Adaptation speed of causal models concerning fairness

Yujie Lin
School of New Media and Communications, Tianjin University
linyujie_22@tju.edu.cn

Chen Zhao
Baylor University
charliezhaoyinpeng@gmail.com

Minglai Shao*
School of New Media and Communications, Tianjin University
shaoml@tju.edu.cn

Xujiang Zhao
NEC Lab America
zhaoxuj32@gmail.com

Haifeng Chen
NEC Labs
haifeng@nec-labs.com

ABSTRACT

When considering machine learning tasks aimed at bidirectional training, it is common practice to employ the source corpus as the target corpus, requiring the training of two models with opposing directions. The prompt question of which model demonstrates superior adaptability to domain shifts holds substantial significance across various disciplines. Specifically, we examine the case wherein an original distribution $\boldsymbol{p}$ undergoes transformations resulting from an unknown intervention, leading to the emergence of a modified distribution $\boldsymbol{p}^*$. Multiple factors, such as causal dependencies among variables within $\boldsymbol{p}$, influence the rate of adaptation when aligning $\boldsymbol{p}$ with $\boldsymbol{p}^*$. Nevertheless, real-life scenarios necessitate the consideration of fairness during the training process, particularly when incorporating a sensitive variable (bias) situated between a cause and an effect variable. To investigate this scenario, we scrutinize a simplified structural causal model (SCM) featuring a cause-bias-effect structure, wherein variable A functions as a sensitive intermediary between the cause and the effect. The two models demonstrate consistent and contradictory cause-effect directions within the cause-bias-effect SCM, respectively. By subjecting variables within the SCM to unknown interventions, we can simulate various domain shifts to facilitate analysis. Consequently, we compare the adaptation speeds of the two models across four shift scenarios while also establishing the connection between their adaptation speeds across all interventions.

KEYWORDS

Domain shift; Adapation speed; Fairness learning

ACM Reference Format:
Yujie Lin, Chen Zhao, Minglai Shao, Xujiang Zhao, and Haifeng Chen. 2023. Adaptation speed of causal models concerning fairness. In *Proceedings of SIGKDD Workshop on Ethical Artificial Intelligence (EAI-KDD'23).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding authors

EAI-KDD'23, August 7, 2023, Long Beach, California

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

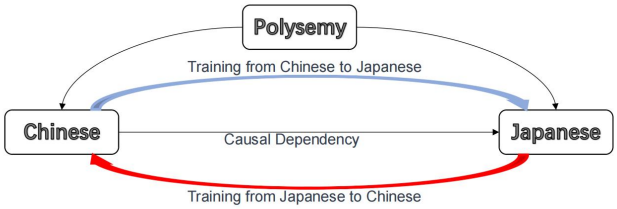


Figure 1: An example of fairness learning. The causal graph illustrates Chinese as the cause and Japanese as the effect, as Japanese developed from Chinese, with polysemy serving as the sensitive variable (bias) between the two.

1 INTRODUCTION

AI systems can be used in many sensitive environments to make important and life-changing decisions; thus, it is crucial to ensure that these decisions do not reflect discriminatory behavior toward certain groups or populations [5]. A sensitive feature is defined as an attribute that contains protected information about individuals or groups within a dataset. Machine learning models have the potential to inadvertently acquire discriminatory patterns if sensitive variables exhibit spurious correlations with the target variable or predictive outcomes. Consequently, this can lead to biased decisions or unfair predictions that disproportionately impact specific individuals or groups. Therefore, it is crucial to ensure that these decisions do not reflect discriminatory behavior towards particular groups or populations [5].

