IMPROVING PRACTICAL COUNTERFACTUAL FAIRNESS WITH LIMITED CAUSAL KNOWLEDGE

Zeyu Zhou, Ruqi Bai & David I. Inouye
Elmore Family School of Electrical and Computer Engineering
Purdue University
{zhou1059, bai116, dinouye}@purdue.edu

ABSTRACT

The widespread application of foundation models across various domains raises significant concerns regarding fairness and bias. In this work, we focus on a specific notion of fairness, Counterfactual Fairness (CF), which posits that an individual's outcome should remain consistent if they had belonged to a different sensitive group. CF is grounded in an underlying causal model, and it typically necessitates either access to the true causal model or the availability of counterfactual pairs. While previous studies have made some progress when such information is available, acquiring it is often challenging in real-world applications. In this paper, we target at achieving CF in a more practical setting where limited causal knowledge is available. We demonstrate that naive adaptations of existing methods are inadequate in such contexts through extensive empirical studies. To bridge the gap, we first introduce a more carefully designed approach for generating counterfactuals in practice, compatible with existing methodologies. Subsequently, we present a technique for utilizing estimated counterfactuals and potentially biased pretraind models. The feasibility of our approaches is validated through both theory and empirical investigation.

1 INTRODUCTION

Machine learning (ML) has been widely used in high-impact domains such as healthcare (Daneshjou et al., 2021), hiring (Hoffman et al., 2018), criminal justice (Brennan et al., 2009), and loan assessment (Khandani et al., 2010), bringing with it critical ethical and social considerations. This issue is particularly alarming in an era where foundation models, commonly trained on noisy data from the internet, are increasingly prevalent (Bommasani et al., 2021; Hellman, 2023). Such models, due to their extensive reach and impact, amplify the potential for widespread and systemic biases. This increasing awareness underscores the need for ML practitioners to integrate fairness considerations into their work, extending their focus beyond merely maximizing prediction accuracy (Bolukbasi et al., 2016; Calders & Verwer, 2010; Dwork et al., 2012; Grgic-Hlaca et al., 2016; Hardt et al., 2016). Various fairness notions have been developed, ranging from group-level measures such as group parity (Hardt et al., 2016) to individual-focused metrics like Individual Fairness (Dwork et al., 2012). Recently, there has been a growing interest in approaches based on causal inference, particularly in understanding the causal effects of sensitive attributes on decision-making (Chiappa, 2019; Galhotra et al., 2022; Khademi et al., 2019). This has led to the proposal of Counterfactual Fairness (CF), which states that prediction for an individual in hypothetical scenarios where their protected attributes differ should remain unchanged (Kusner et al., 2017). Such an approach enables the development of algorithms that do not just ignore protected attributes but also acknowledge and compensate for social biases linked to ethically sensitive attributes effectively.

To achieve CF, Kusner et al. (2017) first propose a naive solution, suggesting that predictions should only use non-descendants of the sensitive attribute in a causal graph. This approach only requires a causal topological ordering of variables and achieves perfect CF by construction. However, it limits the available features for downstream tasks and could be inapplicable in certain cases (Kusner et al., 2017). To mitigate this, they further propose more practical algorithms that leverage exogenous noise but must assume that the causal graph or full causal model are known. Extending this line of work, Zuo et al. (2023) introduce a technique that incorporates additional information by mixing

factual and counterfactual samples. Although CF has been theoretically and empirically established in their work, the efficacy of such sample mixing in preserving prediction accuracy remains an open question. Moreover, this approach requires access to the true causal model for the estimation of counterfactual samples, which are often difficult to obtain in practical settings. Parallel to this, another branch of research employs methods such as regularization and augmentation (Kim et al., 2021; Garg et al., 2019; Stefano et al., 2020), but these cannot provide theoretical guarantees of CF. More discussion on related works can be found in Appendix A. In summary, prior work on CF either has weak machine learning efficacy or assumes knowledge of the underlying causal model from which counterfactuals can be easily approximated.

However, in many practical contexts, the underlying causal model may be unknown. Thus, the question arises: What if methods must estimate counterfactuals *with limited knowledge of the underlying causal model*? To answer this, we illustrate in Figure 1 the difference in performance for the method from Zuo et al. (2023) when ground truth counterfactuals are available compared to when they are naively estimated.In this situation, the lack of ground truth counterfactuals significantly increases *both* the error and unfairness as measured by Error and Total Effect defined in Section 2.2. These issues are explored in more depth in Section 3.

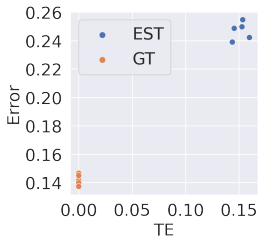


Figure 1: Illustration of the error of CFR (Zuo et al., 2023) without access to the ground truth counterfactual pairs. Different dots represent repetitive experiments. Total Effect is a metric for Counterfactual Fairness, where a lower value indicates greater fairness. The estimation model (EST) employs the same structure as the ground truth (GT) for approximating counterfactual samples. CFR achieves almost 0 Total Effect and lower Error given ground truth (GT) counterfactual samples, as per the original setup described in the referenced study. However, a significant performance degradation is noted when CFR must rely on estimated counterfactuals, despite using a function class that aligns with the GT model. This highlights the importance of practical approach for predicting counterfactuals in practice.

In light of this, we aim to achieve CF while also preserving prediction accuracy in the practical context where neither causal model nor even causal graph are known. In Section 4.1, we introduce a refined approach for generating counterfactual samples, which could be integrated into existing CF methods. Then, in Section 4.2, we introduce a simple post-processing method that effectively combines these estimated counterfactuals and a (unfair) pretrained model to achieve better CF while preserving the accuracy of the pretrained model. As this step can leverage any pretrained model, the approach can follow the prevailing trend in the ML community that favors the use of off-the-shelf pretrained models, such as foundation models. This integration represents a step towards efficiently leveraging existing resources while advancing CF. We summarize our contributions as follows:

1. In the context of limited causal knowledge, we conduct empirical studies to illustrate the ineffectiveness of naively applying previous CF methods.

2. We formalize a generic framework that decomposes prior CF methods, highlighting the importance of effectively estimating counterfactauls. We further propose an approach that leverages recent advancements in counterfactual estimation, which notably does not necessitate identifiability of the causal model.

3. We introduce a simple yet effective CF algorithm that utilizes (estimated) counterfactual samples and pretrained models such as foundation models.

2 PRELIMINARIES

2.1 NOTATION

We use capital letter to represent random variables and lowercase letter to represent the realization of random variables. Now we define a few variables that will be used frequently in this work. A represents the sensitive attribute of an individual, Y represents the target variable to predict, X represents observed features other than A and Y, and U represents unobserved variables which are not caused by any observed variables while a, y, x, u represent their realization respectively. For simplicity, we only consider 1-dimensional binary A and 1-dimensional Y, but the investigation and our method can be naturally extended to multi-dimensional cases.

We assume all data are generated by a causal model (Pearl, 2009) such as the one in Figure 2a. Following this causal graph, data is generated as below

$$A \sim P_A \qquad U \sim P_U \qquad X = F_X(U, A) \qquad Y = F_Y(U, X)$$

where P_A is typically a Bernoulli distribution and P_U depends on the context. We further define Counterfactual Generating Mechanism (CGM) as $G(x, a, a') = x_{a'}$ which typically contains two steps: (1) Estimating posterior distribution of exogenous noise $u \sim P(U|X = a, A = a)$. This step could be deterministic or stochastic and is typically approximated by $u = E(x, a)$. (2) Generating counterfactuals $x_{a'} = D(u, a')$. [1] We define *ground truth* counterfactual pairs as follows: given a fixed u, factual and counterfactual sample are generated by $x_a = D(u, a)$ and $x_{a'} = D(u, a')$. We generally do not consider Y as part of CGM.

2.2 COUNTERFACTUAL FAIRNESS

There are different fairness criteria such as Group Fairness (Calders et al., 2009; vZliobaitė, 2015) , Individual Fairness (Dwork et al., 2012; Zemel et al., 2013) and Counterfactual Fairness (CF) (Kusner et al., 2017). In this work, we focus on CF, which states that intervention on A should not affect the prediction of Y and is formally defined as below.

Definition 2.1. *(Counterfactual Fairness) We say a classifier is counterfactual fair if* $P(\hat{Y}(A = a)|X = x, A = a) = P(\hat{Y}(A = a')|X = x, A = a)$.

Now we introduce a few previous methods of CF that will be investigated in this paper.

Empirical Risk Minimization (ERM). Directly train a classifier on all features without any fainress consideration. Specifically $\widehat{y} = \phi(x, a)$, where ϕ represents the predictor.

Counterfactual Data Augmentation (CDA). The input form of classifer is the same as ERM, however, in the training set, we include counterfactual samples. The counterfactual samples is generated by either ground truth or estimated CGM $x_{a'} = G(x, a, a')$. Multiple previous works have adopted similar approaches. For example, Kim et al. (2021) proposes Disentangled Causal Effect Variational Autoencoder (DCEVAE) to generate counterfactual pairs that are used to train classifier. Zuo et al. (2023) extends their method by including real samples, and they explore the usage of both CVAE (Sohn et al., 2015) and DCEVAE in the case of knowing ground truth causal models. A common characteristic of these two works is that they include counterfactual $y_{a'}$. However, we argue that it could be hard to estimate $y_{a'}$ without access to ground truth model. Hence, in the following investigation, we keep $y_{a'}$ the same as y_a when we need to estimate the causal model. Specifically, the training set is $\mathcal{D}_{\mathcal{D}} = \{x^{(i)}, y^{(i)}, a^{(i)}\}_{i=1}^{N} \cup \{x_{a'}^{(i)}, y^{(i)}, a'^{(i)}\}_{i=1}^{N}$.

