Understanding Shortcut Learning through the Lens of Causality & Robustness

Anonymous

July 20, 2022

Despite tremendous successes, modern machine learning models oftentimes fail to generalize for samples out of distributions where the models are trained. Such failure has been reported as *shortcut learning* (Geirhos et al., 2020), a phenomenon that ML models fail to generalize due to taking unintended features in establishing their decision rules. Notwithstanding that the shortcut learning problem is prevalent in practice, virtually no formal/unified understandings of notions of shortcut learning problems and approaches for addressing the biases have been presented. In this document, we provide an understanding of shortcut learning and present two common approaches for addressing the biases under the rubric of formal causal languages. Finally, we relate the approaches to the causal invariance property. We hope this document will pave the way toward a unified understanding of shortcut learning problems.

Contents

1. Introduction

Consider a task of classifying an object in the set of images in Fig. 1. Specifically, tasks of interest are classifying the objects [boat] and [car] in images in Figs. (1a, 1b, 1c). Obviously, such tasks are not difficult for humans. However, as reported in many cases as in (Geirhos et al., 2020), if a ML model

(a) [*boat*] in the [water]

(b) [*car*] in the [road]

(c) [*boat*] in the [road]

Figure 1: Classification Tasks.

is trained mostly based on images like [boat in the water] as in Fig. 1a and [car in the road] as in Fig. 1c, the model may wrongly classify [boat in the road] in Fig. 1c as "car" by taking unintended features ([water] or [road]) importantly. Such undesirable learning behaviors and induced bias are called '*shortcut learning*':

Definition 1 ((Informal) **shortcut learning** (Geirhos et al., 2020))**.** A shortcut learning is a phenomenon that a ML model fails to be generalized due to taking unintended features, called *shortcut* (e.g., background objects), in establishing decision rules.

In this document, we will understand the notion of shortcut learning and the remedy under the rubric of causality. Specifically,

1. We will formally define the learning problem under the risk of shortcut learning using *structural causal models* (Pearl, 2000).
2. We will formalize two approaches for preventing shortcut learning. Specifically, we will show that natural approaches for shortcut removal will reduce to the ML model trained only using *causal features*, a set of features that *causes* a true label.
3. We will relate the causal features to the invariant features, a set of features whose relations with the label is invariant. Specifically, we will formalize that invariant features are causal features in high probability.

2. Problem Setup

2.1. Structural Causal Model

We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl, 2000). A structural causal model (SCM) is a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{u}) \rangle$ where $\mathbf{V}, \mathbf{U}$ are a sets of endogenous (observerables) and exogenous variables (latents), $\mathbf{F}$ is a set of functions f_{V_i} one for each $V_i \in V$ where $V_i \leftarrow f_{V_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq V$ and $U_{V_i} \subseteq U$, and $P(\mathbf{u})$ is a strictly positive probability measure for $\mathbf{U}$. Each SCM M induces a *semi-Markovian causal graph* G over the node set $\mathbf{V}$ here $V_i \to V_j$ if V_i is an argument of f_{V_j}, and $V_i \leftrightarrow V_j$ if U_{V_i} and U_{V_j} are correlated. Performing an intervention $\mathbf{X} = \mathbf{x}$ is represented through the *do*-operator, $do(\mathbf{X} = \mathbf{x})$ (shortly, $do(\mathbf{x})$), which encodes the operation of replacing the original equations of $\mathbf{X}$ by the constant $\mathbf{x}$ in the SCM M, inducing a submodel $M_{\mathbf{x}}$ and an interventional distribution $P_{\mathcal{M}}(\mathbf{V} = \mathbf{v}|do(\mathbf{x}))$ (shortly, $P(\mathbf{v}|do(\mathbf{x}))$).

2.2. Problem Setup

We assume that an SCM $\mathcal{M}$ is a data generating process for a set of variables $\mathbf{V}$ and a true label Y. For example, we assume that there are possibly unknown functions that generate the target objects (e.g., [boat]) and background objects (e.g., [water]).

