Counterfactual Invariance and Fairness

Sophia Gunluk*
Mila, University of Montreal
sophia.gunluk@mila.quebec

Tejas Vaidhya*
Mila, University de Montreal
tejas.vaidhya@mila.quebec

Abstract

In machine learning, the presence of spurious correlations poses a significant obstacle to accurately learning the true relationship between input and outcome. These correlations arise from hidden associations between features and outcomes, often resulting from confounding by unobserved variables, which may lead to predictors exploiting this association when it should not have an effect on the outcome. To check for such correlations, one can perform stress tests by perturbing the irrelevant parts of the data and assessing whether there are changes in predictions across various irrelevant settings. In this report, we build off of the insights from "*Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests*" to train predictors that are invariant to such spurious associations; we reproduce their perturbative stress test experiments and extend their work to fairness for addressing algorithmic bias. By connecting the notion of counterfactual invariance with observationally-testable conditional independence criteria, we reemphasize the critical importance of understanding the true underlying causal structure of the data, and that the correct regularizer encourages counterfactual invariance. Furthermore, we explore the application of conditionally invariant regularizers in the context of fairness, which promotes specific measures of fairness based on the underlying causal structure, again implying that alignment with the correct causal model is essential for effective fairness promotion.

1 Introduction

In the rapidly evolving field of machine learning, model developers and practitioners are frequently confronted with the challenge of spurious correlations. These are apparent relationships that appear in data but lack any real underlying causal connection, meaning a corresponding predictor should be invariant to changes in the "irrelevant parts." This issue is especially significant in the context of 'black box' testing procedures for machine learning models, such as perturbative stress testing. These tests, which involve changing only the deemed "invariant" parts of the input and observing the resulting changes in the model's output, are popular due to their simplicity and intuitive appeal. However, despite their widespread use, many fundamental questions about these tests remain unanswered. For example, how does a model's performance in these tests relate to its overall performance? How can we design models that consistently pass these tests, particularly when our ability to generate perturbed examples is limited? The ad hoc nature of these tests makes it difficult to provide universally applicable answers.

In this paper, we aim to use their principles of causal inference which provide a structured understanding of what it truly means for a model to pass these stress tests. Causal inference tools play a crucial role in formalizing and examining perturbative stress tests, as demonstrated the "*Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests*". A key insight of the paper lies in finding an invariant predictor by using their connection between counterfactual invariant desiderata and practical testable conditional independence criteria based on the underlying causal

*All members contributed equally.

structure. This connection emphasizes the necessity of obtaining a comprehensive understanding of the true underlying causal structure within the data.

The original research findings align with the present paper's analysis of the semi-synthetic Amazon dataset, affirming that conditional regularization, which aligns with the anti-causal structure, effectively reduces checklist failures. These failures are quantified by the frequency of predicted label alterations caused by perturbations and the resulting mean absolute difference of predictive probabilities being nonzero. Hence, the study suggests that the correct application of regularizers can enhance model robustness against perturbations, thereby preserving counterfactual invariance.

Extending the analysis to the realm of fairness, we show that the causal regularizer is equivalent to promoting demographic parity and the anti-causal regularizer is equivalent to promoting equalized odds, which are two types of fairness metrics. We apply this analysis to training on the COMPAS dataset with counterfactual invariance, using gender and race as sensitive attributes. We found that the causal regularizer, which fits the true data-generating process of the COMPAS data, demonstrates a more stable performance compared to the anti-causal regularizer, which emphasizes the need to select the regularizer according to the correct causal structure. Furthermore, we show that the causal regularizer on causal data indeed promotes demographic parity, which can be considered a form of fairness. However, we do see that when addressing certain biases, the effectiveness of the causal regularizer may plateau beyond a certain coefficient threshold, indicating counterfactual invariance is still bottlenecked by inherent limitations within the dataset itself.

Overall, this report highlights the importance of employing the appropriate regularizer that aligns with the true underlying causal structure of the data, as well as how counterfactual invariance can promote certain types of fairness, also depending on the underlying causal structure of the data.

2 Counterfactual Invariance

2.1 Problem Set Up

To formulate the problem at hand, let us consider the task of training a predictor f which tries to predict a label Y based on a set of covariates X. In the *Counterfactual Invariance* paper, the focus lies in constructing predictors that are invariant to some well-defined perturbations applied to the covariates, specifically, "spurious" perturbations. To capture this idea, we assume there is an additional variable, Z, which contains information that should not influence the predictions, though Z may have a causal effect on X; we denote $X(z)$ as the counterfactual of X that we would have observed if Z had been set to z and kept all other variables fixed. This allows us to interpret "perturbative stress tests" as counterfactual pairs, namely $X(z)$ and $X(z')$, which differ due to an intervention on Z but should not result in changes to the predictions. We can formalize this, denoting the property *counterfactual invariance*:

Definition 1.1 A predictor f is *counterfactually invariant* to Z if $f(X(z)) = f(X(z'))$ (almost) everywhere, for all z, z' in the sample space of Z.