Counterfactual Fairness with exogenous noise (CFE) (Kusner et al., 2017). To achieve Counterfactual Fairness, CFE proposes to use U for prediction. Specifically, $\widehat{y} = \phi(u)$ where $u = E(x, a)$.

Counterfactual Fairness with fair representation (CFR) (Zuo et al., 2023). CFR proposes to use U and a symmetric version of $x, x_{a'}$. Specifically, $\widehat{y} = \phi(\frac{x+x_{a'}}{2}, u)$ where $u = E(x, a), x_{a'} = G(x, a, a')$. Note that CFR would require access to $x_{a'}$ in the test set as well while CDA does not require anything special and CFE requires u.

[1] This follows the three steps Abduction, Action, and Prediction defined in Pearl (2009)

(a) (b)

Figure 2: (a) This illustration provides an example of a causal graph. Not all data distributions examined in our work follow this specific graph. Rather, it represents one of the commonly utilized graphs in the field of fairness. (b) Causal structure of ILD.

Metrics We consider two metrics in this paper: Error and Total Effect (TE). The former evaluates if each method can achieve its goal regardless of fair or not. This is important because we can achieve perfect Counterfactual Fairness by always outputing fixed prediction given whatever input, but that is not useful at all. The latter is a common metric to evaluate Counterfactual Fairness (Zuo et al., 2023). Given a test set $\mathcal{D}_{\mathcal{D}_{\text{test}}}$, Error is defined as Error $= \frac{1}{|\mathcal{D}_{\mathcal{D}_{\text{test}}}|} \sum_{x^{(i)} \in \mathcal{D}_{\mathcal{D}_{\text{test}}}} \ell(\widehat{y}(x^{(i)}), y^{(i)})$ where $y^{(i)}$ is the ground truth target, $\widehat{y}(x^{(i)})$ is the prediction of $x^{(i)}$, and ℓ depends on the task. TE is defined as TE $= \frac{1}{|\mathcal{D}_{\mathcal{D}_{\text{test}}}|} \sum_{x^{(i)} \in \mathcal{D}_{\mathcal{D}_{\text{test}}}} |\widehat{y}(x^{(i)}) - \widehat{y}(x^{(i)}_{a'})|$ where $x_{a'}$ is the ground truth counterfactual corresponding to $x^{(i)}$. Since we only consider binary sensitive attribute, we further define $\text{TE}_0 = \frac{1}{|\{i:a^{(i)}=0\}|} \sum_{i:a^{(i)}=0} |\widehat{y}(x^{(i)}) - \widehat{y}(x^{(i)}_{a'})|$ and $\text{TE}_1 = \frac{1}{|\{i:a^{(i)}=1\}|} \sum_{i:a^{(i)}=1} |\widehat{y}(x^{(i)}) - \widehat{y}(x^{(i)}_{a'})|$ to evaluate Counterfactual Fairness for different group respectively.

2.3 ILD

Anonymous (2024) proposes Invertible Latent Domain Causal Model (ILD) which targets at estimating counterfactuals without fully identifying causal structures. ILD is proposed to deal with domain counterfactuals. For example, what would a medical imaging look like if it had been taken from another hospital? Interpreting A as the domain node in their paper, we can ask a similar question: what would this sample look like if it had been in another sensitive group? ILD assumes invertible latent SCM where all data are generated following the graph in Figure 2b and functions below

$$A \sim P_A \qquad U \sim P_U \qquad Z = F_Z(U, A) \qquad X = F_X(Z)$$

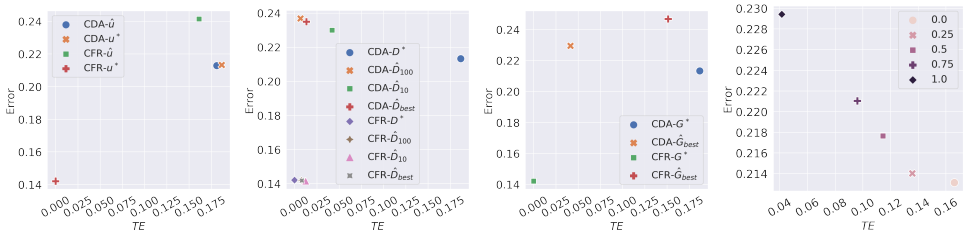
where F_Z and F_X are invertible functions and U serves as the exogenous noise of latent variables Z. They propose a canonical form of model which pushes all effect of A to the last few nodes [2]. This leads to a simpler optimization towards counterfactual equivalence models which, while potentially differing in structure from the ground truth causal model, would generate exactly the same counterfactuals. Thus, the only knowledge required here is sparsity, specifically the number of variables that are directly affected by A.More details could be seen in Appendix C.

3 Issues with Estimating Counterfactuals

The majority of previous research on CF assumes the availability of either the true causal model or a set of counterfactual pairs. While in certain situations, constructing the causal graph may be feasible with expert knowledge, determining the functional form of the causal model can often prove challenging. Moreover, even with such information, it is non-trivial to estimate counterfactuals, further adding to the challenge of obtaining counterfactuals in practice. Here we undertake an empirical investigation into the breakdown of CF methods when using estimated counterfactuals.

Experiment Setup For dataset, we consider UCI Adult Income (Kohavi et al., 1996) and Law School Success (Wightman, 1998) following the choice of Zuo et al. (2023). However, since we need to evaluate TE of each method which requires access to ground truth, we use the simulated version of those datasets. Specifically, we first train a ground truth CGM $G^* = D^* \circ E^*$ that simulates the CGM of those datasets. Then we use E^* and the real X to infer U. After getting samples of U, we constructed the dataset $\mathcal{D}$ and $\mathcal{D}_{\text{counterfactual}}$ by sampling from marginal distribution of A (based on actual distribution of real data), and using D^* to get x_a and $x_{a'}$. We want to emphasize that $\mathcal{D}_{\text{counterfactual}}$, regardless of train or test set, are hidden from downstream models and used for evaluation only. This way, we get access to the ground truth u^* and can generate ground truth counterfactuals without any error. In our investigation, exogenous noise, factual data and counterfactual data are all actually the simulated version of original datasets. However they do follow a fixed mechanism that is close to the real data. For clarity, we called them Adult-Sim and

[2] in terms of topological ordering



(a) Error from estimating U. (b) Error from estimating D. (c) Error from estimating G. (d) Investigation of CDA-$\hat{G}_{best}$.

Figure 3: Error from estimating counterfactuals on Adult-Sim. (a) CFR-u^* is much better than CFR-$\hat{u}$ in terms of both metrics. Importantly, it achieves almost 0 TE. The difference between CDA-u^* and CDA-$\hat{u}$ is not obvious. (b) $\hat{D}_{10}, \hat{D}_{100}, \hat{D}_{best}$ represent checkpoints after 10 epochs, 100 epochs and final checkpoint with best validation loss respectively. TE of CFR-$\hat{D}_{best}$ gets closer to that of CFR-$\hat{D}^*$ but the gap still exists. Through the training process of CDA, we observe a U-shape change of performance. Further analysis might be needed to fully understand how CDA could help or hurt Counterfactual Fairness. (c) Similar to (a), the performance of CFR degrades in both metrics when estimating G. For CDA, estimating $\hat{G}$ leads to a better TE but worse accuracy. (d) The number represents the ratio of counterfactual samples being added to the training set. For example, 0.5 means 50% of the counterfactual samples are used in the training.

Law-Sim throughout the paper. For Adult-Sim, Error is evaluated as the ratio of wrong prediction while for Law-Sim, Error is evaluated as the Root Mean Squared Error (RMSE).

To generate simulated "ground truth" Y, the training of ground truth G^* involves Y, i.e., $x = D^*_X(u, a)$ and $y = D^*_Y(u, x)$. Regarding model structure, we implement CVAE as used in Zuo et al. (2023). We consider two previous methods: CDA and CFR. We do not include CFE here because it only uses U. More details could be found in Appendix E. Due to space issue, results on Law-Sim and numbers containing TE_0 and TE_1 can be found in Appendix F. All numbers are averaged over 5 repetitive experiments.

Error from estimation of U The first step of estimating counterfactuals is to infer the exogenous noise, which models the randomness and unobserved effect in systems (Pearl, 2009). Even given the true data generating mechanism, we cannot estimate perfect counterfactuals without being able to identify U. Counterfactuals are generated as follows:

$$\text{Oracle } U: x^*_{a'} = D^*(u^*, a') \qquad \text{Estimated } U: \hat{u} = \hat{E}(x, a),\ \ \hat{x}_{a'} = D^*(\hat{u}, a')$$

where u^* is the exogenous noise used to generate the datasets and $\hat{E}$ is trained jointly in $\hat{G}$. Note that $x^*_{a'}$ is exactly the same as the counterfactual samples in $\mathcal{D}_{\text{counterfactual}}$.

Observation. In Figure 3a, we present a comparison of the Error and TE with CFE and CFR. The results show that when ground truth u^* is accessible, CFR achieves almost perfect CF, which is also demonstrated in Zuo et al. (2023). However, despite utilizing the correct D^*, the efficacy significantly diminishes when attempting to infer U, which lead to bad performance in terms of both Error and TE. Besides, we find that the Error of CFR could get as bad as predicting a fixed value (which would be around 0.24). [3] The difference between CDA-$\hat{u}$ and CDA-u^* in unclear which could because there is no guarantee if CDA will help or not. The performance degradation of CFR highlights a practical challenge: even with knowledge of the data generating mechanism D, reverse engineering the exogenous noise is not straightforward without certain assumptions, such as invertibility. This simple yet important observation underscores the need for more cautious handling of u in future algorithm design.