Let's consider a common situation. When translating Chinese to Japanese, the impact of polysemous words cannot be overlooked. Polysemous words, or words with multiple meanings, can lead to ambiguity in translation, especially when the target language has different meanings for the same word. Such ambiguity can be particularly problematic when the polysemous word serves as a sensitive variable in fairness learning, where the translation may have unintended consequences on the fairness of the resulting model. Therefore, it is crucial to carefully consider the impact of polysemous words when translating between languages, especially when such words serve as sensitive variables in fairness learning (Fig 1). Furthermore, modern machine learning methods encounter surprising failures when the test distribution differs from the training distribution, which is commonly known as *domain shift*. Although many domain adaptation [7] and domain generalization [8] methods have been proposed in recent years to mitigate the problem of domain shift, in reality, we often

need to achieve the best possible results and for the convenience of training, we still need to retrain the data in the new domain. In the process of relearning from the original distribution to the new distribution, the relative speed of adaptation between the causal model and the anti-causal model may differ. Considering this SCM ($X \to Y$), the previous work [4] analyzed the adaptation speed of two models trained in the direction of the causal dependency (i.e., from X to Y) and in the reverse direction of the causal dependency (i.e., from Y to X) concerning domain shift. While they have obtained some promising findings, the analysis of adaptation speed did not take into account fairness considerations (*i.e.,* sensitive variables), which are crucial in real-life scenarios. There are currently theoretical and experimental gaps in the analysis of this aspect, which is precisely what our work aims to address.

Now let us provide a more formal definition. As shown in Fig 1, the three variables (bias: A, cause: X, effect: Y) related to fairness can form a cause-bias-effect structural causal model ($A \to X$ and $A, X \to Y$). We consider two training models with sensitive variables (bias): one that has the consistent cause-effect direction of the cause-bias-effect SCM (i.e., causal model), and the other that has the opposite direction(i.e., anti-causal model). Domain shift can be described as the transition from the initial joint distribution $\boldsymbol{p}(A, X, Y)$ in the training set to the joint distribution $\boldsymbol{p}^*(A, X, Y)$. We establish different distribution shift to explain the relationship between the adaptation speed of the two models in situations. In conclusion, our **contributions** can be summarized as follows:

- In summary, our work provides a significant contribution to the understanding of the adaptation speed of causal models in scenarios involving sensitive variables. The common types of domain shift can be categorized into covariate shift and concept shift. However, when considering fairness factors, there can also be a drift in the distribution of sensitive variables. As demonstrated in the appendix, when the distribution of sensitive variables drifts without altering the relative adaptation speed caused by other variable shifts, we focus only on the scenario where the distributions of cause and bias variables change simultaneously. We considered domain shift caused by four types of interventions on the data distribution, including interventions on bias, cause, interventions on both bias and cause, and interventions on effect. And we conducted theoretical analyses for each of the four cases.
- We analyze the relationship between the adaptation speed of two different causal models and provide insights into this topic, using synthetic real data. Our work is the first of its kind and offers valuable information on this topic.

2 PRELIMINARY

2.1 Interventions on SCMs

Structural causal models (SCMs) are widely used in causal inference to model the causal relationships among variables. An SCM consists of a directed acyclic graph (DAG) and a set of structural equations that define the causal relationships among the variables in the graph [6]. Interventions on SCMs involve changing the value of a variable to a specified value. This can be represented mathematically using the do-operator, denoted by $\mathbf{do}(X = x)$. The do-operator separates the effect of an intervention from the effect of other variables in the system. For example, if we want to investigate the effect of drug treatment on a disease outcome, we might use the do-operator to set the value of the treatment variable to "treated" and observe the effect on the outcome variable. In the following narrative, we will use $\boldsymbol{p}^*$ to represent this modified distribution, such as

$$\boldsymbol{p}^*(a, x, y) = \boldsymbol{p}(a, x, y|\mathbf{do}(x = t)). \tag{1}$$

By controlling one or several variables in this way, we simulate domain shifts under different scenarios. And $\boldsymbol{p}^*$ is the outcome after domain shifts.