To derive properties of a counterfactually invariant predictor, they consider two causal structural that capture most situations that involve such spurious association between protected attribute Z and outcome Y. Specifically, they distinguish between the causal direction, when Z causes some covariates in X which cause Y, and the anti-causal direction, where Y and Z both cause the covariates. In both cases, the covariate variables X can be divided into 3 groups $X_Z^\perp$, $X_{Y \wedge Z}$, and $X_Y^\perp$; $X_Z^\perp$ is the subset of features that are independent from, or not caused by, Z, $X_Y^\perp$ is the subset of features that are independent from Y, and $X_{Y \wedge Z}$ are the features that are dependent on both Y and Z. The following diagrams capture the causal structures for each direction:

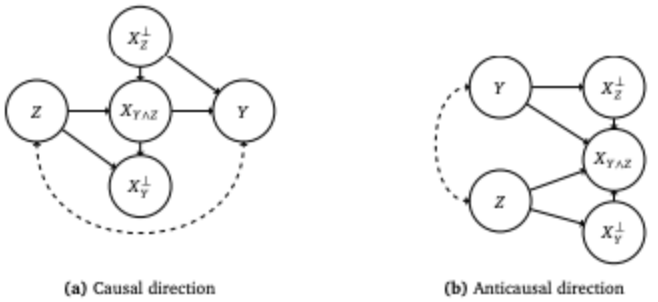

Figure 1: Causal models for the data generating process

2.2 Observable Signature

With a well-defined problem, they can now consider how to achieve counterfactual invariance in practice. The main challenge now is that counterfactual invariance is defined based on the predictor's behavior on counterfactual data, which, in practice, is impossible to simultaneously observe. However, we can overcome this issue by deriving a measurable signature of counterfactual invariance, which can be evaluated and enforced in practice using standard datasets where the variable Z is directly measured.

Intuitively, a predictor f has counterfactual invariance if it depends only on $X_Z^\perp$, which is the only part of the "inputs" X that is not causally affected by Z. First they formally define $X_Z^\perp$:

Lemma 3.1 Let $X_Z^\perp$ be a X-measurable random variable such that, for all measurable functions f, we have that f is counterfactually invariant if and only if $f(X)$ is $X_Z^\perp$-measurable. If Z is discrete then such a $X_Z^\perp$ exists.

This means our objective is to identify a predictor that solely relies on the covariates $X_Z^\perp$. To accomplish this, they leverage the causal structures in Figure 1 to derive the conditional independence relationships that lead to an invariant predictor. These relationships are testable through observed data, so they provide a signature for counterfactual invariance, depending on the underlying causal structure:

Theorem 3.2 If f is a *counterfactually invariant predictor*:

1. Under the anti-causal graph, $f(X) \perp Z \mid Y$
2. Under the causal-direction graph, if Y and Z are not subject to selection (but possibly confounded), $f(X) \perp Z$.
3. Under the causal-direction graph, if the association is purely spurious, $Y \perp X \mid X_Z^\perp, Z$, and Y and Z are not confounded (but possibly selected), $f(X) \perp Z \mid Y$.

Even though we cannot directly enforce counterfactual invariance as it would require access to counterfactual pairs, we can try to encourage a trained model to satisfy the signatures found in Theorem 3.2 by adding a corresponding regularization term to the loss. By regularizing the model to satisfy the appropriate conditional independence condition, the model in theory should be encouraged to be counterfactually invariant. For simplicity, they only consider binary Y and Z. They employ the maximum mean discrepancy (MMD) loss, which serves as a metric for comparing probability measures. Minimizing the difference between the two probability measures through MMD encourages equality and, consequently, the desired conditional independence. This gives them the (infinite data) regularization terms:

$$\begin{aligned}
\text{marginal regularization} &= \text{MMD}(P(f(X) \mid Z=0), P(f(X) \mid Z=1)) \\
\text{conditional regularization} &= \text{MMD}(P(f(X) \mid Z=0, Y=0), P(f(X) \mid Z=1, Y=0)) \\
&\quad + \text{MMD}(P(f(X) \mid Z=0, Y=1), P(f(X) \mid Z=1, Y=1)).
\end{aligned}$$

In practice, they approximate the MMD with finite data samples (2), and during training with stochastic gradient descent, they compute the penalty on each minibatch.

A key point is that the choice of regularizer depends on the underlying causal structure, as the conditional and marginal independence conditions are often incompatible. Consequently, enforcing a condition that does not align with the true causal structure may fail to promote counterfactual invariance in the model, or may throw away more information than is necessary.

Identifiability: The conditional independence signature in Theorem 3.2 serves as a necessary condition for counterfactual invariance, but it is not sufficient for two reasons. First, counterfactual invariance pertains to individual realizations of data points, whereas the signature itself is distributional in nature. Specifically, the invariance $P(f(X)|do(Z = z)) = P(f(X)|do(Z = z'))$ for all z, z' would satisfy the conditional independence signature yet it would not fit their definition of counterfactual invariance since this invariance is actually weaker as it does not require access to actual counterfactual realizations. Second, albeit unlikely, there may be cases where certain values of Z are unobservable in practice or there may be other unobserved variables that confound X and Z, which would mean the independence between $f(X)$ and Z in the training data, denoted as $f(X) \perp Z$, does not generally imply that $f(X) \perp Z$ is independent of Z and thus that Z is not a cause of $f(X)$.

3 Fairness

3.1 Demographic Parity:

We say that a predictor $\hat{Y}$ satisfies demographic parity with respect to protected attribute Z if $\hat{Y}$ and Z are independent, i.e:

$$\mathbb{P}\left(\hat{Y} = 1 \mid Z = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \mid Z = 1\right)$$

A predictor with demographic parity has equal positive, and consequently negative, prediction rates across the protected and unprotected groups.

3.2 Equalized Odds:

We say that a predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute Z and true outcome Y, if $\hat{Y}$ and Z are independent conditional on Y, i.e: (3)

$$\mathbb{P}\left(\hat{Y} = 1 \mid Y = y, Z = 0\right) = \mathbb{P}\left(\hat{Y} = 1 \mid Y = y, Z = 1\right) \quad \forall y \in \{0, 1\}$$

A predictor with equalized odds has equal true positive rates across the protected and unprotected groups (denoted by the equation where $y = 1$), and equal false positive rates across the protected and unprotected groups (denoted by the equation where $y = 0$).