We assume that there is a set of submodels $\mathcal{E} = \mathcal{E}(\{\mathbf{v}_i\}_{i=1}^N) \coloneqq \{\mathcal{M}_{\mathbf{v}_i}\}_{i=1}^N \cup \{\mathcal{M}\}$ induced by intervening $do(\mathbf{V}_i = \mathbf{v}_i)$, where each $\mathcal{M}_{\mathbf{v}_i} \in \mathcal{E}$ generates samples that a ML model could neglect in training. We will call $\mathcal{E}$, a set of data generating processes, as an *environment*. We permit that the intervened sets $\mathbf{V}_i$ are possible unknown. For example, we assumed that the image of [boat in the water] in Fig. 1a is generated by some SCM $\mathcal{M}$, and in contrast, [boat in the road] in Fig. 1c is generated by some SCM $\mathcal{M}_{\mathbf{v}_i}$ where $\mathbf{V}_i$ is a set of features corresponding to the background, and $\mathbf{v}_i$ is a set of realized values corresponding to [road]. If the ML model neglects the sample [boat in the road] in training, then the model could fall into the shortcut learning pitfalls.

Presented nomenclatures provide informational interpretation of the shortcut learning in Def. 1 – a phenomenon when a ML model doesn't perform well in some submodels in $\mathcal{E}$. Then, our task is to design a ML model that works well in all environments. Specifically,

Task 1. *Assume that the data generating process is an environment set $\mathcal{E}$. Our task is to construct the performant ML model for predicting the true label Y for all environment in $\mathcal{E}$ from samples generated by submodels in $\mathcal{E}$.*

3. Preventing shortcut learning

3.1. Two Approaches for Preventing shortcut learning

We present two approaches to achieve Task 1.

Approach 1. Our first approach is motivated by the fact that human's classification rule is robust to the shortcut learning. For example, we note that classifying [boat] and [car] in Fig. 1 is an easy classification even for babies. Specifically, humans will classify the [boat] objects in Figs. (1a, 1c) successfully regardless of their different background objects (e.g., [water], [road]), because humans will only use the [boat] object to produce a label (*human label* or *true label*) "boat". We will call such objects that *causes* a human to produce the true labels as *causal features*:

Unlike humans, a ML model trained from images of (1) [boat] in the [water] as in Fig. 1a and (2) [car] in the [road] as in Fig. 1b may fail for correctly classifying Fig. 1c. Such failures happen because, unlike humans, ML models take *unintended* or *causally-irrelevant* features (e.g., background features like [water] and [road]) other than causal features (e.g., [boat] and [car]) in establishing their decision rules. This observation motivates a learning approach that avoids to take non-causal features as much as possible:

Principle 1 (**Avoidance of causally-irrelevant features**). *To prevent shortcut learning, the ML model must be designed without using causally-irrelevant features as much as possible.*

Approach 2. Our second approach is motivated by the fact that a ML model working well for all environment is a robust model to the shortcut learning, since the bias occurs when there exists environments that the model could fail. For example, if the model works well in classifying [boats] in different

environments [water] or [road], then we can say that the model is robust to the shortcut learning. This observation motivates a learning approach that design a ML model working well for all environments:

Principle 2 (**Performant ML models for all environments**). *To prevent shortcut learning, the ML model must work well for all possible environments.*

We will show that two approaches commonly encourage that the ML predictors must be trained only using a set of features that directly causes the true label. Throughout this section, we assume the existence of causal graphs G induced by an SCM $\mathcal{M}$.

3.2. Approach 1. Avoidance of Causally-Irrelevant Features

We first provide the formal definition for the notion of causal irrelevance. A causal irrelevance between a pair of sets of variables in $\mathbf{V}$ is invariance of one set of variables even if other sets of variables are changed. Formally,

Definition 2 (**Causal Irrelevance** (Pearl, 2000, Def. 7.3.7)). For disjoint sets $(\mathbf{X}, \mathbf{Y}, \mathbf{W}) \subseteq \mathbf{V}$, a set of variables $\mathbf{X}$ is said to be *causally irrelevant* to $\mathbf{Y}$ given $\mathbf{W}$ if

$$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{w})) = P(\mathbf{y}|do(\mathbf{x}'), do(\mathbf{w})),$$

for all possible realizations of $\mathbf{Y}, \mathbf{X}, \mathbf{W}$ denoted as $\mathbf{y}, \mathbf{x}, \mathbf{w}$ and $\mathbf{x}'$, a realization of $\mathbf{X}$ s.t. $\mathbf{x} \neq \mathbf{x}'$.