2.2 Reference and Transfer Distributions

We obtain the initial reference distribution $\boldsymbol{p}$ by sampling the triad (A, X, Y) from a structural causal model (SCM) constructed as follows: A is a bias, X is the cause, and Y is the effect. The SCM is defined by the following two equations:

$$A \to X, \tag{2}$$

$$A, X \to Y. \tag{3}$$

If the intervention is on the bias, we sample A from a different marginal distribution, while X and Y are sampled from the reference conditional distribution:

$$\boldsymbol{p}^*(a, x, y) = \boldsymbol{p}^*(a)\boldsymbol{p}(x|a)\boldsymbol{p}(y|a, x). \tag{4}$$

If the intervention is on the cause, A is sampled from the reference marginal distribution, X is sampled from a different marginal distribution independently of A, and Y is sampled from the reference conditional distribution:

$$\boldsymbol{p}^*(a, x, y) = \boldsymbol{p}(a)\boldsymbol{p}^*(x)\boldsymbol{p}(y|a, x). \tag{5}$$

If the intervention is on both the bias and the cause, A is sampled from a different marginal distribution, X is sampled from a different marginal distribution independently of A, and Y is sampled from the reference conditional distribution:

$$\boldsymbol{p}^*(a, x, y) = \boldsymbol{p}^*(a)\boldsymbol{p}^*(x)\boldsymbol{p}(y|a, x). \tag{6}$$

If the intervention is on the effect, A is sampled from the reference marginal distribution, X is sampled from the reference conditional distribution, and Y is sampled from a different marginal distribution independently of A and X:

$$\boldsymbol{p}^*(a, x, y) = \boldsymbol{p}(a)\boldsymbol{p}(x|a)\boldsymbol{p}^*(y). \tag{7}$$

Thus, we obtain all the transfer joint distributions that arise from interventions on some of the variables.

2.3 Fairness-aware Models for Training

The models mentioned in this section, which are distinct from the SCM and are used for training, are referred to as causal models and anti-causal models, respectively. The causal model and the anti-causal model are constructed with the variables (A,X,Y). The causal model can be described as

$$\boldsymbol{p}_{\theta_{\rightarrow}}(a, x, y) = \boldsymbol{p}_{\theta_A}(a)\boldsymbol{p}_{\theta_{X|A}}(x|a)\boldsymbol{p}_{\theta_{Y|A,X}}(y|a, x). \tag{8}$$

Meanwhile, the anti-causal model can be described as

$$\boldsymbol{p}_{\theta_{\leftarrow}}(a, x, y) = \boldsymbol{p}_{\theta_A}(a)\boldsymbol{p}_{\theta_{Y|A}}(y|a)\boldsymbol{p}_{\theta_{X|A,Y}}(x|a, y), \tag{9}$$

where the $\theta_{\rightarrow}$ and $\theta_{\leftarrow}$ represent parameters of the two models respectively (e.g. $\theta_{\rightarrow}$ includes θ_A, $\theta_{X|A}$, and $\theta_{Y|A,X}$).

2.4 Model Adaptation

This section explains the most fundamental issue of this work. Assuming the initial distribution of two models is both p, and due to certain factors, the distribution drifts to p^*, the training process is to make the models approach the distribution p^*. When the training is about to start, the training term is $T := 0$. The two models will be initialized to fit the same reference distribution $\mathbf{p}$ like

$$\boldsymbol{p}_{\theta_{\rightarrow}^{(0)}} = \boldsymbol{p}_{\theta_{\leftarrow}^{(0)}} = \boldsymbol{p}. \tag{10}$$

