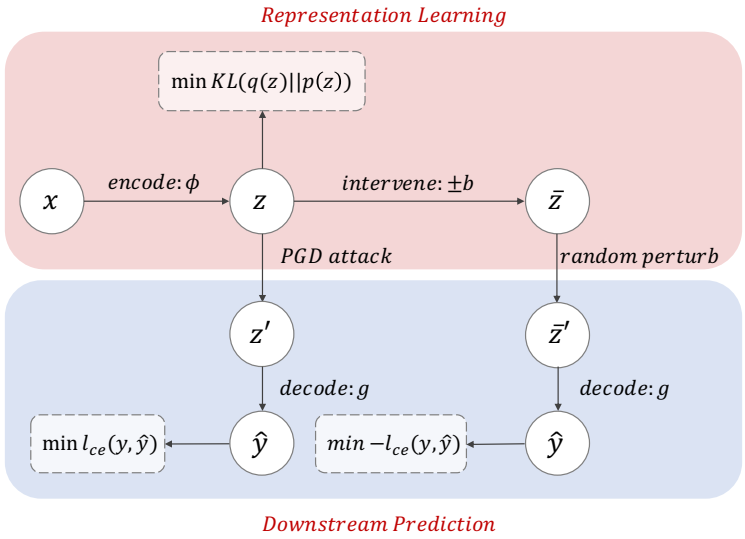

Figure 3: The figure demonstrates the model architecture of CaRI

The figure shows the model architecture.

11 ADDITIONAL RESULTS

Due to the page limit in main text, we present the additional test results and analysis in this section. Table 3 overall results with standard error via 5 tims runs. Fig. 5 and 6 compare the distance correlation metric given by the training under standard and robust mode. It shows that our method performs consistently better compared with base methods in both modes, with a higher distance correlation, under smaller variance. The gap is obvious especially in the learning of parental information, which is the main focus of our approaches.

Fig. 7 and 8 record the results along optimization process and until convergence, under different settings of the pertubation degree β, considering the dataset CelebA-anno. The annotation smile is used as the label to be prediced, and other features are the source data. It shows that when the optimization process is not finished, both approaches have similar performance, with unstability evidenced by large variance of the DC metric. However, our method outperforms the baseline when the optimization converges, owning a higher DC with smaller variations. The results also show that β is an important factor for training the model. Larger β often leads to higher variance of the training of the model.

Fig.4 demonstrates how robust training degree ($\beta = \{0.1, 0.3, 0.5, 0.7, 1.0\}$) influences the downstream prediction under adversarial settings. We conduct the experiments on the attacked real-world dataset by PGD attacker. From Fig.4, we find that our method is better than base method, because the base model's ability on standard prediction is broken by adversarial training. When β is small, our method behaves closely to the r-CVAE in all the datasets. When β gets larger, the difference between performance of CaRI and that of r-CVAE continuously enlarges in Yahoo!R3. In PCIC, the gap becomes the largest among all when $\beta = 0.5$, and narrows down to 0 when $\beta = 0.7$. This is because in our framework, we explicitly deploy a model to achieve more robust representations, while others fail.

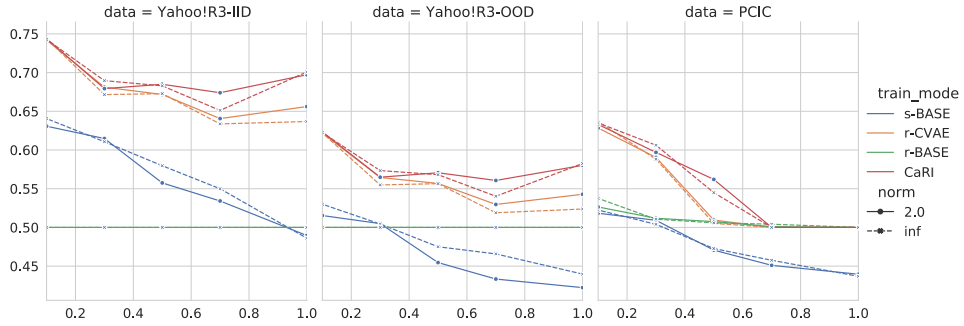

Figure 4: Results under different adversarial perturbations β on three datasets. Axis-x is the attack degree β. Axis-y is the adv-AUC under attacked test datasets.