Remark 1 (**Causal Irrelevance Set**). *We will say that* $\mathbf{X}$ *is causally irrelevant set to* $\mathbf{Y}$ *iff*

$$P(\mathbf{y}|do(\mathbf{x}), do(\mathbf{v}\backslash\mathbf{x})) = P(\mathbf{y}|do(\mathbf{x}'), do(\mathbf{v}\backslash\mathbf{x})).$$

For the example of Fig. 1, we note that the background label $B \in \{$[water], [road]$\}$ is causally irrelevant to the true label Y ("boat") given the target object $T \in \{$[boat, car]$\}$ since perturbing the data generating process to generate an perturbed samples like [boat in the road] doesn't affect the label.

Under the assumption that the data generating process is an SCM, the meaning of the causal irrelevance becomes inferrentially clearer:

Lemma 1 (**Graphical Interpretation of Causal Irrelevance**). *Let* G *denote the graph induced by* $\mathcal{M}$. *For* $(\mathbf{X}, \mathbf{Y}, \mathbf{W}) \subseteq \mathbf{V}$, *a set of variables* $\mathbf{X}$ *is causally irrelevant to* $\mathbf{Y}$ *given* $\mathbf{W}$ *if*

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{W})_{G_{\overline{\mathbf{X},\mathbf{W}}}},$$

where $G_{\overline{\mathbf{X},\mathbf{W}}}$ *is a graph induced from* G *by cutting all incoming edges to the variables in* $\mathbf{X}, \mathbf{W}$. *Also,* $\mathbf{X}$ *is causally irrelevant set to* $\mathbf{Y}$ *iff*

$$(\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{V}\backslash\mathbf{X})_{G_{\overline{\mathbf{V}}}}.$$

Equipped with the notion of causally irrelevance, we now formalize our learning strategy in Principle 1. We translate Principle 1, avoiding the causally irrelevant variables as much as possible, as a strategy of identifying the largest sets of causally irrelevant variables and ruling out these variables when training ML models.

A following result presents that a ML model agreeing with Principle 1 can be derived by learning the model with a parental set of the true label, denoted PA_Y. Formally,

Theorem 1 (**Identification of the Largest Causally Irrelevant Sets**). *Let PA_Y denote the parents of the true label Y in G. Then, the largest sets of variables that is causally irrelevant features to Y is $\mathbf{V}\backslash PA_Y$. Formally, $\mathbf{V}\backslash PA_Y$ is the unique solution of the following optimization problem*

$$\max_{\mathbf{X}\subseteq\mathbf{V}} |\mathbf{X}| \;\; \textit{subject to } P(\mathbf{y}|do(\mathbf{v}\backslash\mathbf{x}), do(\mathbf{x})) = P(\mathbf{y}|do(\mathbf{v}\backslash\mathbf{x}), do(\mathbf{x}')),$$

for any arbitrary realizations $\mathbf{y}, \mathbf{v}\backslash\mathbf{x}, \mathbf{x}, \mathbf{x}'$ s.t. $\mathbf{x} \neq \mathbf{x}'$.

Obviously, the only remaining variables after ruling out the largest causally irrelevant sets $\mathbf{V}\backslash PA_Y$ is PA_Y, a parental set of the true label Y. Therefore, Thm. 1 engenders the following principles, which particularizes Principle 1:

Principle 3 (**ML models with causal predictors**). *To prevent shortcut learning, the ML model must be trained with a set of variables PA_Y, a causal predictor for Y.*

3.3. Approach 2. Performant ML models for all environments

We translate the Principle 2 as the problem of finding the best ML predictor even in the most perturbed example ("worst environment"). Then, the learning problem can be rewritten as a problem of finding the solution function for the following problem:

$$\min_{f\in\mathcal{F}} \max_{P\in\mathcal{P}(\mathcal{E})} \mathbb{E}_P \left[\ell(Y, f(\mathbf{Y}))\right], \tag{1}$$

where $\mathcal{F}$ is a ML model class, $\mathcal{P}(\mathcal{E})$ is a set of distributions induced by submodels in $\mathcal{E}$, and $\ell(Y, f(\mathbf{V}))$ is a predefined loss function of $f(\mathbf{V})$. Then, following result formalizes that one of the solution function is the ML predictor that is only dependent on the causal predictors PA_Y:

Theorem 2 (**The ML models learned with causal features work well for all environments – *Regression*** (Rojas-Carulla et al., 2018, Theorem 4)). *Let $f_0(\mathbf{V}) \coloneqq \mathbb{E}\left[Y|PA_Y\right]$. Then,*

$$f_0 \in \arg\min_{f\in\mathcal{C}_0} \max_{Q\in\mathcal{P}(\mathcal{E})} \mathbb{E}_Q \left[\ell(Y, f(\mathbf{V}))\right], \tag{2}$$

where $\mathcal{P}(\mathcal{E})$ is a set of distributions induced by a set of environments $\mathcal{E}$, and $\mathcal{C}_0$ is a set of continuous functions.

Theorem 3 (**The ML models learned with causal features work well for all environments – *Classification.***). *Suppose Y is a discrete variable. $f_0(\mathbf{V}) \coloneqq \arg\max_y P(Y = y|PA_Y)$. Then,*

$$f_0 \in \arg\min_{f\in\mathcal{C}_0} \max_{Q\in\mathcal{P}(\mathcal{E})} \mathbb{E}_Q \left[\mathbb{1}(f(\mathbf{V}) \neq Y)\right], \tag{3}$$

Therefore, for making the ML system that works well (minimizes statistical risks) in all environment, one must design the prediction model based on the causal predictors, which agrees with Principle 3.

4. Identification of Causal Features through Invariance

So far, we shows that ML models trained using causal predictors PA_Y will be robust to the shortcut learning since they agree with Principles (1, 2). If we have a causal graph G induced by $\mathcal{M}$, then the problem of identifying causal predictors become trivial. In practical settings, however, such causal

graphs are oftentimes absent. Therefore, a strategy for identifying causal features from samples generated by multiple heterogeneous environment in $\mathcal{E}$ is required. In this section, we present such strategy by finding features whose relation with the true label Y invariant in all environments. Throughout the section, we assume that there are no bidirected edges connected to Y in $G(\mathcal{M})$; equivalently, $Y = f_Y(PA_Y, U_Y)$ and $(U_Y \perp\!\!\!\perp \mathbf{V})$ where f_Y is an arbitrary structural function.

We first define the test function checking whether relation between a set $\mathbf{X} \subseteq \mathbf{V}$ and Y is invariant:

Definition 3 (**Test function**). $T_{\mathcal{E}}(\mathbf{X}, Y)$ is called a test function if it satisfies the follow: $T_{\mathcal{E}}(\mathbf{X}, Y) = 1$ if, for all environment in $\mathcal{M}_{\mathbf{v}_i} \in \mathcal{E}$, the relation between a pair $(Y^i, \mathbf{X}^i)$ (which denotes a pair $(\mathbf{X}, Y)$ generated by the environment $\mathcal{M}_{\mathbf{v}_i}$) remains the same; and $T_{\mathcal{E}}(\mathbf{X}, Y) = 0$ otherwise.

Examples of test functions are the following:

Example 1 (Peters et al., 2016) $T_{\mathcal{E}}(\mathbf{X}, Y) = 1$ if, for any pairs of environments $(\mathcal{M}_{\mathbf{v}_i}, \mathcal{M}_{\mathbf{v}_j}) \in \mathcal{E}$, $P(Y^i|\mathbf{X}^i) = P(Y^j|\mathbf{X}^j)$.

Example 2 (Heinze-Deml et al., 2018) $T_{\mathcal{E}}(\mathbf{X}, Y) = 1$ if, for any environment $\mathcal{M}_{\mathbf{v}_i} \in \mathcal{E}$, $P(Y^i|\mathbf{X}^i) = P(Y^i|\mathbf{V}^i)$ (equivalently, $Y^i \perp\!\!\!\perp \mathbf{V}^i \backslash \mathbf{X}^i | \mathbf{X}^i$).