Then we get samples from the transfer distribution p^*. Letting these samples join the training process, the log-likelihood will gradually increase in every step of stochastic gradient descent (SGD). The distribution $\boldsymbol{p}_\theta$ adapts to p^* closest until we get the minimal log-likelihood loss. Taking the causal model for example, the loss is

$$\begin{aligned}\mathcal{L}_{\text{causal}}(\theta_{\rightarrow}) =& \mathbb{E}_{(A,X,Y)\sim\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{\rightarrow}}(A,X,Y)\right] \\ =& \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_A}(A)\right] + \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{A|X}}(A|X)\right] \\ &+ \mathbb{E}_{\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{Y|A,X}}(Y|A,X)\right],\end{aligned} \tag{11}$$

where the log-likelihood suboptimality is equal to the KL-divergence

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) = D_{\text{KL}}(\boldsymbol{p}^* || \boldsymbol{p}_\theta). \tag{12}$$

2.5 Parameters in Trainable Models

We assume that the bias A, the cause X, and the effect Y in the two models are multiclass variables with K classes. The natural parameter $\boldsymbol{\theta} \in \mathbb{R}^K$ is generated from the distribution p by the inverse function of the softmax function

$$p_z = \frac{e^{s_z}}{\sum_{z'} e^{s_{z'}}}. \tag{13}$$

We set $\boldsymbol{s}$ as the trainable parameter in SGD. Taking causal model for example, the model has parameters $\boldsymbol{s}_A := (s_a)_{a=1\ldots K}$, $\boldsymbol{s}_{X|A} := (s_{x|a})_{a,x=1\ldots K}$ and $\boldsymbol{s}_{Y|A,X} := (s_{y|a,x})_{a,x,y=1\ldots K}$. The parameters of causal model can be represented as $\theta_{\rightarrow} = (\boldsymbol{s}_A, \boldsymbol{s}_{X|A}, \boldsymbol{s}_{Y|A,X})$, while that's $\theta_{\leftarrow} = (\boldsymbol{s}_A, \boldsymbol{s}_{Y|A}, \boldsymbol{s}_{X|A,Y})$ in anti-causal model. Using the parameter $\boldsymbol{s}$, the loss (11) can be written as

$$\begin{aligned}\mathcal{L}_{\text{causal}}(\theta_{\rightarrow}) &= \mathbb{E}_{(A,X,Y)\sim\boldsymbol{p}^*}\left[-\log \boldsymbol{p}_{\theta_{\rightarrow}}(A,X,Y)\right] \\ &= \mathbb{E}_{\boldsymbol{p}^*}[-s_A + \log\sum_a e^{s_a} - s_{X|A} + \log\sum_x e^{s_{x|A}} \\ &\quad - s_{Y|A,X} + \log\sum_y e^{s_{y|A,X}}].\end{aligned} \tag{14}$$

3 SPEED ANALYSIS

In this section, we will discuss the parameters involved in the training process and present our findings on the comparative speed of the causal and anti-causal models.

3.1 Distance Inequality

Based on Average Stochastic Gradient Descent (ASGD) [2], the previous work [4] proves that the average parameter's $\bar{\theta}^{(T)} = \frac{1}{T}\sum_{t=0}^{T-1}\theta^{(t)}$ suboptimality is upper bounded by

$$\mathbb{E}\left[D_{\text{KL}}(\boldsymbol{p}^* || \boldsymbol{p}_{\bar{\theta}^{(T)}})\right] \leq \frac{c^{-1}\|\theta^{(0)} - \theta^*\|^2 + cB^2}{2\sqrt{T}}, \tag{15}$$

where c is a small enough constant. This inequality indicates that the upper bound of the convergence is mainly determined by the distance $\triangle := \|\theta^{(0)} - \theta^*\|^2$ between the reference distribution and the transfer distribution. Specifically, the initial distance of the causal model and the anti-causal model is respectively denoted by $\triangle_{\text{causal}} = \|\theta_{\rightarrow}^{(0)} - \theta_{\rightarrow}^*\|^2$ and $\triangle_{\text{anticausal}} = \|\theta_{\leftarrow}^{(0)} - \theta_{\leftarrow}^*\|^2$. And the two distances are the core basis of the subsequent discussion.