Table 3: Additional overall results with standard error.

dataset				AUC	std	ACC	std	adv_AUC	std	adv_ACC	std
PCIC	standard	p=2	CaRI	0.6416	0.0078	0.6803	0.0014	0.619	0.004	0.6625	0.0041
			r-CVAE	0.6328	0.0023	0.6725	0.0042	0.5893	0.0419	0.6429	0.0201
		p=∞	CaRI	0.6447	0.0041	0.6817	0.0043	0.6148	0.011	0.664	0.0104
			r-CVAE	0.6358	0.014	0.6779	0.0066	0.6138	0.0062	0.6601	0.0048
	robust	p=2	CaRI	0.6363	0.0045	0.6709	0.0042	0.6332	0.0024	0.6576	0.0006
			r-CVAE	0.63	0.0075	0.674	0.0069	0.6187	0.0051	0.6493	0.0013
		p=∞	CaRI	0.639	0.007	0.6761	0.0024	0.6225	0.0057	0.6638	0.001
			r-CVAE	0.6363	0.0066	0.6733	0.0058	0.6088	0.0098	0.6596	0.0124
Yahoo!R3 OOD	standard	p=2	CaRI	0.6276	0.0001	0.6255	0.0022	0.5917	0.0071	0.5917	0.0072
			r-CVAE	0.6233	0.0005	0.6243	0.002	0.5865	0.0022	0.5872	0.0025
		p=∞	CaRI	0.629	0.0011	0.6257	0.0002	0.5966	0.0049	0.5965	0.0042
			r-CVAE	0.6253	0.0023	0.6249	0.0014	0.5855	0.0016	0.5863	0.0019
	robust	p=2	CaRI	0.6242	0.0009	0.6307	0.0012	0.6008	0.0009	0.601	0.0016
			r-CVAE	0.6191	0.0013	0.6241	0.0051	0.5882	0.0014	0.5907	0.0009
		p=∞	CaRI	0.6238	0.0011	0.6284	0.0017	0.5993	0.0019	0.5999	0.0026
			r-CVAE	0.6186	0.001	0.6235	0.0028	0.5886	0.0014	0.5912	0.0012
Yahoo!R3 i.i.d.	standard	p=2	CaRI	0.7493	0.0004	0.7495	0.0015	0.7188	0.0015	0.7072	0.0013
			r-CVAE	0.7487	0.0001	0.7529	0.0027	0.7202	0.0029	0.7099	0.0027
		p=∞	CaRI	0.7497	0.0004	0.7503	0.0019	0.7191	0.0023	0.7099	0.0026
			r-CVAE	0.7488	0.0001	0.7515	0.0008	0.7191	0.0021	0.7072	0.0015
	robust	p=2	CaRI	0.7374	0.0024	0.7158	0.0061	0.7247	0.0026	0.7159	0.0036
			r-CVAE	0.7376	0.0018	0.7151	0.0045	0.7194	0.0020	0.7082	0.0021
		p=∞	CaRI	0.7378	0.0015	0.7168	0.0015	0.7210	0.0031	0.7107	0.0040
			r-CVAE	0.7341	0.0007	0.7093	0.0035	0.7180	0.0017	0.7080	0.0016
Coat OOD	standard	p=2	CaRI	0.5725	0.0005	0.5732	0.0005	0.5608	0.0003	0.5601	0.0004
			r-CVAE	0.5671	0.0005	0.5649	0.0006	0.5586	0.0002	0.554	0.0001
		p=∞	CaRI	0.5705	0.0013	0.5718	0.0017	0.5643	0.0001	0.5659	0.0006
			r-CVAE	0.5656	0.0005	0.5643	0.0007	0.5527	0.0074	0.5478	0.0081
	robust	p=2	CaRI	0.5705	0.0015	0.5675	0.0015	0.5674	0.0002	0.565	0.0012
			r-CVAE	0.5634	0.0014	0.5591	0.0018	0.5572	0.0009	0.5522	0.0003
		p=∞	CaRI	0.5707	0.0017	0.5681	0.0024	0.5653	0.0019	0.5659	0.0011
			r-CVAE	0.5629	0.0017	0.5586	0.0028	0.559	0.0004	0.5544	0.0007
Coat i.i.d.	standard	p=2	CaRI	0.7248	0.0011	0.7305	0.0016	0.7069	0.0023	0.7125	0.0036
			r-CVAE	0.7129	0.0009	0.7206	0.0022	0.7023	0.0041	0.7059	0.0061
		p=∞	CaRI	0.7283	0.0013	0.7355	0.0015	0.7125	0.0007	0.7196	0.001
			r-CVAE	0.7106	0.0029	0.7184	0.0033	0.7029	0.0008	0.7106	0.0094
	robust	p=2	CaRI	0.7265	0.0032	0.7331	0.0027	0.7196	0.0046	0.7261	0.0042
			r-CVAE	0.7087	0.0005	0.7169	0.0016	0.7058	0.002	0.7141	0.0036
		p=∞	CaRI	0.7276	0.0028	0.7339	0.002	0.7208	0.0023	0.727	0.0019
			r-CVAE	0.7147	0.0023	0.7222	0.0026	0.7105	0.0039	0.7181	0.0043