Remark 2 (**Sufficient Condition for the test function**). *By modularity property of the SCM and the assumption that there are no bidirected edges connected to Y, a function $g_{\mathcal{E}}(\mathbf{X}, Y)$ is a valid test function if $g_{\mathcal{E}}(\mathbf{X}, Y) = 1$ whenever $PA_Y \subseteq \mathbf{X}$, and 0 otherwise, because the relation between $(Y, \mathbf{X})$ remains invariant if $PA_Y \subseteq \mathbf{X}$.*

Now, we will use this test function for identifying causal features. We relax Def. 3 and consider a *high-probability test function* $T(\mathbf{X}, Y)$, which can identify the invariant features with high probability; i.e., for some $\alpha \in (0, 1)$,

$$P(T_{\mathcal{E}}(\mathbf{X}, Y) = 1) > 1 - \alpha \text{ if the relation } (\mathbf{X}, Y) \text{ is invariant over environments in } \mathcal{E}. \tag{4}$$

Equipped with such oracle function, we can identify causal features in high probability. Formally,

Theorem 4 (**Identifying causal features with high probability**). *Let $T_{\mathcal{E}}(\mathbf{X}, Y)$ denote the high-probability test function in Eq. (4). Let*

$$\widehat{PA}_Y \coloneqq \bigcap_{\mathbf{X} \subseteq \mathbf{V}} \{\mathbf{X} \subseteq \mathbf{V} \text{ such that } T_{\mathcal{E}}(\mathbf{X}, Y) = 1\}.$$

Then, with high probability, $\widehat{PA}_Y$ identifies the causal features; i.e.,

$$P\left(\widehat{PA}_Y \subseteq PA_Y\right) > 1 - \alpha.$$

We note that Thm. 4 implies that invariance over environments implies causal features. Therefore, Thm. 4 justifies the approach of using the invariance property for identifying the causal features.

5. Summary

In this document, we introduce the task of preventing the shortcut learning in Task 1. We then present two approaches for achieving Task 1:

Approach 1. Avoid causally irrelevant features in training the ML predictors as much as possible.

Approach 2. Design the estimator that works best in the 'worst' environment w.r.t. prediction.

We then formalize Approaches (1,2) as Theorems (1, 2, 3), and show that both approaches imply that the ML model must be trained with causal features to prevent the shortcut learning, as presented in Principle 3. Finally, we relate the task of identifying causal features with identifying features invariant over environments in Theorem 4. This result justifies the approa, ch of establishing a ML model with invariant features.

References

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. (2020), "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, 2, 665–673.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018), "Invariant causal prediction for nonlinear models," *Journal of Causal Inference*, 6.

Pearl, J. (2000), *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press, 2nd edition, 2009.

Peters, J., Bühlmann, P., and Meinshausen, N. (2016), "Causal inference by using invariant prediction: identification and confidence intervals," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78, 947–1012.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. (2018), "Invariant models for causal transfer learning," *The Journal of Machine Learning Research*, 19, 1309–1342.

Appendix

A. Proofs

Proof of Lemma 1. By *do*-calculus (Pearl, 2000) Rule 3.

Proof of Theorem 1. Let $\mathbf{W}_a \coloneqq PA_Y$ in this proof. Let $\mathbf{W}_b \coloneqq \mathbf{V} \backslash \mathbf{W}$. We first note that $\mathbf{V} \backslash \mathbf{W}$ satisfies the constraints; i.e.,

$$P(\mathbf{y}|do(\mathbf{w}_a, \mathbf{w}_b)) = P(\mathbf{y}|do(\mathbf{w}_a, \mathbf{w}'_b)).$$

This is obvious by Lemma 1 and *do*-calculus (Pearl, 2000) Rule 3.

We now show that $\mathbf{W}_b$ is a maximizer. To witness, consider $\mathbf{W}_b \cup V_k$ for $V_k \in PA_Y$. This set doesn't satisfy the constraints since

$$\mathbb{E}\left[Y|do(\mathbf{w}_a \backslash v_k, \mathbf{w}_b, v_k)\right] \neq \mathbb{E}\left[Y|do(\mathbf{w}_a \backslash v_k, \mathbf{w}'_b, v'_k)\right],$$

since V_k is causally relevant to Y. This concludes the proof.