3.2 Adaptation Speeds of Two Models

Furthermore, comparing the convergence speed of two models can be equivalently expressed as comparing the magnitudes of $\triangle_{\text{causal}} = \|\theta_{\rightarrow}^{(0)} - \theta_{\rightarrow}^*\|^2$ and $\triangle_{\text{anticausal}} = \|\theta_{\leftarrow}^{(0)} - \theta_{\leftarrow}^*\|^2$, the model with a smaller initial distance ($\triangle$) has a faster convergence speed.

Intervention on bias A, $s_A \leftarrow s_A^*$. In this scenario, both the causal and anti-causal models modify only the same sensitive marginal $\boldsymbol{s}_A$. The initial distance between the two models can be expressed as

$$\triangle_{\text{causal}} = \triangle_{\text{anticausal}} = \|\boldsymbol{s}_A - \boldsymbol{s}_A^*\|^2, \tag{16}$$

which implies that the two models converge simultaneously.

Intervention on cause X, $\forall a, \boldsymbol{s}_{X|a} \leftarrow \boldsymbol{s}_X^*$. The conditional $\boldsymbol{s}_{Y|A,X}$ remains unchanged while $\boldsymbol{s}_{Y|A}$ and $\boldsymbol{s}_{X|A,Y}$ of anti-causl model are modified.The initial distance can be written as

$$\triangle_{\text{causal}} = \sum_a \|\boldsymbol{s}_{X|a} - \boldsymbol{s}_X^*\|^2, \tag{17}$$

$$\triangle_{\text{anticausal}} = \sum_a \|\boldsymbol{s}_{Y|a} - \boldsymbol{s}_{Y|a}^*\|^2 + \sum_a\sum_y \|\boldsymbol{s}_{X|a,y} - \boldsymbol{s}_{X|a,y}^*\|^2. \tag{18}$$

We can compare the initial distances of the two models based on the aforementioned distance and arrive at the following proposition.

PROPOSITION 1. *When the intervention is on the cause,*

$$\triangle_{anticausal} \geq K\triangle_{causal}, \tag{19}$$

where the specific proof process will be explained in Appendix.

Intervention on both bias A and cause X, $s_A \leftarrow s_A^*$ and $\forall a, \boldsymbol{s}_{X|a} \leftarrow \boldsymbol{s}_X^*$. Compared to the case of intervening on cause X, the intervention on bias A introduces an additional equal distance to the initial distance of the two models. Hence, we can obtain a similar result as before and derive the following inequality:

$$\triangle_{\text{anticausal}} \geq \triangle_{\text{causal}}, \tag{20}$$

which will be simply explained in Appendix.It can be observed that in all three scenarios, the initial distance of the causal model is consistently smaller than that of the anti-causal model, indicating that the causal model adapts to the domain more quickly. However, if the domain shift is induced by interventions on the effect variable (Y), interestingly, different conclusions can be drawn.

Intervention on effect Y, $\forall a, x, \boldsymbol{s}_{Y|a,x} \leftarrow \boldsymbol{s}_Y^*$. The marginal $\boldsymbol{s}_A$ and the conditional $\boldsymbol{s}_{X|a}$ remain unchanged under this intervention. On the other hand, the conditional $\boldsymbol{s}_{Y|A}$ and $\boldsymbol{s}_{Y|A,X}$ in the anti-causal model change with respect to effect Y. The initial distances for the two models can be expressed as:

$$\triangle_{\text{causal}} = \|\boldsymbol{s}_Y - \boldsymbol{s}_{Y|A,X}^*\|^2, \tag{21}$$

$$\triangle_{\text{anticausal}} = \sum_a \|\boldsymbol{s}_{Y|a} - \boldsymbol{s}_Y^*\|^2 + \sum_a\sum_y \|\boldsymbol{s}_{X|a,y} - \boldsymbol{s}_{X|a,y}^*\|^2, \tag{22}$$