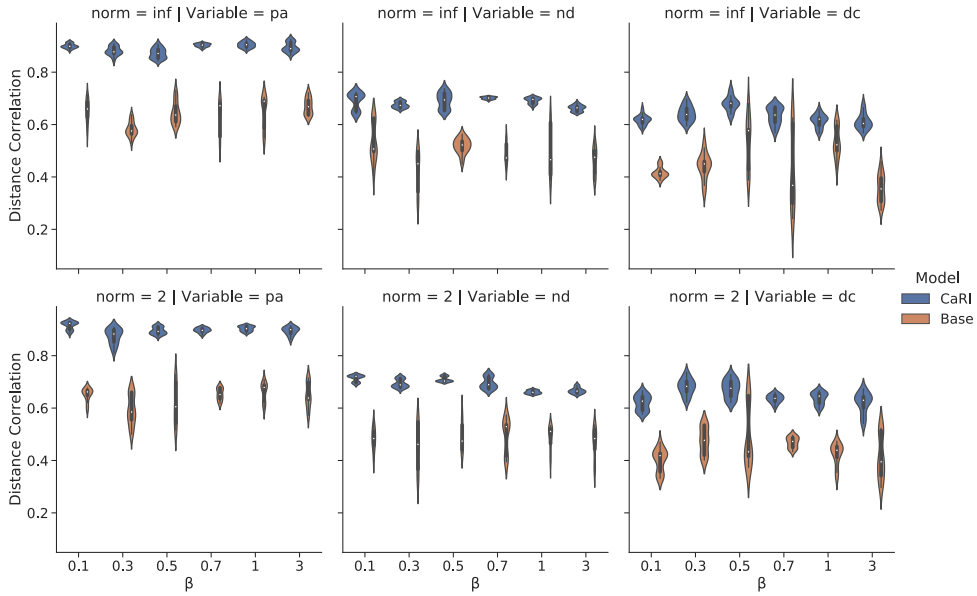

Figure 5: Identify results on synthetic dataset over different range of β under robust training.