3.3 Connection to Counterfactual Invariance

Veitch notes in Remark 3.3 that their notion of counterfactual invariance can be connected to fairness when considering Z to be a protected attribute, such as race or gender. Based on the equations, we can see that demographic parity is equivalent to the causal condition $f(X) \perp Z$, and equalized odds is equivalent to the anti-causal condition $f(X) \perp Z \mid Y$. To formally show this,

In the causal cause, we have $f(X) \perp Z$ or $\hat{Y} \perp Z$, and therefore $P(\hat{Y}, Z) = P(\hat{Y})P(Z)$. Thus, by starting by with a necessarily true equation, we can derive the demographic parity equation, meaning it must always hold given the causal condition:

$$
\begin{aligned}
\mathbb{P}(\hat{Y}=1) &= \mathbb{P}(\hat{Y}=1) \\
\frac{\mathbb{P}(Z=0)}{\mathbb{P}(Z=0)}\mathbb{P}(\hat{Y}=1) &= \mathbb{P}(\hat{Y}=1)\frac{\mathbb{P}(Z=1)}{\mathbb{P}(Z=1)} \\
\frac{\mathbb{P}(\hat{Y}=1)\mathbb{P}(Z=0)}{\mathbb{P}(Z=0)} &= \frac{\mathbb{P}(\hat{Y}=1)\mathbb{P}(Z=1)}{\mathbb{P}(Z=1)} \\
\frac{\mathbb{P}(\hat{Y}=1, Z=0)}{\mathbb{P}(Z=0)} &= \frac{\mathbb{P}(\hat{Y}=1, Z=1)}{\mathbb{P}(Z=1)} \\
\mathbb{P}(\hat{Y}=1 \mid Z=0) &= \mathbb{P}(\hat{Y}=1 \mid Z=1)
\end{aligned}
$$

The same can be done for the anti-causal condition $f(X) \perp Z \mid Y$ or $\hat{Y} \perp Z \mid Y$, which implies that $P(\hat{Y}, Z \mid Y=y) = P(\hat{Y} \mid Y=y)P(Z \mid Y=y)$ for all $y \in \{0,1\}$. We can similarly derive the equalized odds equation from an inherently true equation:

$$
\begin{aligned}
\mathbb{P}(\hat{Y}=1 \mid Y=y) &= \mathbb{P}(\hat{Y}=1 \mid Y=y) \\
\frac{\mathbb{P}(Z=0 \mid Y=y)}{\mathbb{P}(Z=0 \mid Y=y)}\mathbb{P}(\hat{Y}=1 \mid Y=y) &= \mathbb{P}(\hat{Y}=1 \mid Y=y)\frac{\mathbb{P}(Z=1 \mid Y=y)}{\mathbb{P}(Z=1 \mid Y=y)} \\
\frac{\mathbb{P}(\hat{Y}=1 \mid Y=y)\mathbb{P}(Z=0 \mid Y=y)}{\mathbb{P}(Z=0 \mid Y=y)} &= \frac{\mathbb{P}(\hat{Y}=1 \mid Y=y)\mathbb{P}(Z=1 \mid Y=y)}{\mathbb{P}(Z=1 \mid Y=y)} \\
\frac{\mathbb{P}(\hat{Y}=1, Z=0 \mid Y=y)}{\mathbb{P}(Z=0 \mid Y=y)} &= \frac{\mathbb{P}(\hat{Y}=1, Z=1 \mid Y=y)}{\mathbb{P}(Z=1 \mid Y=y)} \\
\mathbb{P}(\hat{Y}=1 \mid Z=0, Y=y) &= \mathbb{P}(\hat{Y}=1 \mid Z=1, Y=y)
\end{aligned}
$$

This implies that in the context of fairness, if we want to prevent the predictor from exploiting spurious associations from protected attributes, then counterfactual invariance either guarantees demographic parity or equalized odds, depending on which regularizer we use and whether it is appropriate for the underlying causal structure. One insight this can provide is when we should apply each fairness metric since they oftentimes contradict one another. This also brings up a question of what it means when both demographic parity and equalized odds hold; the underlying structural causal model would be implied to be both causal and anti-causal, though this is not possible.

4 Experiments

We reproduce experiments from paper, *"Counterfactual Invariance to Spurious Correlations"*. We also apply their notion of counterfactual invariance to fairness with the COMPAS dataset and analyze the two aforementioned fairness measures of the resulting predictor.

4.1 Amazon Experiments

In our research, we specifically replicate a synthetic confounding experiment from the original paper titled Counterfactual Invariance to Spurious Correlations. Our objective is to corroborate the assertion made in the referenced work (8), which posits that "Regularizing conditional MMD improves counterfactual invariance on synthetic anticausal data".

4.1.1 Synthetic Dataset

Synthetic counterfactuals in product review Data: We employ a Synthetic counterfactuals setup as outlined in the study by Vietch V et al. (8). In this synthetic confounding experiment, each review is associated with a Bernoulli random variable, denoted as Z. A perturbation is then introduced into

the review text, referred to as X, which effectively transforms common words such as "the" and "a" into carriers of information about Z. For instance, when Z equals 1, the token "the" is substituted with the token "thexxxxx". Comprehensive details regarding the data generation process are provided in the appendix.