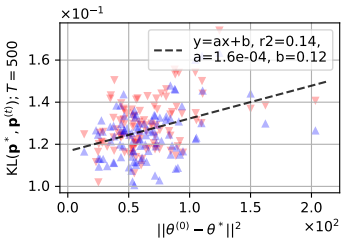
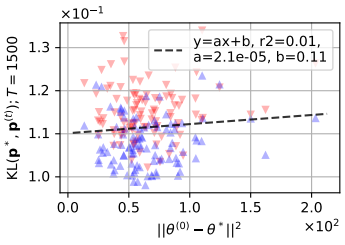
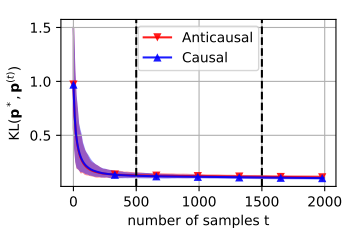
Intervention	Scatter Plot T=500	Scatter Plot T=1500	Convergence Curve
Bias			

Table 1: Results on synthetic data. The scatter plots in the first column demonstrate the positive correlation between the KL-divergence after one-quarter of the training steps, while the second column shows the correlation after three-quarters of the training steps. Each point on the scatter plots represents a pair $(p^{(0)}, p^*)$ in the causal model (blue) or the anti-causal model (red). The curve in the third column shows the relative speeds of the two models. The shaded area indicates the 5th and 95th percentiles of the KL-divergence. The value of $K_{synthetic}$ is set to 20 in this experiment.

where the distance above does not maintain a constant relationship compared to the other three situations.

PROPOSITION 2. *When the intervention is on the effect, there will be two situations. If the following inequality:*

$$\|\boldsymbol{s}_Y^* - \boldsymbol{c}\|^2 < R^2 \tag{23}$$

is satisfied, then $\triangle_{anticausal} \geq \triangle_{causal}$ *(proved in Appendix). Let us illustrate this inequality (23), where* R^2 *(see Appendix for specific formula) is a constant related with* $\boldsymbol{s}_X$, $\boldsymbol{s}_{Y|A}$, $\boldsymbol{s}_{X|A,Y}$ *and* $\boldsymbol{s}_{Y|A,X}$ *, and* $\boldsymbol{c} = \frac{(\sum_x \boldsymbol{s}_{Y|A,x}) - \boldsymbol{s}_{Y|A}}{K-1}$.

The inequality (23) implies that the causal model has a comparative advantage only within a certain range, where the modified marginal $\boldsymbol{s}_Y$ is sufficiently close to $\boldsymbol{c}$. However, if the distance goes beyond this range, the anti-causal model will converge faster.

4 EXPERIMENTS

4.1 Data

In this section, we evaluate the criterion of adaptation speed using a synthetic dataset. The first we need is to get the distributions $\boldsymbol{p} = \boldsymbol{p}_{\theta^{(0)}}$ which is called *prior*. Specifically, we get $\boldsymbol{p}_A$, $\boldsymbol{p}_{X|A}$ and $\boldsymbol{p}_{Y|A,X}$ from the Dirichlet distribution. The three distributions can be represented as:

$$\begin{aligned} \boldsymbol{p}_A &\sim Dirichlet(\mathbf{1}_K), \\ \forall a, \boldsymbol{p}_{X|a} &\sim Dirichlet(\mathbf{1}_K), \\ \forall a, x, \boldsymbol{p}_{Y|a,x} &\sim Dirichlet(\mathbf{1}_K), \end{aligned}$$

where $\mathbf{1}_K$ is the all-one vector of K-dimension. Such a prior has been previously adopted by the work [1]. And distributions sampled from this prior exhibit some asymmetry between X and Y [3]. We can say that by using the aforementioned distributions, we ensure that these three distributions are mutually independent. Now we have thus obtained an initial joint distribution.