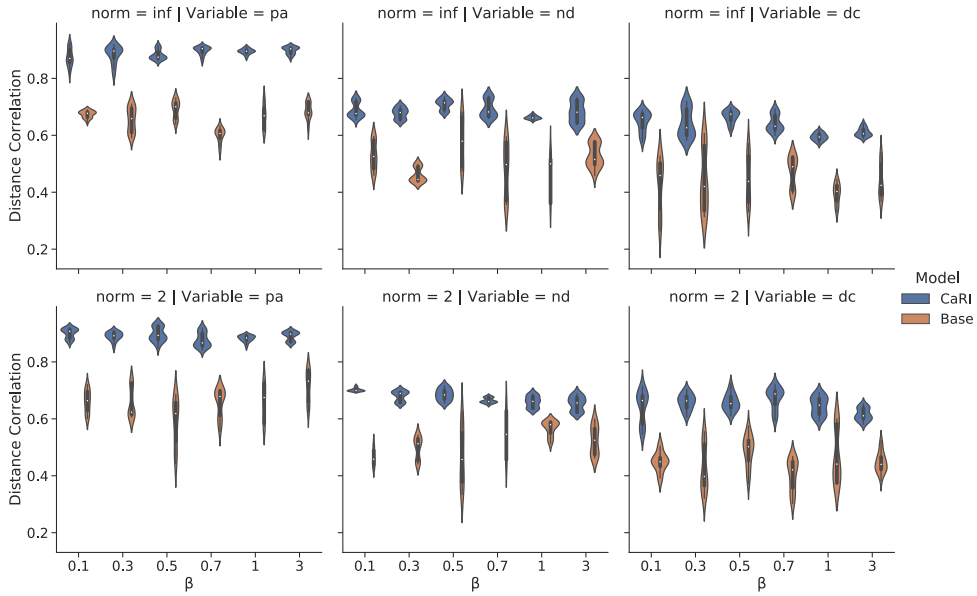

Figure 6: Identify results on synthetic dataset over different range of β under standard training.

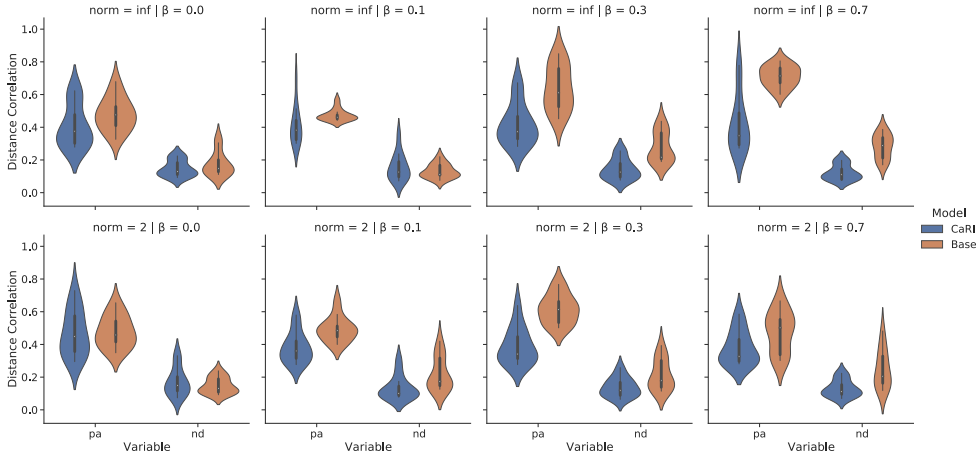

Figure 7: Identify results on CelebA-anno dataset over different range of β during early optimization step.

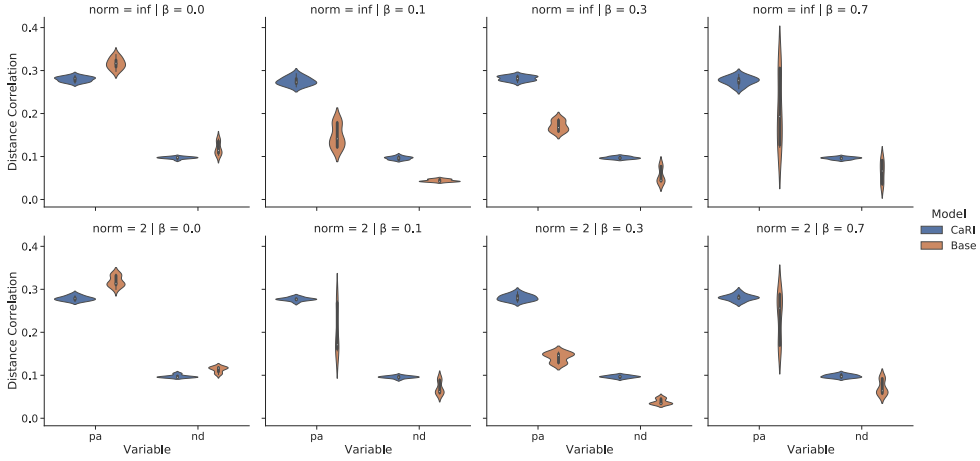

Figure 8: Identify results on CelebA-anno dataset over different range of β after converging.