Error from estimation of D Here we investigate what if we have access to the ground truth exogenous noise but not the ground truth D^*. In contrast with last section, here we generate counterfactuals as follows:

$$\text{Oracle } D: x^*_{a'} = D^*(u^*, a') \qquad \text{Estimated } D: \hat{x}_{a'} = \hat{D}(u^*, a')$$

[3] Note that in this case, it is not predicting fixed output, otherwise TE would become 0.

where $\hat{D}$ is trained jointly in $\hat{G}$ that has a same CVAE structure as G^*.

Observation. To investigate how CFR and CDA's performance change through the training, we show the performance with different $\hat{D}$ checkpoints. As shown in Figure 3b, for CFR, as the model is fitted better, TE could improve but can never perform as well as that with D^*. Besides, we notice that despite the gap between estimated $\hat{D}$ and D^* still exists, it is relatively smaller than that with estimated $\hat{u}$. We conjecture this is because u is also an input to CFR. Since the algorithm sees the ground truth u^*, the effect of bad estimation $x_{a'}$ is less severe (in contrast, bad estimation of u^* would also lead to bad $x_{a'}$ even if we know D^*). For CDA, we observe that the final model tends to predict fixed value regardless of input which is indicated by Error around 0.24 and very low TE. This suggests it may not be optimal to naively use the same y for factual and counterfactual pair.

Error from estimation of G Finally, what if we don't have knowledge of G^* at all, i.e., a combination of previous errors? Counterfactuals are generated as follows:

$$\text{Oracle } G: x^*_{a'} = D^*(u^*, a') \qquad \text{Estimated } G: \hat{u} = \hat{E}(x, a),\ \ \hat{x}_{a'} = \hat{D}(\hat{u}, a')$$

Observation. In Figure 3c, we first notice that CFR completely fails with estimated $\hat{G}$, which is as expected based on previous observations. For CDA, the reason why CDA-G^* gets better Error but worse TE is still unclear. Essentially, we cannot prove whether CDA could really help with TE given G^*. Besides, we notice that they are both much worse than the performance of CFR-G^*. To further investigate the behavior of CDA, we add an investigation in Figure 3d where we add different portion of augmented counterfactual data to the training set. Here we observe a trend of trade-off between TE and Error. This makes sense as we use the same y for both x_a and $x_{a'}$, and as we add more samples to the training set, the downstream model would learn to predict the same output for counterfactuals samples. However, this does not necessarily lead to a better prediction performance.

Conclusion The three exploratory studies discussed in this section demonstrate that the performance of state-of-the-art method for CF is substantially compromised when there is incomplete understanding of the CGM, including exogenous noise and causal models. This finding challenges the applicability of previous methods that rely on such knowledge, indicating they may not be suitable for practical scenarios. Besides, the observation that CDA outperforms CFR in the case of estimating $\hat{G}$ also suggests the necessity of revaluating the performance of each method under this situation. These two findings highlight the need for improvements in counterfactual estimation and usage.

4 Fair Algorithm with estimated counterfactuals

To get started, we propose a generic framework for achieving CF, facilitating a more clear comparison between our methods and previous work. This framework is detailed in Algorithm 1. It's important to note that not every algorithm within this framework requires all the inputs listed there and discussion on how previous methods fits in this framework can be seen in Appendix D. While most prior works assumes the availability of G^* and skip Step 1, in the previous section, we identified significant challenges in applying current methods when such assumptions do not hold. To tackle these issues, we propose two key solutions: (1) Development of enhanced methods for counterfactual estimation. (2) Refinement of strategies for employing estimated counterfactuals. The first solution alters the foundational step (Step 1), essential for establishing CF in real-world scenarios, while the second solution (Step 3), building upon this groundwork, facilitates the practical application of pre-trained models, which may originally be unfair.

4.1 Estimating Counterfactuals

ILD is one of the recent methods for generating counterfactuals (Anonymous, 2024). This section will first explain why ILD is particularly apt for Step 1 in Algorithm 1. Following this, we will delve into several challenges encountered when applying ILD to achieve CF.

Benefits of ILD for CF There are two major merits of ILD: (1) It is capable of handling *latent* causal variables. (2) It does not need knowledge of the causal model or even the causal graph. ILD was originally designed under the framework of latent causal models, where all causal variables lie in a latent space and all observables are connected with latent causal factors via a shared observation function. As depicted in Figure 2b, variables Z are causally significant yet unobserved while X

represents the set of observed variables. At first glance, this might seem less advantageous. In contrast with image datasets investigated by the original ILD paper, on tabular datasets, where many fairness concerns are concentrated, the setup of latent causal model appears less straightforward. Nonetheless, it is important to first recognize that there are no theoretical constraints preventing the use of ILD in such contexts. A naive way to address the gap would be to use trivial F_X such as identity or shuffling function. Furthermore, even though each observed variable in tabular datasets is more interpretable than pixels in images, they are not necessarily the key variables in the causal model. In other words, the features we observed could be the mixture of or noisy version of other causal observables. For instance, blood pressure and heart rate measure the "healthiness" of the heart which is a latent variable. Thus, adopting the ILD framework makes it a more generic approach.

Algorithm 1 CF with counterfactuals

Input: Counterfactual Estimation Algorithm $\mathcal{A}_{\text{counterfactual}}$, CF training algorithm $\mathcal{A}_{\text{CF-Train}}$, CF inference algorithm $\mathcal{A}_{\text{CF-Infer}}$, training dataset for $\hat{G}$ $\mathcal{D}_{\text{counterfactual-train}} = \{(x_i, a_i)\}_{i=1}^{n}$, training dataset for ϕ $\mathcal{D}_{\text{cf-train}} = \{(x_i, a_i, y_i)\}_{i=1}^{n'}$, test point $(x_{\mathcal{D}_{\text{test}}}, a_{\mathcal{D}_{\text{test}}})$
Output: Prediction result
Step 1: Obtain CGM $\hat{G}$
$\hat{G} \leftarrow \mathcal{A}_{\text{counterfactual}}(\mathcal{D}_{\text{counterfactual-train}})$
Step 2: Train the predictor ϕ
$\phi \leftarrow \mathcal{A}_{\text{CF-Train}}(\mathcal{D}_{\text{cf-train}}, \hat{G})$
Step 3: Make the final prediction $\hat{y}$
$\hat{y} \leftarrow \mathcal{A}_{\text{CF-Infer}}(x_{\mathcal{D}_{\text{test}}}, a_{\mathcal{D}_{\text{test}}}, \phi, \hat{G})$

The second benefit regarding true causal model is also crucial, as in practical scenarios, we often lack such knowledge. Furthermore, identifying the causal graph itself can be challenging. These tasks have been well studied in the field of causal discovery (Chickering, 2002; Colombo et al., 2014) and causal representation learning (Schölkopf et al., 2021). Solutions to this problem typically rely on strong assumptions, such as the linearity of Structural Causal Models (SCMs) or additive noise (Shimizu et al., 2006; Hoyer et al., 2008; Peters et al., 2014). Since our focus is on generating appropriate counterfactual samples rather than identifying the underlying causal model, ILD becomes a natural fit by avoiding estimating causal models and estimating counterfactuals directly instead .

Challenges of ILD for CF Despite ILD's benefits, its application in our context faces challenges. Initially designed for image datasets, adapting ILD to tabular data is non-straightforward. Additionally, ILD's assumption of invertibility for functions F_Z and F_X poses a three-fold difficulty. First, even though invertibility requirement is more practical than assumptions such as linearity or additive noise, it could still be restrictive in some sense. For example, F_Z being invertible constrains the randomness that U can handle, and could thus restrict it capability to model counterfactual distribution. However, as most prior works typically employ a lower dimensional U in practice, we follow that trend and leave theoretical advancement of ILD for further work. Second, training an invertible model, particularly as data dimensions increase, is a non-trivial task in practice. In our case, we implement ILD using a pseudo-invertible VAE structure, where E and D are parameterized by an encoder and decoder respectively as is done in Anonymous (2024). Finally, the prevalence of categorical variables in tabular data presents another obstacle to the invertibility assumption. We approach this by converting categorical variables using one-hot encoding and modeling them with Gumbel-Softmax (Jang et al., 2016), in line with the methodology used in Xu et al. (2019).

4.2 Postprocessing algorithm with pretrained models

With a trained ILD or other counterfactual generation methods, we now explore the possibility of leveraging (unfair) pretrained models to achieve CF while maintaining the performance of the pretrained model. This approach saves resources from retraining a downstream model, especially for complicated tasks. Additionally, this enables the use of off-the-shelf powerful pretrained models such as foundations models. Foundation models are often trained on vast quantities of unstructured and noisy internet data which could contain biases that adversely affect CF (Bommasani et al., 2021). Towards this end, we first theoretically prove several results that will motivate our postprocessing algorithm. All proofs are in Appendix B. For theoretical analysis, we assume that $F_X^*(u, a)$ [4] is

[4] F_X corresponds to D^* and we use this notation to distinguish theory and implementation.

invertible w.r.t. u given a, i.e., $\exists {F_X^*}^{-1}, {F_X^*}^{-1}(x,a) = {F_X^*}^{-1}(F_X^*(u,a),a) = u, \forall (x,a)$. [5] This ensures that CGM is a deterministic function as the exogenous noise can be recovered exactly from x and a. The following lemma characterizes the constraint on ϕ that is equivalent to perfect TE.