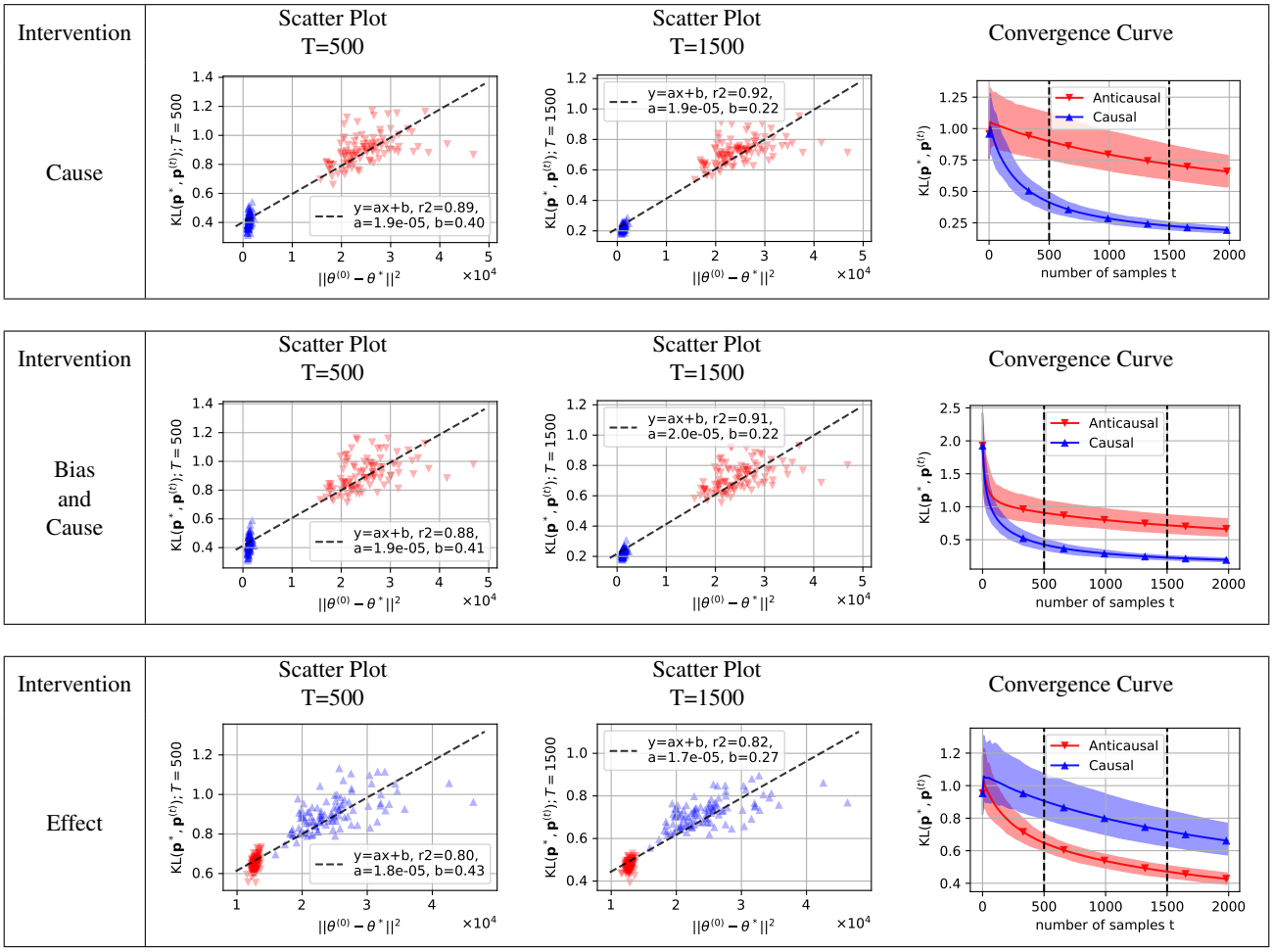
4.2 Results

This section demonstrates the adaptation speed of the causal and anti-causal models using the data from section 4.1, as well as the positive relationship between the KL divergence and the initial parameter distance. Using $K_{synthetic} = 20$, the results for the three datasets are presented in Table.1 to demonstrate the adaptation speed of the causal and anti-causal models. Additionally, scatter plots are used to depict the positive correlation between the KL divergence and the initial parameter distance during the training process. The results after one-quarter and three-quarters of the training steps are shown.