The review score, denoted as Y, is taken as the dependent variable, and a subsampling process is performed to ensure a balanced distribution of Y. The data exhibits an anti-causal structure, wherein the text X is designed to explicate the score Y. It is hypothesized that any association between Y and Z is purely coincidental, stemming from the fact that the words "the" and "a" carry minimal information concerning the label.

The models are trained on a dataset where the probability of Y equalling Z, denoted as $P(Y = Z)$, is fixed at 0.3. To evaluate the resilience of the models, we follow the perturbed stress-test datasets as describe in the paper. This is achieved by transforming each instance $X_i(z)$ to its counterfactual $X_i(1 - z)$, as prescribed by the synthetic model (8).

By evaluating the performance of each model on the perturbed dataset, we assess whether the distributional properties imposed by the regularizers result in counterfactual invariance at the instance level. This experiment provides insights into the effectiveness of the regularizers in maintaining the stability of the model outcomes under perturbations.

4.1.2 Model and Implementation Details

We implemented our own codebase using Pytorch (7) and HuggingFace transformers (9). We use BERT (1) as the base model for all experiments in the Amazon review section, specifically employing the 'bert_en_uncased_L-12_H768_A-12' configuration from the Huggingface Hub (9) with no parameter alterations. Following standard procedures, a linear transformation is applied to the representation layer for making predictions. we train identical architectures using CrossEntropy + $\lambda \cdot$ Regularizer as the objective function, where we vary λ and take Regularizer as either the marginal penalty, or conditional penalty. Training is conducted using stochastic gradient descent with a batch size of 80 and a learning rate set to 1e-5. To mitigate overfitting, we implement early stopping on validation risk with a patience of 10. The models were trained on a single A100 GPU. Our codebase is flexible enough to support all the model architecture from huggingface universe without any change in the codebase.

For the MMD regularizer, we employ the estimator developed by Gretton et al. (2), using the Gaussian RBF kernel. We set the kernel bandwidth at 10.0.

4.1.3 Results

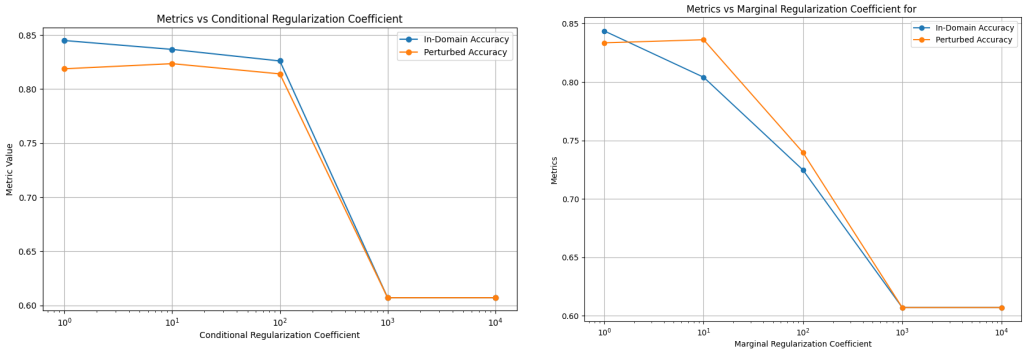

Figure 2: Comparison of In-Domain and Perturbed Accuracy as a function of Regularization Coefficients. As the regularization coefficient increases, both the in-domain and perturbed accuracy decrease, indicating the model's performance is sensitive to the regularization strength.

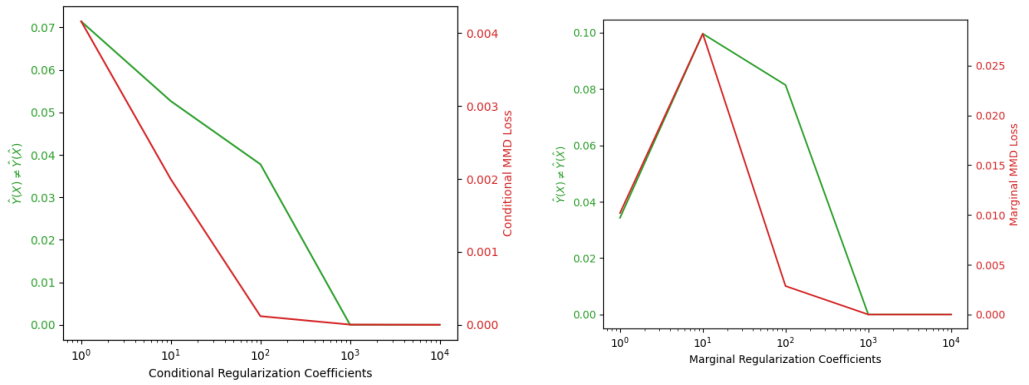

Figure 3: The generated plot represents the relationship between regularization coefficients and two metrics: label flip rate and the Maximum Mean Discrepancy (MMD) test.