Lemma 4.1. *A predictor is perfectly counterfactually fair w.r.t. TE if and only if the predictor returns the same value for a sample and its counterfactuals, i.e.,* $\mathrm{TE}(\phi) = 0 \Leftrightarrow \phi(x,a) = \phi(x_{a'}, a'), \quad \forall (x,a,a')$

The proof is straightforward from the definition of $\mathrm{TE}(\phi)$. Given this, we now prove that the optimal fair predictor, i.e., the minimal loss under the constraint of perfect fairness, is a simple weighted average of the optimal (unfair) predictor.

Theorem 4.2. *If F_X^* is invertible, the optimal fair predictor (w.r.t. square L_2 loss for regression and cross-entropy loss for classification), i.e., the best possible model under the constraint of perfect CF, is the average of the optimal (unfair) predictor on all possible counterfactuals:*

$$\phi^*_{\mathrm{CF}}(x,a) \triangleq \underset{\phi:\mathrm{TE}(\phi)=0}{\arg\min}\ \mathbb{E}[\ell(\phi(X,A),Y)] = \textstyle\sum_{\tilde{a}\in\{0,1\}} p(A=\tilde{a})\phi^*(x_{\tilde{a}},\tilde{a})$$

where $\phi^*(x,a)$ is the optimal predictor without a fairness constraint, i.e., $\phi^*(x,a) \triangleq \arg\min_\phi \mathbb{E}[\ell(\phi(X,A),Y)] = \mathbb{E}[Y|X=x, A=a]$

This result suggests that, if G^* could be estimated reasonably well, a simple postprocessing algorithm of an unfair model could achieve strong fairness and accuracy. We propose a simple algorithm as summarized in Algorithm 2, which serves as $\mathcal{A}_{\text{CF-Infer}}$ in Step 3 of Algorithm 1. For general task

Algorithm 2 Postprocessing for CF (PCF)

Input: Pretrained probabilistic prediction model $\phi : \mathcal{X} \to \mathcal{Y}$, CGM G, factual test point (x,a), prior distribution p of A
Output: Predicted output $\hat{\mu}$
for $\tilde{a} \in \{0,1\}$ **do**
  $\hat{x}_{\tilde{a}} \leftarrow G(x,a,\tilde{a})$
end for
$\hat{\mu} \leftarrow \sum_{\tilde{a}\in\{0,1\}} p(A=\tilde{a})\phi(\hat{x}_{\tilde{a}},\tilde{a})$

such as regression, $\hat{\mu}$ is the final output while for classification, $\hat{\mu}$ is equivalent to the probability of $Y=1$, i.e., $p(Y=1|X=x,A=a) = \mathbb{E}[Y|X=x,A=a]$. It is important to highlight that PCF may use any model ϕ which may not depend on G or be trained in a fair manner. In fact, with access to the oracle CGM G^*, PCF would achieve perfect CF as proved in the next result.

Proposition 4.3. *If F_X^* is invertible and G is the ground truth CGM, i.e., $G(x,a,a') = x_{a'}, \forall (x,a,a')$, then Algorithm 2 achieves perfect CF for any pretrained model ϕ.*

This indicates that PCF could achieve perfect CF with ground truth CGM G^* *regardless of the pretrained model* ϕ. And if ϕ achieves strong accuracy, then the corresponding PCF is likely to achieve strong accuracy, which we empirically validate in our experiments. From another perspective, there are consistent estimators for the optimal predictor.Thus, we also mention the case where we do have access to the optimal (unfair) predictor, i.e., $\phi^*(x,a)$ and consider the TE and excess risk of Algorithm 2 incurred by using an imperfect CGM $\hat{G}$ instead of the perfect CGM G^*. We conjecture that simple bounds on TE and excess risk could be made based on the Lipschitz smoothness of ϕ^* and the amount of counterfactual error, i.e., $\epsilon_{\max} = \max_{x,a,a'} \|G^*(x,a,a') - \hat{G}(x,a,a')\|_2$. Intuitively, if the counterfactuals are not too far away and ϕ^* is smooth, then TE and accuracy will not be affected significantly because $\hat{\mu}$ will be close to the ideal μ.

5 Experiments

5.1 Estimated counterfactuals

Experiment Setup Here we empirically investigate if ILD method in Section 4.1 leads to better performance when combined with CF algorithms. We consider two different G: (1) G_{CVAE}: CVAE

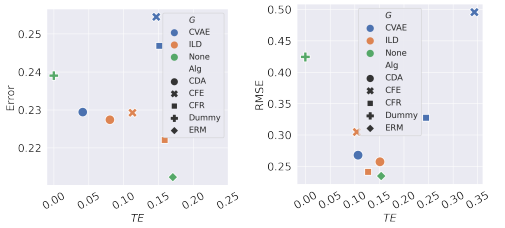
[5] This simplifies the theoretic analysis but we conjecture similar results would hold for approximately invertible F_X^* as well though it would require different analysis tools.

used in Section 3. (2) G_{ILD}: ILD. Details on model design and parameter choice can be found in Appendix E. We investigate how CDA, CFE and CFR perform given counterfactuals provided by (1) and (2). We also include ERM and Dummy as a baseline. For Dummy it always predict 0 on Adult-Sim and always predict the mean of y in the training set on Law-Sim. Regarding dataset, for a fair comparison, we use the G^* used in Section 3, which has the same structure as G_{CVAE}.

Observation In Figure 4a, we observe that on the Adult-Sim dataset, ILD and CVAE show a trade-off between TE and Error for CDA and CFR while ILD outperforms CVAE in terms of both metrics when integrated with CFE. It's particularly noteworthy that CFR-CVAE exhibits significantly higher error compared to CFR-ILD, with only a marginal gain in TE. In fact, its performance is even inferior to predicting the fixed output. In Figure 4b, for the Law-Sim dataset, ILD demonstrates superior performance over CVAE when integrated with CFE and CFR. For CFA, a similar trade-off pattern is observed. In summary, ILD generally provides more effective counterfactual estimation than CVAE, despite CVAE sharing the same model class as the ground truth G^*. Futhermore, even though being more fair, all of them have higher Error than ERM, which suggests that we find better ways of utilizing counterfactuals in this context. Results involving TE_0 and TE_1 can be seen in Appendix F.

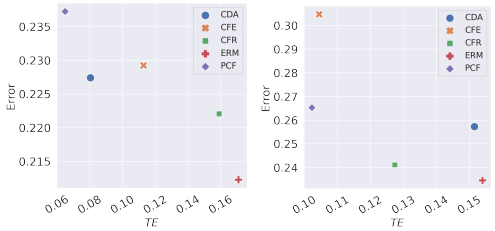
5.2 Postprocessing

Experiment Setup Here we investigate the effectiveness of PCF proposed in Section 4.2. Built upon our investigation in last section, we only use ILD to generate counterfactual samples. We compare the performance of CDA-ILD, CFE-ILD, CFR-ILD, and PCF-ILD. The pretrained prediction models used in PCF-ILD is trained via ERM, and we also include performance of that model without any postprocessing, marked as ERM.



(a) Adult-Sim (b) Law-Sim

Figure 4: Comparison between CVAE and ILD integrated with CF algorithms. Color indicates $\hat{G}$ and shape indicates algorithms.



(a) Adult-Sim (b) Law-Sim

Figure 5: Effectiveness of PCF in comparison to previous CF algortihms with estimated counterfactual samples via ILD on both dataset.

Observation In Figure 5a and Figure 5b, we observe that PCF achieves best TE. However, we notice that PCF cannot achieve good ML performance. This could result from either predictor for each group not being optimal or counterfactual estimation error. We leave further investigation as future works. Additional results and expanded figures can be found in Appendix F.

6 Conclusion and Future Works

In our study, we evaluate the efficacy of existing CF algorithms in situations with limited causal knowledge. As a first step, we enhance their performance using ILD for superior counterfactual sample generation. These simple yet insightful empirical investigations highlight the importance of refining counterfactual estimation methodologies for fairness. Inspired by a few theorems on the optimal prediction model, we further propose a new algorithm that utilizes estimated counterfactuals and off-the-shelf pretrained models.

Despite the merits of our method and investigations, there are limitations that present intriguing avenues for future research. For instance, while ILD's effectiveness is evident in our study, it doesn't assure that we could always achieve counterfactual equivalence solely by fitting the observed distribution. Investigating methods to further narrow down the search space of ILD is a promising direction for future work. Besides, our current empirical study does not incorporate foundation models. Therefore, examining the practicality of our approach with actual foundation models presents an intriguing pathway for future investigations. This might also close the gap in ML efficacy that we observed in the empirical study.

ACKNOWLEDGEMENT

Z.Z., R.B. and D.I. acknowledge support from ARL (W911NF-2020-221) and ONR (N00014-23-C-1016).

REFERENCES

Junaid Ali, Matthäus Kleindessner, Florian Wenzel, Kailash Budhathoki, Volkan Cevher, and Chris Russell. Evaluating the fairness of discriminative foundation models in computer vision. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 809–833, 2023.

Anonymous. Towards characterizing domain counterfactuals for invertible latent causal models. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=v1VvCWJAL8`.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and behavior*, 36(1):21–40, 2009.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE international conference on data mining workshops*, pp. 13–18. IEEE, 2009.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7801–7808, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.