Proof of Theorem 2 It suffices to show that, for any distribution $P \in \mathcal{P}(\mathcal{E})$ and a function f, there exists $Q \in \mathcal{P}(\mathcal{E})$ s.t.

$$\mathbb{E}_P\left[(Y - f_0(\mathbf{V}))^2\right] \leq \mathbb{E}_Q\left[(Y - f(\mathbf{V}))^2\right].$$

Choose $Q(\mathbf{V}, Y) \coloneqq P(PA_Y, Y)P(\mathbf{V} \backslash PA_Y)$. Then, Y and $\mathbf{V} \backslash PA_Y$ is independent conditioned on PA_Y in the distribution Q. Then,

$$\mathbb{E}_P\left[(Y - f_0(\mathbf{V}))^2\right] = \mathbb{E}_Q\left[(Y - f_0(\mathbf{V}))^2\right] \leq \mathbb{E}_Q\left[(Y - f(\mathbf{V}))^2\right].$$

To witness the second inequality, it suffices to show that $f_0(\mathbf{V}) = \mathbb{E}_Q\left[Y|\mathbf{V}\right]$ since the minimizer of the mean squared loss is its conditional expectation. Since

$$\mathbb{E}_Q\left[Y|\mathbf{V}\right] = \mathbb{E}_Q\left[Y|PA_Y\right] = \mathbb{E}_P\left[Y|PA_Y\right],$$

it concludes the proof.

Proof of Corollary 3. It suffices to show that, for any distribution $P \in \mathcal{P}(\mathcal{E})$ and a function f, there exists $Q \in \mathcal{P}(\mathcal{E})$ s.t.

$$\mathbb{E}_P\left[\mathbb{1}(f_0(\mathbf{V}) \neq Y)\right] \leq \mathbb{E}_Q\left[\mathbb{1}(f(\mathbf{V}) \neq Y)\right].$$

Choose $Q(\mathbf{V}, Y) \coloneqq P(PA_Y, Y)P(\mathbf{V} \backslash PA_Y)$. Then, Y and $\mathbf{V} \backslash PA_Y$ is independent conditioned on PA_Y in the distribution Q. Then,

$$\mathbb{E}_P\left[\mathbb{1}(f_0(\mathbf{V}) \neq Y)\right] = \mathbb{E}_Q\left[\mathbb{1}(f_0(\mathbf{V}) \neq Y)\right] \leq \mathbb{E}_Q\left[\mathbb{1}(f_0(\mathbf{V}) \neq Y)\right]$$

To witness the second inequality, it suffices to show that $f_0(\mathbf{V}) = \arg\max_y Q(Y = y|\mathbf{V})$ since the minimizer of the mean squared loss is the Bayes optimal classifier. Since

$$Q(Y = y|\mathbf{V}) = Q(Y = y|PA_Y) = P(Y = y|PA_Y),$$

it concludes the proof.

Proof of Theorem 4. The following holds:

$$P\left(\widehat{PA}_Y \subseteq PA_Y\right) \geq P\left(T_{\mathcal{E}}(PA_Y, Y) = 1\right) > 1 - \alpha.$$

The second inequality is obvious since the relation (Y, PA_Y) is invariant, and the test function fails w/ probability α. To show the first inequality, it suffices to show that the event $T_{\mathcal{E}}(PA_Y, Y) = 1$ implies the event $\widehat{PA}_Y \subseteq PA_Y$. Suppose $T_{\mathcal{E}}(PA_Y, Y) = 1$. Note,

$$PA_Y = \bigcap_{\mathbf{X} \subseteq \mathbf{V}} \{\mathbf{X} \subseteq \mathbf{V} \text{ such that } T_{\mathcal{E}}(\mathbf{X}, Y) = 1\},$$

since no set of variables missing variables in PA_Y is invariant to Y; i.e., $P(Y|\mathbf{X})$ is invariant if $\mathbf{X}$ misses a variable in PA_Y. Therefore, if $T_{\mathcal{E}}(PA_Y, Y) = 1$ then $P\left(\widehat{PA}_Y \subseteq PA_Y\right)$. This completes the proof.