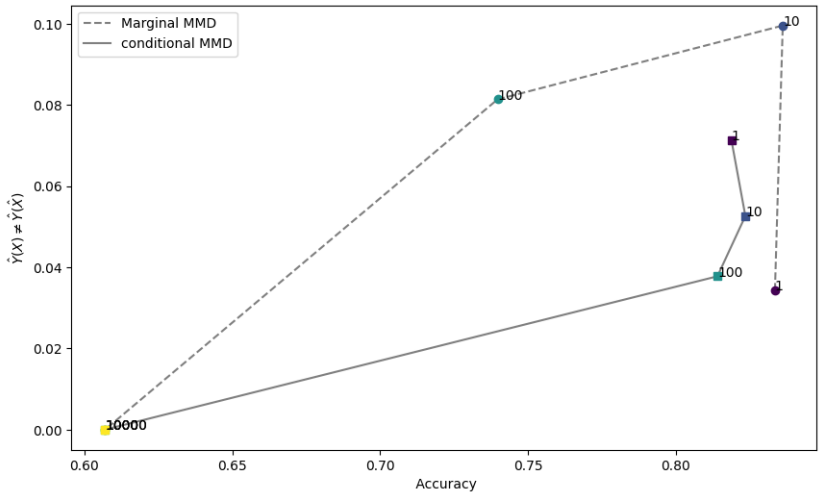

Figure 4: MMD regularization reduces the rate of predicted label flips on perturbed data, with little affect on in-domain accuracy

In-Domain Accuracy vs Regularization Coefficient: The In-Domain Accuracy of the model decreases as the regularization coefficient increases. This suggests that higher regularization leads to a simpler model which may not capture the complexity of the training data as effectively, thereby reducing its performance on in-domain tasks.

Perturbed Accuracy vs Regularization Coefficient: Initially, as the regularization coefficient increases from 1.0 to 10.0, the Perturbed Accuracy of the model shows a slight increase. This suggests that a certain degree of regularization may help the model to generalize better and handle perturbations more effectively. However, as the regularization coefficient continues to increase beyond 10.0, Perturbed Accuracy also starts to decline, reflecting the model's reduced ability to handle complexity.

Label Flip Rate: The Label Flip Rate initially increases with the regularization coefficient, indicating that higher regularization may lead to more instability in the model's predictions. However, when the regularization coefficient is extremely high (1000.0 and 10000.0), the Label Flip Rate reduces to zero, implying the model's output becomes invariant, regardless of input perturbations.

Conditional MMD Test and Perturbed Test: These metrics are very low for all values of the regularization coefficient, suggesting that the model's predictions are relatively consistent across different instances of the same input. However, it should be noted that these values increase slightly

when the regularization coefficient is extremely high (1000.0 and 10000.0), indicating a slight decrease in consistency.

In summary, it seems that there is a trade-off in the choice of the regularization coefficient. The results suggest that a moderate level of regularization (in this case, around 10.0) might be the most effective. However, the optimal level of regularization likely depends on the specific context and requirements of the task at hand.

4.2 COMPAS Experiments

We then extend their work on counterfactual invariance to algorithmic fairness by applying the aforementioned regularization to training on COMPAS data. Our objective is to find a counterfactually invariant predictor with respect to protected attributes such as gender and race, which should imply a "fair" predictor. Since the outcome of whether a person recidivates is thought to be a result of an individual's characteristics, which are the input features, we consider the data-generating process to be causal rather than anti-causal.

4.2.1 COMPAS Background

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm is a machine-learning tool, used by the courts in multiple U.S. states, to predict the likelihood of recidivism for criminal offenders, which they use to determine pretrial release, sentencing, and parole. In 2016, ProPublica, a non-profit investigative newsroom, published an article concluding that the algorithm was biased against black defendants, as it was more likely to falsely flag them as having a higher risk of re-offending than white defendants (6). Though there were previous questions about the potential bias of the algorithm, this confirmed concerns about the fairness of blindly using machine learning models in criminal justice decisions.

This can be related to counterfactual invariance in the context of fairness; the issue with the COMPAS algorithm is that it seems to exploit spurious correlations between sensitive features (such as race) and the outcome due to historic and systemic biases. This causes the model to output two very different scores even when the rest of the attributes are held the same except for race, which to an analyst looks unfair or discriminatory.

4.2.2 Model and Implementation Details

Though we do not have access to the exact details of the COMPAS algorithm, we can train a basic MLP model without any adjustment for counterfactual invariance and use it as a baseline since it achieves similar accuracy as the reported accuracy of the COMPAS algorithm. For our experiments, we use a pre-processed (filtered and binarized) data set that ProPublica released as part of their studies (5) and trained an MLP model to predict whether the individual would recede or not, where the loss is the sum of the cross entropy loss and the weighted MMD loss of the marginal/conditional probabilities, based on the regularizer and its coefficient. To derive the MMD loss, we employ the same estimator developed by Gretton et al. (2), using the Gaussian RBF kernel with bandwidth set to 10.0. The specific details of the MLP model and training are in the appendix (Section 7).

For fairness metrics, we consider the difference between the probabilities so the demographic parity metric is defined as

$$\mathbb{P}\left(\hat{Y} = 1 \mid Z = 0\right) - \mathbb{P}\left(\hat{Y} = 1 \mid Z = 1\right)$$

and equalized odds metric is defined as

$$\frac{1}{2} \sum_{y \in \{0,1\}} \left[\mathbb{P}\left(\hat{Y} = 1 \mid Z = 0, Y = y\right) - \mathbb{P}\left(\hat{Y} = 1 \mid Z = 1, Y = y\right)\right]$$

To collect data for the graphs, we average the results (loss, accuracy, and fairness metrics) of 10 models trained with each regularizer coefficient, and we tested about 70-80 coefficients in total.