Roxana Daneshjou, Kailas Vodrahalli, Weixin Liang, Roberto A Novoa, Melissa Jenkins, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai: assessments using diverse clinical images. *arXiv preprint arXiv:2111.08006*, 2021.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Causal feature selection for algorithmic fairness. In *Proceedings of the 2022 International Conference on Management of Data*, pp. 276–285, 2022.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor (eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pp. 219–226. ACM, 2019. doi: 10.1145/3306618.3317950. URL https://doi.org/10.1145/3306618.3317950.

Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Mach. Learn.*, 112(3):741–763, 2023. doi: 10.1007/S10994-022-06206-8. URL https://doi.org/10.1007/s10994-022-06206-8.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, pp. 11. Barcelona, Spain, 2016.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Deborah Hellman. Big data and compounding injustice. *Journal of Moral Philosophy*, 1(aop):1–22, 2023.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *arXiv preprint arXiv:2303.15715*, 2023.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Mitchell Hoffman, Lisa B Kahn, and Danielle Li. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2):765–800, 2018.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Wenyue Hua, Yingqiang Ge, Shuyuan Xu, Jianchao Ji, and Yongfeng Zhang. Up5: Unbiased foundation model for fairness-aware recommendation. *arXiv preprint arXiv:2305.12090*, 2023.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Aria Khademi, Sanghack Lee, David Foley, and Vasant Honavar. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, pp. 2907–2914, 2019.

Amir E Khandani, Adlar J Kim, and Andrew W Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767–2787, 2010.

Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8128–8136, 2021.

Ron Kohavi et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Christopher Meek. Causal inference and causal explanation with background knowledge. *arXiv preprint arXiv:1302.4972*, 2013.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual identifiability of bijective causal models. *arXiv preprint arXiv:2302.02228*, 2023.

Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *International Conference on Machine Learning*, pp. 16848–16887. PMLR, 2022.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. 2014.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Lucas Rosenblatt and R. Teal Witter. Counterfactual fairness is basically demographic parity. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pp. 14461–14469. AAAI Press, 2023. doi: 10.1609/AAAI.V37I12.26691. URL https://doi.org/10.1609/aaai.v37i12.26691.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Abhin Shah, Raaz Dwivedi, Devavrat Shah, and Gregory W Wornell. On counterfactual inference with unobserved confounding. *arXiv preprint arXiv:2211.08209*, 2022.

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, Antti Kerminen, and Michael Jordan. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(10), 2006.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Pietro G. Di Stefano, James M. Hickey, and Vlasios Vasileiou. Counterfactual fairness: removing direct effects through regularization. *CoRR*, abs/2002.10774, 2020. URL https://arxiv.org/abs/2002.10774.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 214–229, 2022.

Linda F Wightman. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. *Advances in neural information processing systems*, 32, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Indré vZliobaitė. On the relation between accuracy and fairness in binary classification. In *The 2nd Workshop on Fairness, Accountability, and Transparency in Machine Learning, Lille, France*, volume 452, 2015.

Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. Counterfactual fairness with partially known causal graph. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Zhiqun Zuo, Mohammad Mahdi Khalili, and Xueru Zhang. Counterfactually fair representation. *Advances in neural information processing systems*, 2023.

A Related works

Fairness of Foundation Models Foundation models refer to the machine learning models that are trained on broad data and can be adapted to a wide range of downstream tasks (Bommasani et al., 2021). The significant impact of foundation models has led to increased focus on its social implications regarding fairness and bias (Henderson et al., 2023; Bender et al., 2021; Weidinger et al., 2022). Ali et al. (2023) use relaxed fair PCA as a postprocessing method for CLIP (Radford et al., 2021). Their analysis is limited to CLIP-based models and does not address Counterfactual Fairness. Hua et al. (2023) propose a Counterfactually-Fair-Prompting method for recommendation systems based on Large Language Models. Similarly, this is specific to one type of foundation models. In contrast with most prior works, our paper focuses on estimating counterfactual samples and post-processing output, which could be broadly applied for different types of foundation models.

Counterfactual Fairness Counterfactual Fairness, first introduced by Kusner et al. (2017), has recently gained traction (Nilforoshan et al., 2022; Rosenblatt & Witter, 2023), with many prior studies relying on ground truth causal models. While Garg et al. (2019) regularize counterfactual sample predictions, they depend on counterfactual pairs and lack solid fairness guarantees. Furthermore, their performance may be suboptimal, as shown by Zuo et al. (2023). Zuo et al. (2022) make progress by not fully relying on the causal graph but still require a maximally partially directed acyclic graph (Meek, 2013), contrasting with ILD relying solely on sparsity knowledge. Other research, like Kim et al. (2021) and Grari et al. (2023), explore empirical methods such as fairness loss in VAE training or adversarial loss, but their guarantee on CF remains uncertain.

Counterfactual Generation A main branch of research concentrates on identifying counterfactual queries (Nasr-Esfahany et al., 2023; Shah et al., 2022). For instance, Nasr-Esfahany et al. (2023) establish counterfactaul identifiability for certain type of causal structures. However, in comparison to ILD, their approach doesn't accommodate latent causal settings and require assumptions on the causal structure. A less stringent approach to generating counterfactuals involves the use of generative models without explicit usage of causality (Zhu et al., 2017; Choi et al., 2018). Typically, these models find a way to incorporate group information into the generative process and prioritize the generation of high-quality samples. However, due to their lack of integration with causal reasoning and the absence of guarantees on the generated samples, these approaches run the risk of introducing significant bias concerns.

B Proofs

B.1 Proof of Theorem 4.1

Proof of Theorem 4.1.

$$\mathrm{TE}(\phi) = 0 \Leftrightarrow \mathbb{E}[|\phi(X, A) - \phi(X_{A'}, A')|] = 0 \Leftrightarrow \phi(x, a) \overset{\text{a.s.}}{=} \phi(x_{a'}, a'), \quad \forall (x, a, a')\,, \quad (1)$$

where the first equality is by definition and the second equality is because absolute value is always non-negative for any (x, a, a'). Thus, the predictions must be almost surely equal for all (x, a, a'). Similarly, if they are all equal on the non-zero metric set, then the expectation must be 0. □

B.2 Proof of Theorem 4.2

Before proving the main theorem, we first provide one well-known lemma that reminds the reader of the well-known result of the optimal predictor, which is denoted by ϕ^* in the theorem statement.

Lemma B.1 (Optimal Predictor is Conditional Mean)**.** *The conditional mean $\mathbb{E}[Y|X = x]$ is the optimal predictor without fairness constraints for classification with cross entropy loss and for regression with MSE loss.*

Proof. First, let's establish that the optimal predictor without constraints is in fact $\mathbb{E}[Y|X=x]$. For squared error, we have that derivative :

$$\begin{aligned}
&\mathbb{E}[\ell(\phi(X),Y)] \\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[(Y-\phi(X))^2]] \\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[Y^2] - 2\mathbb{E}_{Y|X}[Y\phi(X)] + \mathbb{E}_{Y|X}[\phi(X)^2]] \\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[Y^2] - 2\phi(X)\mathbb{E}_{Y|X}[Y] + \phi(X)^2\mathbb{E}_{Y|X}[1]] \\
&= \mathbb{E}_X[\mathbb{E}_Y[Y^2] - 2\phi(X)\mathbb{E}[Y|X] + \phi(X)^2]
\end{aligned}$$

Taking the derivative of the inside expectation w.r.t. $\phi(X)$ and setting to 0 yields $\phi^*(X) = \mathbb{E}[Y|X]$.

Now let's look at cross entropy loss for classification:

$$\begin{aligned}
&\mathbb{E}[\ell(\phi(X),Y)] \\
&= \mathbb{E}_X[\mathbb{E}_{Y|X}[-Y\log(\phi(X)) - (1-Y)\log(1-\phi(X))]] \\
&= \mathbb{E}_X[-\log(\phi(X))\mathbb{E}[Y|X] - \log(1-\phi(X))\mathbb{E}[(1-Y)|X]]]
\end{aligned}$$

Again, if you take the derivative w.r.t. $\phi(X)$ and set to 0, we see that $\phi^*(X) = \mathbb{E}[Y|X]$. □

Now we seek to prove Theorem 4.2.

Proof. First, we decompose the factual error across the sensitive attribute A given the exogenous noise U.

$$\begin{aligned}
&\mathbb{E}_{X,A,Y}[\ell(\phi(X,A),Y)] \\
&= \mathbb{E}_{U,A,Y}[\ell(\phi(F_X^*(U,A),A),Y)] \\
&= \mathbb{E}_U[\mathbb{E}_A[\mathbb{E}_{Y|U,A}[\ell(\phi(F_X^*(U,A),A),Y)]]] \\
&= \mathbb{E}_U[p(A=a)\mathbb{E}_{Y|U,A=a}[\ell(\phi(F_X^*(U,a),a),Y)] + p(A=a')\mathbb{E}_{Y|U,A=a'}[\ell(\phi(F_X^*(U,a'),a'),Y)]]\,.
\end{aligned}$$

Consider $U=u$, inside the expectation we have

$$\begin{aligned}
&p(A=a)\mathbb{E}_{Y|U=u,A=a}[\ell(\phi(F_X^*(u,a),a),Y)] + p(A=a')\mathbb{E}_{Y|U=u,A=a'}[\ell(\phi(F_X^*(u,a'),a'),Y)] \\
&= p(A=a)\mathbb{E}_{Y|X=x,A=a}[\ell(\phi(x,a),Y)] + p(A=a')\mathbb{E}_{Y|X=x_{a'},A=a'}[\ell(\phi(x_{a'},a'),Y)]\,,
\end{aligned}$$

where w.l.o.g., x is viewed as the factual and $x_{a'}$ is viewed as the counterfactual. Because of invertibility, these two terms are unique for every (u,a) or correspondingly (x,a) combination and thus the problem decomposes across U. Thus, the factual loss can be viewed as a combination of the factual loss from one specific a plus the counterfactual loss for a' for each point x.