Intervention on bias A. When intervening on the bias A, both the causal and anti-causal models undergo the same change with respect to the bias, while holding other variables constant. Therefore, both models exhibit overlapping points on the scatter plots, resulting in the initial distance $\delta_{\text{causal}} = \delta_{\text{anticausal}}$ and coinciding curves in the third column. As a result, the convergence speed of both models is the same. This is in line with the observation that interventions on biases do not change the causal structure of the model, and thus do not provide any new information that can be used to distinguish between the two models.

Intervention on cause X. In the case of an intervention on the cause X, the causal model has an advantage over the anti-causal model, as the points of the causal model cluster towards the bottom left as compared to the points of the anti-causal model. This relative positioning reflects the formula $\delta_{\text{causal}} < \delta_{\text{anticausal}}$, which is derived in our previous work. This advantage of the causal model is most pronounced in four intervention scenarios, where the causal model exhibits a significant advantage in terms of the curves.

Intervention on both bias A and cause X. When intervening on both the bias and the cause, the relative positions of points in the causal and anti-causal models maintain their relative positions, with the anti-causal model being on the upper right of the causal model as a whole. However, the difference in adaptation speed between the two models shrinks, as the two curves have a short overlap in the early training steps. This result is somewhat counterintuitive, as one might expect that intervening on both the bias and the cause would provide more information that could be used to distinguish between the two models. However, our results suggest that this is not always the case, and that the advantage of the causal model may depend on the specific structure of the model.



Intervention on effect Y. When intervening on the effect Y, the relative positions of points in the causal and anti-causal models are reversed, with the points of the anti-causal model being concentrated in the lower-left corner. This advantage of the anti-causal model corresponds to the proposition that the anti-causal model is better suited to handle interventions on effect variables. Although the anti-causal model may not always have this advantage, it typically retains it in most cases. The curves also reflect this result, as the anti-causal model consistently converges faster when using the three datasets. Our results suggest that interventions on effect variables provide valuable information that can be used to distinguish between the causal and anti-causal models.

5 CONCLUSION

We investigate the adaptation speed of both causal and anti-causal models in the presence of bias and build upon a theory that explains the relationship between the initial distance of parameters and the adaptation speed. Furthermore, it's a challenge to extend our analysis to models with more complex structural differences, such as those with varying numbers of variables and edges. To analyze the adaptation speed in such cases, an indicator to measure the difference between the models is necessary. However, finding an appropriate indicator can be challenging to analyze in the future.

REFERENCES

[1] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. 2019. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912* (2019).

[2] Léon Bottou. 2012. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*. Springer, 421–436.

[3] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. 2016. Estimating causal direction and confounding of two discrete variables. *arXiv preprint arXiv:1611.01504* (2016).

[4] Rémi Le Priol, Reza Babanezhad Harikandeh, Yoshua Bengio, and Simon Lacoste-Julien. 2020. An Analysis of the Adaptation Speed of Causal Models. *arXiv e-prints*, Article arXiv:2005.09136 (May 2020), arXiv:2005.09136 pages. arXiv:2005.09136 [stat.ML]

[5] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[6] Judea Pearl. 2009. *Causality*. Cambridge University Press.

[7] Mei Wang and Weihong Deng. 2018. Deep visual domain adaptation: A survey. *Neurocomputing* 312 (2018), 135–153.

[8] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2022. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).