4.2.3 Results

Gender: Though ProPublica does not focus on gender, it is a commonly used sensitive attribute for stress tests on models. We applied the counterfactual invariance algorithm with the sensitive

attribute Z as the binary variable representing whether or not the individual was female, and found the following results:

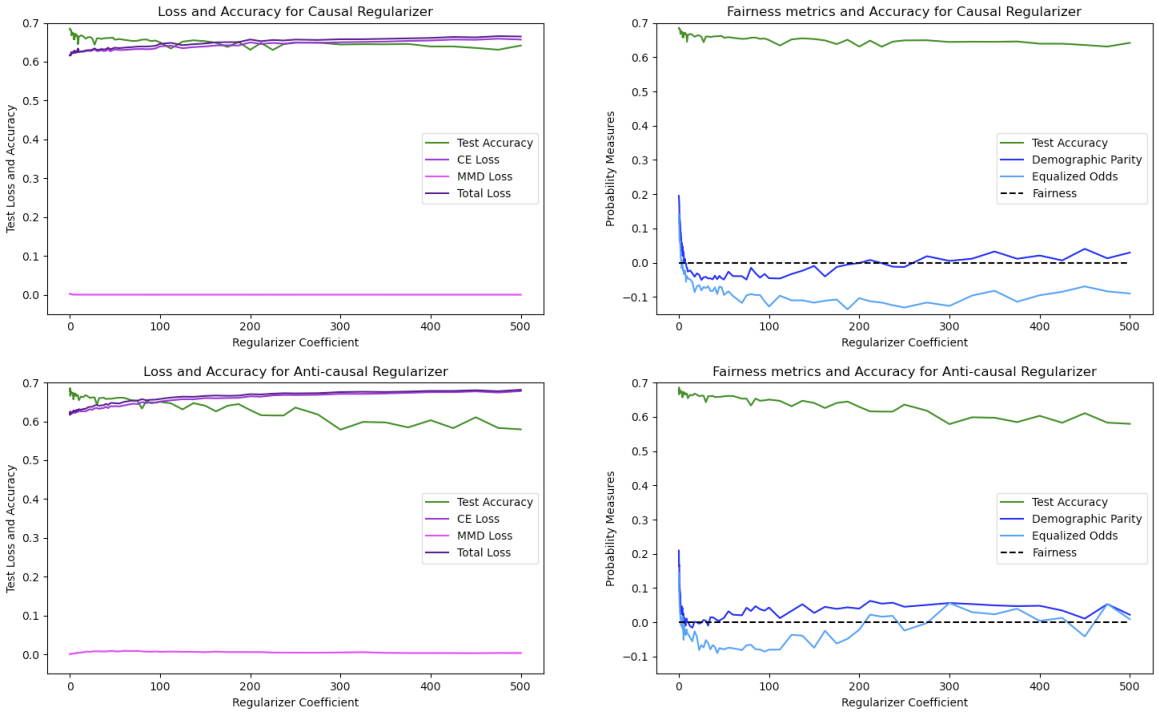

Figure 5: COMPAS Experiments with the sensitive attribute Gender

When the causal regularizer is applied, we observe that increasing the coefficient has minimal impact on the performance of the predictor, though it dies slightly decrease. In contrast, the anti-causal regularizer exhibits a more pronounced decrease in performance as the coefficient is increased. This suggests that the anti-causal regularizer leads the model to become a trivial predictor that makes arbitrary guesses. This finding aligns with our intuition that applying an incorrect regularizer based on the underlying structural causal model can incentivize the model to sacrifice predictive accuracy in favor of satisfying the wrong regularizer. Conversely, using the correct regularizer yields more consistent performance.

In terms of fairness, the demographic parity and equalized odds of the unregularized model (i.e., with a coefficient of 0) both exhibit positive and relatively high values. This indicates a significant disparity in prediction outcomes between males and females; specifically, the unregularized model has a higher likelihood of predicting males as recidivists and its true positive rate is higher for males compared to females, and thus the false negative rate is higher for females than males. However, as we increase the coefficient, which is the weight of the regularizer, the demographic parity moves toward 0. In contrast, equalized odds initially decrease but then somewhat stabilize, presumably because the predictor the model finds that satisfies demographic parity has a higher true positive rate for females than males. This validates our earlier intuition that the causal regularizer effectively promotes demographic parity, i.e., minimizing the disparity, which also happens to naturally influence the equalized odds metric. Conversely, the anti-causal regularizer leads to both the equalized odds and demographic parity metrics approaching 0 as the coefficient increases. This outcome likely stems from the fact that, as previously mentioned, the only models satisfying the anti-causal regularizer are relatively trivial predictors that exhibit poor accuracy.

Race: The main focus of ProPublica's findings was the racial bias in the COMPAS algorithm, so we apply the same idea of counterfactual invariance with race as the sensitive attribute. Specifically, we assigned 1, the protected group, to represent Black individuals and 0, the unprotected group, to be all other races.