We have the following subproblems indexed by u: The factual loss can be viewed as a combination of the factual loss from one specific a plus the counterfactual loss for a' for each point x. Notice that the constraint is $\phi(x,a) \overset{a.s.}{=} \phi(x_{a'},a')$ from Theorem 4.1. We can directly push the constraint into the optimization problem by optimizing over $\phi_0 \triangleq \phi(x,a) \overset{a.s.}{=} \phi(x_{a'},a')$:

$$\arg\min_{\phi} p(A=a)\mathbb{E}_{Y|X=x,A=a}[\ell(\phi_0,Y)] + p(A=a')\mathbb{E}_{Y|X=x_{a'},A=a'}[\ell(\phi_0,Y)] \tag{2}$$

Taking ℓ as squared L_2 loss: we have

$$\begin{aligned}
&\arg\min_{\phi} p(A=a)\mathbb{E}_{Y|X=x,A=a}[(Y-\phi_0)^2] + p(A=a')\mathbb{E}_{Y|X=x_{a'},A=a'}[(Y-\phi_0)^2] \\
&= \arg\min_{\phi} p(A=a)\{\mathbb{E}_{Y|X=x,A=a}[Y^2] - 2\phi_0\mathbb{E}_{Y|X=x,A=a}[Y] + \phi_0^2\} \\
&+ p(A=a')\{\mathbb{E}_{Y|X=x_{a'},A=a'}[Y^2] - 2\phi_0\mathbb{E}_{Y|X=x_{a'},A=a'}[Y] + \phi_0^2\}\,.
\end{aligned}$$

Similarly, if we take ℓ as (binary) cross entropy loss: we have

$$\begin{aligned}
&\arg\min_{\phi} p(A=a)\mathbb{E}_{Y|X=x,A=a}[-(Y\log(\phi) + (1-Y)\log(1-\phi))] \\
&+ p(A=a')\mathbb{E}_{Y|X=x_{a'},A=a'}[-(Y\log(\phi) + (1-Y)\log(1-\phi))]
\end{aligned}$$

It is simple to see that both loss functions are convex, thus could obtain a unique solution by taking the derivative. Thus, for each $x, x_{a'}$ induced by $U = u$, we could get the optimal ϕ_0:

$$\phi_0 = \sum_{a' \in \{0,1\}} p(A = a')\phi^*(x_{a'}, a')\,,$$

where ϕ^* is the optimal predictor from the lemma above. This result holds for every u and thus gives the final result. □

B.3 Proof of Theorem 4.3

Proof.

$$\begin{aligned}
\mathbb{E}[\hat{Y}|X = x, A = a] &= \hat{\mu}(x, a) \\
&= \sum_{\tilde{a} \in \{0,1\}} p(A = \tilde{a})\phi(G(x, a, \tilde{a})) \\
&= \sum_{\tilde{a} \in \{0,1\}} p(A = \tilde{a})\phi(x_{\tilde{a}}) \\
&= \sum_{\tilde{a} \in \{0,1\}} p(A = \tilde{a})\phi(G(x_{a'}, a', \tilde{a})) \\
&= \hat{\mu}(x_{a'}, a') \\
&= \mathbb{E}[\hat{Y}|X = x_{a'}, A = a']\,,
\end{aligned}$$

where the middle equalities are by the properties of the deterministic and ground truth CGM. Because the factual output for Algorithm 2 is the same as the counterfactual output, then the TE must be 0 by Theorem 4.1. □

C More Details of ILD

Invertible Latent Domain Causal Model (ILD) was originally proposed in Anonymous (2024) to solve domain counterfactual problems. They assume all data are generated in the form of

$$\begin{aligned}
A &\sim P_A & U &\sim P_U \\
X &= F_X(U, A) & Y &= F_Y(U, X)
\end{aligned}$$

where F_Z and F_X are invertible functions.

Assumptions There are three main assumptions of ILD. (1) Invertibiltiy of the models. This is key to estimating exogenous noise U from the observations. (2) Soft intervention. This states that latent causal models belonging to different group are generated by soft intervention on another SCM, which changes the causal mechanism without breaking the causal relationship with respect to their exogenous noise. (3) Sparse Mechanism Shift (Schölkopf et al., 2021). This says that A change a sparse number of causal mechanisms. Such an assumption makes sense in most cases as sentitive attribute typically only affects a few certain features directly. Though it's worth emphasizing that this assumption only constrains the number of variables that are directly caused by A instead of all its descendants.

Main Theorems Here we list two key theorems that will be useful for our task. We refer the readers to their original paper for more careful theoretical discussion.

Theorem C.1. *Given an ILD, there exists a canonical ILD that is both counterfactually and distributionally equivalent while the sparsity is maintained.*

Here distributional equivalence means the distribution of X matches while counterfactual equivalence means given an observation, they will generate the same counterfactual sample. This theorem provides insights on model design. Suppose the sparsity is k and latent dimension is d, Canonical ILD means that group attribute only affects the last k nodes. Besides, the latent SCM belonging to

one of the group will be identity function. Note that this implies that the first $k - d$ nodes are identical to their corresponding exogenous noise. They further propose Relaxed Canonical ILD where they remove the constraint on one of the latent SCM being identity.

Theorem C.2. *Given an ILD, all ILDs that are counterfactually and distributionally equivalent share the same sparsity.*

This theorem further provides confidence in how ILD could improve counterfactual estimation by just fitting the observed distribution.

D Discussion of prior works

In this section, we connect previous methods with our generic framework in Algorithm 1.

ERM ERM does not require any $\hat{G}$ so it can naturally skip Step 1. The second step uses a conventional ERM algorithm $\mathcal{A}_{\text{CF-Train}}(\mathcal{D}_{\text{cf-train}},) = \mathcal{A}_{\text{ERM}}(\mathcal{D}_{\text{cf-train}})$. It also skips the Step 3.

CDA There are multiple different ways of implementing CDA. Zuo et al. (2023) assumes access to G^* and thus skip Step 1. In Step 2, $\mathcal{D}_{\text{cf-train}}$ is augmented with $\hat{G}$. They do nothing special in Step 3.

CFE Similarly, they skip Step 1 by assuming access to G^*. In Step 2 they generate u based on $\mathcal{D}_{\text{cf-train}}$ and $\hat{G}$. In Step 3, they need to convert the input x into u via $\hat{G}$.

CFR They also skip Step 1 by assuming access to G^*. In Step 2 they generate counterfactual samples based on $\mathcal{D}_{\text{cf-train}}$ and $\hat{G}$. In Step 3, they also need to convert the input x into $x_{a'}$ via $\hat{G}$.

E Experiment Details

For a fair investigation of previous methods, most setups follow the convention in the original CFR paper (Zuo et al., 2023). We have also included the code to reproduce all results. All experiments are run on RTX A5000.

E.1 Model and Training

There are several models used in the paper. We will explain each of them separately below.

CVAE On Adult-Sim, we use the same design as that in (Zuo et al., 2023). It maps from X and A to U through an encoder in the form $u = E(x, a)$ where the dimension of U is chosen to be 7. It further maps from U to X via separate decoder. Specifically $x_\alpha = D_\alpha(u)$ and $x_\beta = D_\beta(u, a)$ where x_α and x_β represents features directly affected by A and features not directly affected by A. Here they use expert knowledge to distinguish x_α and x_β. In the case of Y being involved, the encoder is in the form $u = E(x, a, u)$ and there will be a separate decoder $y = D_Y(u, a)$. The training of CVAE follows the objective in beta-VAE (Higgins et al., 2017). When there is Y (i.e. when training G^*), there is a fairness loss that enforces Y prediction to be the same for different groups. The intuition behind this design could be found in Kim et al. (2021) and Zuo et al. (2023). The regularization term of fairness loss is 0.15.

On Law-Sim, the model design remains the same, though we change the dimension of U from 7 to 3. The reason is that the dimension of X is only 3 for this dataset. Even though it becomes 10 after one-hot encodeing, we believe it makes less sense to use such a high dimensional latent space. Besides, we remove fairness loss in the training of G^* to increase the difficulty of the task. Note that this won't cause any issue to the validness of our experiment as G^* is fixed and considered as ground truth.

ILD We employ ILD-Relax-Can on Adult-Sim and ILD-Can for Law-Sim. The sparsity is chosen as 1. The dimension of Z and U are 7 on Adult-Sim and 3 on Law-Sim, which is consistent with the

choice for CVAE. Both of encoder and decoder are composed of two-layer multi-layer perceptrons which are close to the model used for CVAE in terms of number of parameters. The training of ILD follows the objective in Beta-VAE and β is chosen to be 1.

Prediction Model For Adult-Sim, the classifer is Logistic Regression, which follows the choice in Zuo et al. (2023). For most results on Law-Sim, the regressor is Ridge Regression with Cross Valiation over [0.1,1,10,100,1000,10000]. We also investigate different choices of regressor in Appendix F.