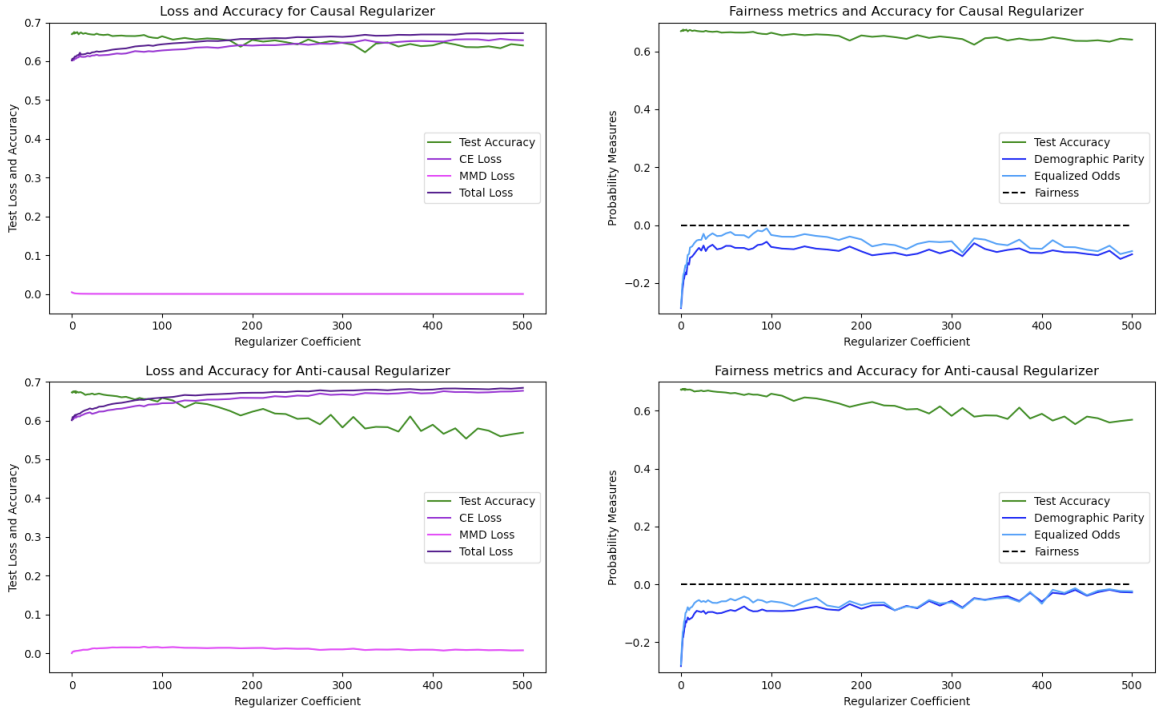

Figure 6: COMPAS Experiments with the sensitive attribute Race

Like the previous example, when we apply the causal regularizer, increasing the coefficient has minimal impact on the performance of the predictor, again with a slight decrease. In contrast, the performance of models trained with the anti-causal regularizer more significantly decreases as the coefficient is increased, which again suggests that the anti-causal regularizer leads the model to become a trivial predictor. This finding similarly confirms our intuition that models trained with the correct regularizer, according to the underlying structural causal model, yield relatively consistent performance as we increase the coefficient whereas applying an incorrect regularizer can incentivize the model to satisfy the incorrect regularizer while forfeiting its predictive accuracy.

However, this example differs a bit more in the context of fairness. The demographic parity and equalized odds of the unregularized model (i.e., with a coefficient of 0) both exhibit relatively high and negative values this time, implying a difference in prediction outcomes between Black individuals and others; specifically, the unregularized model has a higher likelihood of predicting Black individual will recidivate, and its true positive rate is higher for Black individuals compared to other individuals, and thus the false negative rate is higher for other individuals than black individuals. This agrees with Propublica's findings that an unregularized model will result in a biased predictor, most likely due to historic and systemic biases that are reflected in the data. However, as we increase the coefficient, which is the weight of the regularizer, the demographic parity jumps toward 0 but then remains relatively constant. This implies that with the data provided, it is not possible to find an accurate predictor that gets the demographic parity metric closer to 0 beyond about -0.1. In contrast, equalized odds metric initially drastically increases and then stabilizes as we increase the coefficient, presumably because the predictor that satisfies demographic parity (or does as best as it can) has a smaller difference between true positive and false negative rates between Black individuals and the other individuals. This supports our prior intuition that the causal regularizer promotes a predictor with demographic parity, which just happens to naturally influence the equalized odds of the predictor.

Conversely, the anti-causal regularizer leads to both the equalized odds and demographic parity metrics approaching 0 as the coefficient increases, and the accuracy decreases quite significantly, like before. Similarly, this outcome likely stems from the fact that the only predictors the model can find that satisfy the anti-causal regularizer are relatively trivial predictors with poor accuracy. This is interesting because, with the causal regularizer, we were able to find a predictor whose equalized odds

metric was very close to 0 (about 0.02), but when we apply the anti-causal regularizer, it does not find the same predictor with high accuracy. This could be due to the fact the model cannot properly traverse the gradient of the incorrect regularizer since it does not match the data-generating process, while the appropriate regularization term encourages the model to correctly converge to a predictor that actually learns from the data while regularizing.

5 Conclusion

In this paper, we utilized the tools of causal inference to formalize and scrutinize perturbative stress tests. A primary insight of our work is the connection between counterfactual desiderata and observationally-testable conditional independence criteria. Such a connection necessitates a thorough understanding of the true underlying causal structure of the data. Our analysis of the COMPAS algorithm, applying counterfactual invariance to gender and race as sensitive attributes, leads to key conclusions that reinforce the importance of this concept in addressing algorithmic bias.

5.1 Counterfactual Invariance Paper

Our findings matches with the original paper, indicate that conditional regularization, which aligns with the anti-causal structure, leads to a reduction in checklist failures. Checklist failures are quantified by the frequency of the predicted label altering due to the perturbation and the mean absolute difference in predictive probabilities that the perturbation incites. Therefore, our study suggests that the application of regularizers can enhance the robustness of models against perturbations, thereby ensuring the preservation of counterfactual invariance.

There is a trade-off in the choice of the regularization coefficient. Too low a value may not sufficiently prevent overfitting, leading to poorer performance on perturbed data. Too high a value, on the other hand, may lead to underfitting, reducing performance on both in-domain and perturbed data. The results suggest that a moderate level of regularization (in this case, around 10.0) might be the most effective. However, the optimal level of regularization likely depends on the specific context and requirements of the task at hand.

5.1.1 Application to Fairness

From our comprehensive analysis of the COMPAS algorithm with counterfactual invariance using gender and race as sensitive attributes, we draw several significant conclusions.

Our evaluation of the COMPAS algorithm led to significant conclusions. The causal regularizer yielded more stable performance than the anti-causal regularizer, underscoring the need to align the regularizer with the correct causal model. The causal regularizer also promoted fairness, as indicated by demographic parity and equalized odds. However, in the case of racial bias, the effectiveness of the causal regularizer plateaued beyond a certain threshold, suggesting inherent limitations within the dataset.

In conclusion, we underscore the importance of using suitable regularizers that align with the true underlying causal structure of the data in machine learning models. This practice not only sustains the model's predictive performance but also significantly enhances fairness, aiding in the mitigation of bias. Nevertheless, it is also essential to recognize that even the most appropriate regularizer may face limitations in fully eliminating disparities, especially when dealing with data reflecting deep-rooted societal biases. Thus, the pursuit of algorithmic fairness must be a multi-faceted approach, combining robust model techniques, careful scrutiny of data sources, and continuous monitoring of algorithmic outcomes.

6 Acknowledgements

We express our sincere gratitude to our course instructor, Prof. Dhanya Sridhar, for her invaluable guidance, constructive discussions, and unwavering support throughout this project.

We also extend our appreciation to the MILA Cluster team for their assistance in facilitating the running of this experiment on the MILA Cluster. Their help was crucial in the smooth progress of our project.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[2] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL: http://jmlr.org/papers/v13/gretton12a.html.

[3] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016. URL: https://arxiv.org/abs/1610.02413.pdf, arXiv:1610.02413.

[4] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.

[5] Jeff Larson. Compas analysis. https://github.com/propublica/compas-analysis, 2017.

[6] Jeff Larson, Julia Angwin, Lauren Kirchner, and Surya Mattu. How we analyzed the compas recidivism algorithm, May 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[7] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. arXiv:1912.01703.

[8] Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests, 2021. arXiv:2106.00545.

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. arXiv:1910.03771.

7 Appendix

7.1 Implementation details of Synthetic Data set

In order to generate data, We preprocess the data by selecting a subset of the reviews and converting the review score into a binary format (1 for positive reviews with scores 4 or 5, and 0 for all others). We also truncate each review text to the first 20 words to simplify the problem.

Next, we create an additional binary variable, Z, following a Bernoulli distribution. This variable is used to control the modification of the review text. Specifically, we replace the words 'and' and 'the' in the review text, adding unique suffixes based on the value of Z. This process serves to create distinct versions of the review text.

In the subsequent step, we induce an association between the outcome variable (Y) and the introduced binary variable (Z). This is achieved through data resampling, ensuring that the proportion of positive reviews is higher when Z=1 and lower when Z=0, or vice versa, based on a defined parameter, gamma.

7.2 Results on Synthetic Counterfactuals in product review data

In-Domain Accuracy: This is a measure of how well a model performs on the specific dataset it was trained on, also known as the training set.

Perturbed Accuracy: This is a measure of how well the model performs on a perturbed dataset, which is a dataset that has been slightly modified or 'noised' in some way. It is used to evaluate the robustness of the model to slight changes in the input data. If a model has high perturbed accuracy, it suggests that the model is robust and can generalize well to new or slightly different data.

Regularization Coefficient	In-Domain Accuracy	Perturbed Accuracy	Label Flip Rate	Marginal MMD Test	Marginal MMD Perturbed Test	Probability Difference
1.0	0.843	0.833	0.034	0.019	0.010	0.005
10.0	0.804	0.836	0.099	0.001	0.028	0.014
100.0	0.724	0.739	0.081	0.001	0.003	0.003
1000.0	0.607	0.607	0.0	2.145e-07	2.384e-07	0.0
10000.0	0.607	0.607	0.0	1.549e-07	4.053e-07	0.0
Regularization Coefficient	In-Domain Accuracy	Perturbed Accuracy	Label Flip Rate	Conditional MMD Test	Conditional MMD Perturbed Test	Probability Difference
1.0	0.84486	0.81875	0.07138	0.00415	0.00856	-0.01805
10.0	0.83666	0.82347	0.05263	0.00199	0.00381	-0.00986
100.0	0.82597	0.81402	0.03777	0.00011	0.00019	-0.00277
1000.0	0.60708	0.60708	0.0	0.000003	0.000004	0.0
10000.0	0.60708	0.60708	0.0	0.000000083	0.000000083	0.0

Table 1: Comparison of model performance metrics at different regularization coefficients

Label Flip Rate: This is a measure of how often the predicted label of an instance changes when the sensitive attribute is flipped. A high label flip rate indicates that the model's predictions are significantly influenced by the sensitive attribute, suggesting potential bias.

Probability Difference: This is a measure of the difference in the probabilities of a certain prediction between the two groups defined by the sensitive attribute.

7.3 Implementation details of COMPAS training

The following table contains details about various hyperparameters of the training process and MLP model being trained.

Figure 7: MLP Training Information

hyper-parameter	**value**	**hyper-parameter**	**value**
epochs	10	**hidden layers**	$(100, 30, 10)$
batch size	256	**training set**	0.65
learning rate	10^{-3}	**validation set**	0.15
optimizer	Adam (4)	**test set**	0.20

7.4 Released Code Details

We release our code under MIT License at: https://github.com/tejasvaidhyadev/Causal_Inference_Project. The model's weights, data, and other dependencies required for the experiment are at https://github.com/tejasvaidhyadev/Causal_Inference_Project/releases.

Our code is designed for academic research and has been made publicly available for others to use for their own research or other purposes. As per our knowledge, it is the only public implementation of paper. (8). The code README file on the GitHub repository provides detailed instructions on how to reproduce the results.