E.2 DATASETS

For dataset, we consider UCI Adult Income (Kohavi et al., 1996) and Law School Success (Wightman, 1998) following the choice of Zuo et al. (2023). For Adult, the sensitive attribute is sex and the target is the whether the income is greater than 50K. Other features contain age, race, native country, workclass, education, martial status, occupation, relationship, hours per week. For Law, the sensitive attribute is gender and the target is first-year grade. Other features contain race, LSAT and GPA. However, since we need to evaluate TE of each method which requires access to ground truth, we use the simulated version of those datasets. Specifically, we first train a ground truth CGM $G^* = D^* \circ E^*$ that simulates the CGM of those datasets. Then we use E^* and the real X to infer U. After getting samples of U, we constructed the dataset $\mathcal{D}$ and $\mathcal{D}_{\text{counterfactual}}$ by sampling from marginal distribution of A (based on actual distribution of real data), and using D^* to get x_a and $x_{a'}$. We want to emphasize that $\mathcal{D}_{\text{counterfactual}}$, regardless of train or test set, are hidden from downstream models and used for evaluation only. This way, we get access to the ground truth u^* and can generate ground truth counterfactuals without any error. In our investigation, exogenous noise, factual data and counterfactual data are all actually the simulated version of original datasets. However they do follow a fixed data generating mechanism that is close to the real data. For clarity, we called them Adult-Sim and Law-Sim throughout the paper. For Adult-Sim, Error is evaluated as the ratio of wrong prediction while for Law-Sim, Error is evaluated as the Root Mean Squared Error (RMSE).

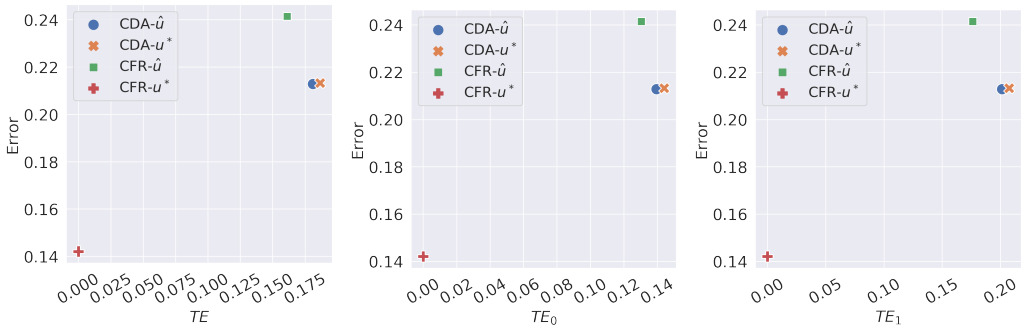
To generate simulated "ground truth" Y, the training of ground truth G^* involves Y, i.e., $x = D_X^*(u, a)$ and $y = D_Y^*(u, x)$. Regarding model structure, we implement CVAE as used in Zuo et al. (2023). In this investigation we consider two previous methods: CDA and CFR. We do not include CFE here because it only uses U. More details could be found in Appendix E. Due to space constraint, results on Adult-Sim and numbers containing TE_0 and TE_1 can be found in Appendix F. All numbers are averaged over 5 repetitive experiments.

F ADDITIONAL RESULTS

Here we include additional results on Adult-Sim and expanded figures with TE_0, TE_1 on Law-Sim. In most cases, the trend with TE_0 and TE_1 is the same as that with TE, so we won't explicitly discuss about it unless there is anything special.

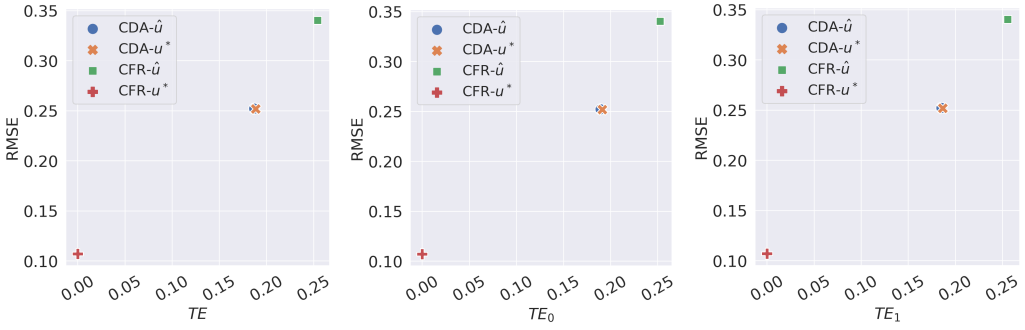
F.1 ADDITIONAL RESULTS FOR SECTION 3

Figure 6, Figure 8, Figure 10 and Figure 12 are the expanded figures of Figure 3a, Figure 3b, Figure 3c and Figure 3d respectively. For Law-Sim, in Figure 7, we observe a similar trend as that in Figure 3a. The trend in Figure 9 is similar to that in Figure 3b for CFR. One difference we observe is that as we train the model, CDA seems to lead to better performance in terms of both metric. Again, the relation between CDA-D^* and CDA-$\hat{D}_{best}$ is not clear. It's important to note that they are both much worse than CFR-D^* and even CFR-$\hat{D}$, which indicates that CDA might not fit well for this task. The trends we observe in Figure 11 and Figure 13 are also very close to that in Figure 3c and Figure 3d. Overall, our investigations on both datasets validate our claim that naively estimating counterfactuals might lead to significant failure for prior works.



(a) TE (b) TE_0 (c) TE_1

Figure 6: Error from estimating U on Adult-Sim

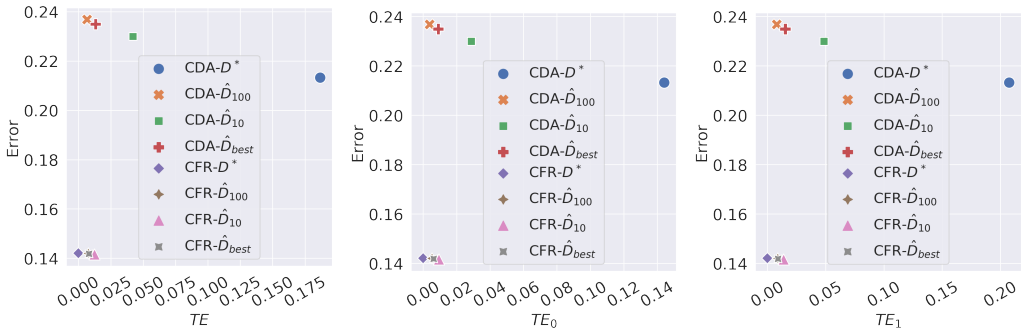


(a) TE (b) TE_0 (c) TE_1

Figure 7: Error from estimating U on Law-Sim

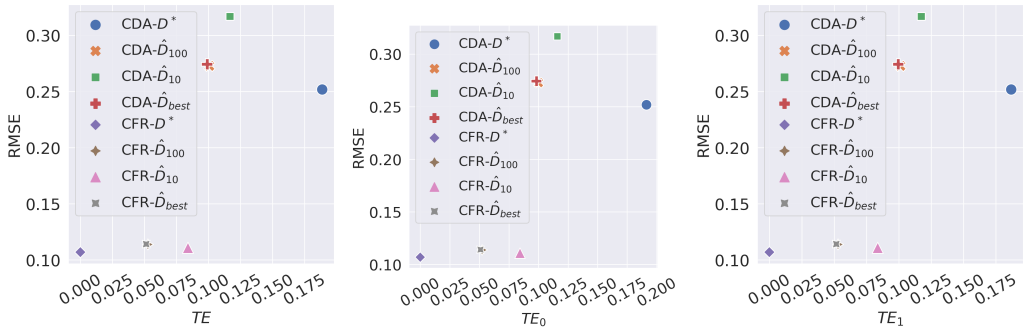
F.2 Additional results for Section 5

Figure 14 and Figure 16 are the expanded figures of Figure 4a and Figure 5a. Figure 15 and Figure 17 are the expanded figures of Figure 4b and Figure 5b. As an extra investigation, we explore how choices of different regressors would make a difference. In Figure 18, we test with Linear Regression, Ridge Regression and MLP. We notice that the trend remains similar in most cases.



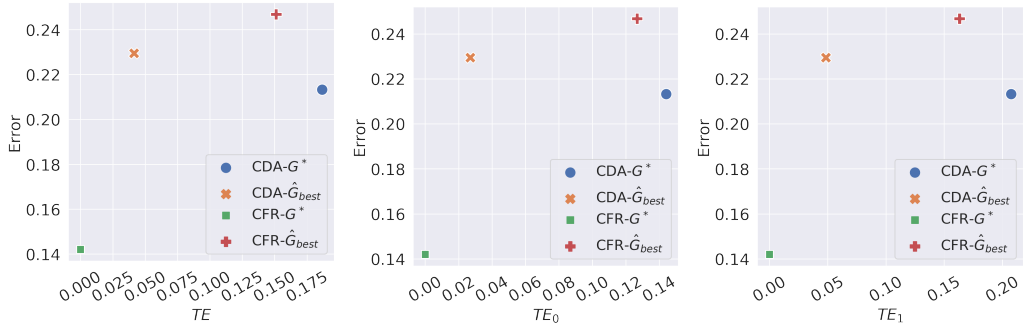
(a) TE (b) TE_0 (c) TE_1

Figure 8: Error from estimating D on Adult-Sim



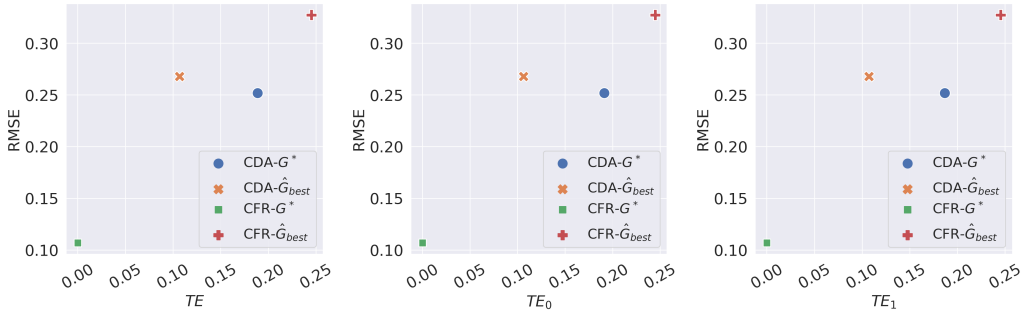
(a) TE (b) TE_0 (c) TE_1

Figure 9: Error from estimating D on Law-Sim



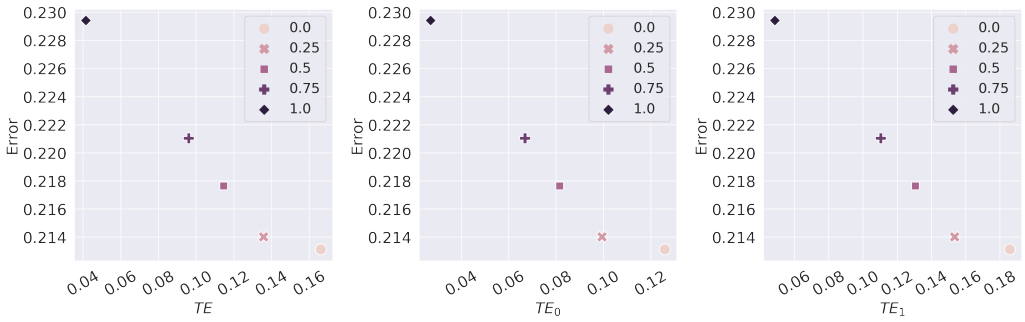
(a) TE (b) TE_0 (c) TE_1

Figure 10: Error from estimating G on Adult-Sim



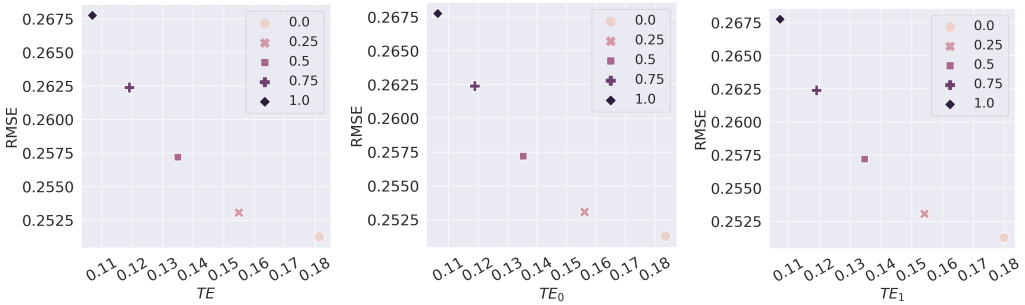
(a) TE (b) TE_0 (c) TE_1

Figure 11: Error from estimating G on Law-Sim



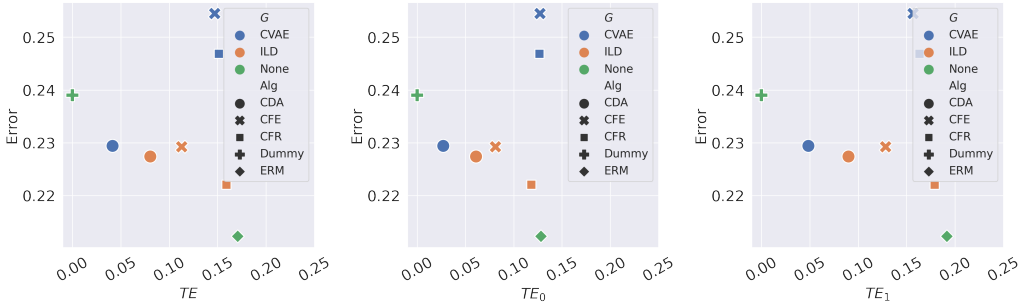
(a) TE (b) TE_0 (c) TE_1

Figure 12: Investigation of CDA-$\hat{G}_{best}$ with different number of augmented counterfactual samples on Adult-Sim.



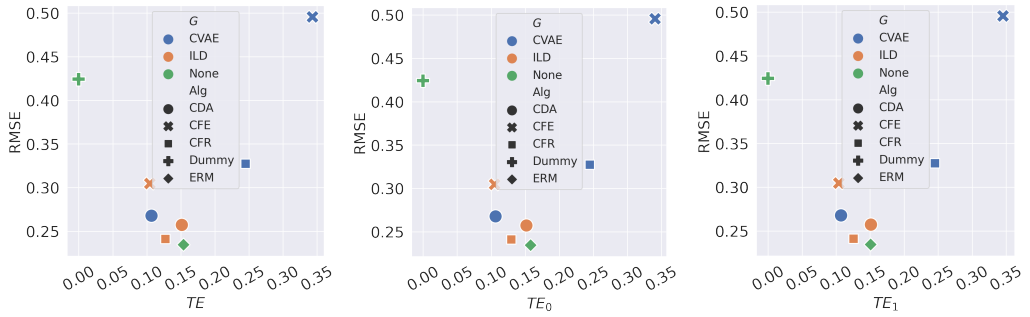
(a) TE (b) TE_0 (c) TE_1

Figure 13: Investigation of CDA-$\hat{G}_{best}$ with different number of augmented counterfactual samples on Law-Sim.



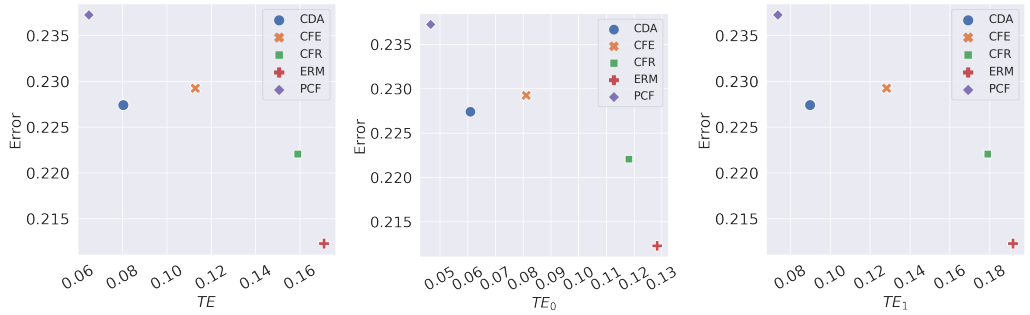
(a) TE (b) TE_0 (c) TE_1

Figure 14: Comparison between CVAE and ILD integrated with different CF algorithms on Adult-Sim.



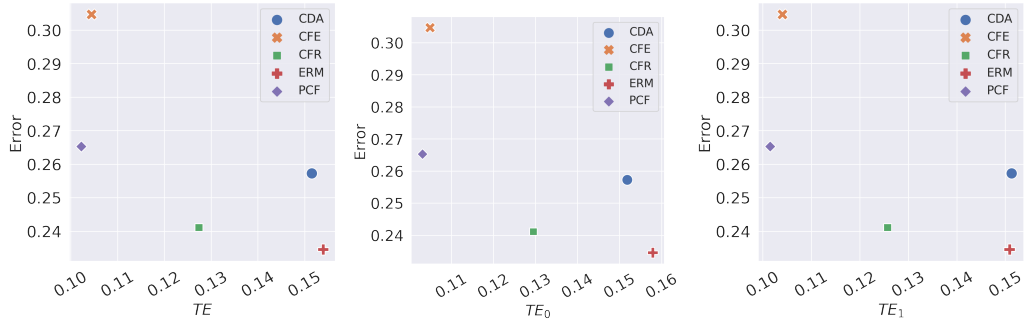
(a) TE (b) TE_0 (c) TE_1

Figure 15: Comparison between CVAE and ILD integrated with different CF algorithms on Law-Sim.



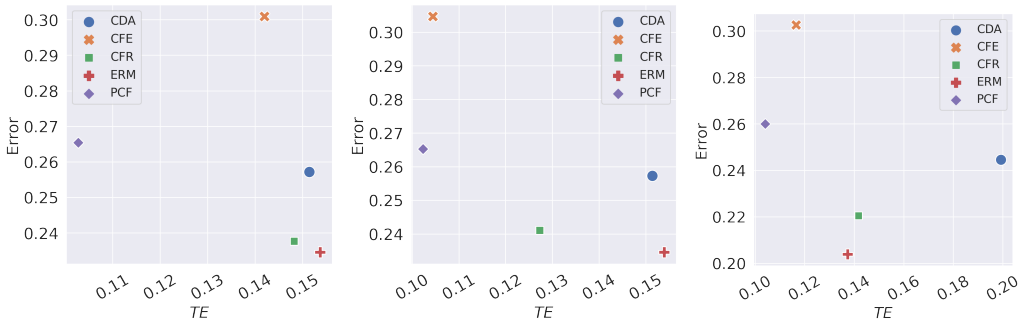
(a) TE (b) TE_0 (c) TE_1

Figure 16: Effectiveness of PCF in comparison to previous CF algortihms with estimated counterfactual samples on Adult-Sim.



(a) TE (b) TE_0 (c) TE_1

Figure 17: Effectiveness of PCF in comparison to previous CF algortihms with estimated counterfactual samples on Law-Sim.



(a) Linear Regression (b) Ridge Regression (c) MLP

Figure 18: Exploration of different regressors on Law-Sim.