A Proofs

Theorem 1. *Let $\mathcal{F}^{CF}$ be the set of all counterfactually invariant predictors. Let ℓ be a proper scoring rule (e.g., square error, cross entropy loss). Let the counterfactually fair predictor that minimizes risk on the training distribution $X, Y, A \sim P$ be:*

$$f^*(X) := \operatorname*{argmin}_{f \in \mathcal{F}^{CF}} \mathbb{E}_P[\ell(f(X), Y)]$$

Then, f^* also minimizes risk on the target distribution $X, Y, A \sim Q$ with no selection effects, i.e.,

$$f^*(X) = \operatorname*{argmin}_{f} \mathbb{E}_Q[\ell(f(X), Y)]$$

if either of the following conditions hold:

1. *The association between Y and A is due to selection on label and the marginal distribution of the label Y is the same in each distribution, i.e., $P(Y) = Q(Y)$.*
2. *The association between Y and A is due to selection on predictors.*

Proof. Counterfactual fairness is a case of counterfactual invariance. By Lemma 3.1 in Veitch et al. [47], this implies X is $X_A^\perp$-measurable. Therefore,

$$\operatorname*{argmin}_{f \in \mathcal{F}^{\text{CF}}} \mathbb{E}_P[\ell(f(X), Y)] = \operatorname*{argmin}_{f \in \mathcal{F}^{\text{CF}}} \mathbb{E}_P[\ell(f(X_A^\perp), Y)]$$

Following the same reasoning as Theorem 4.2 in Veitch et al. [47], it is well-known that under squared error or cross entropy loss the risk minimizer is $f^*(x_A^\perp) = \mathbb{E}_P[Y \mid x_A^\perp]$. Because the target distribution Q has no selection (and no confounding because A is exogenous in the case of counterfactual fairness), the risk minimizer in the target distribution is the same as the counterfactually fair risk minimizer in the target distribution, i.e., $\mathbb{E}_Q[Y \mid x] = \mathbb{E}_Q[Y \mid x_A^\perp]$. Thus our task is to show $\mathbb{E}_P[Y \mid x_A^\perp] = \mathbb{E}_Q[Y \mid x_A^\perp]$.

Selection on label is shown in Figure 2b. Because $X_A^\perp$ does not d-separate Y and A, $f^*(X)$ depends on the marginal distribution of Y, so we need an additional assumption that $P(Y) = Q(Y)$. We can use this with Bayes' theorem to show the equivalence of the conditional distributions,

$$\begin{aligned}
Q(Y \mid X_A^\perp) &= \frac{Q(X_A^\perp \mid Y)Q(Y)}{\int Q(X_A^\perp \mid Y)Q(Y)\mathrm{d}y} && \text{(A.1)}\\
&= \frac{P(X_A^\perp \mid Y)Q(Y)}{\int P(X_A^\perp \mid Y)Q(Y)\mathrm{d}y} && \text{(A.2)}\\
&= \frac{P(X_A^\perp \mid Y)P(Y)}{\int P(X_A^\perp \mid Y)P(Y)\mathrm{d}y} && \text{(A.3)}\\
&= P(Y \mid X_A^\perp), && \text{(A.4)}
\end{aligned}$$

where the first and fourth lines follow from Bayes' theorem, the second line follows from the causal structure (X causes Y), and the third line follows from the assumption that $P(Y) = Q(Y)$. This equality of distributions implies equality of expectations.

Selection on predictors is shown in Figure 2c. Because $X_A^\perp$ d-separates Y and A, $f^*(X)$ does not depend on the marginal distribution of Y, so we immediately have an equality of conditional distributions, $Q(Y \mid X_A^\perp) = P(Y \mid X_A^\perp)$, and equal distributions have equal expectations, $\mathbb{E}_P[Y \mid x_A^\perp] = \mathbb{E}_Q[Y \mid x_A^\perp]$. □

Theorem 2. *Let the causal structure be known and represented as a faithful causal DAG with $X_Y^\perp$, $X_A^\perp$, Y, and A, such as in Figure 2, then:*

1. *Counterfactual fairness is equivalent to* **demographic parity** *if and only if there is no unblocked path between $X_A^\perp$ and A.*
2. *Counterfactual fairness is equivalent to* **equalized odds** *if and only if all paths between $X_A^\perp$ and A, if any, are either blocked by a variable other than Y or unblocked and contain Y.*

(a) Example of an unblocked path between A and $X_A^\perp$ (b) Example of a blocked path between A and $X_A^\perp$

Figure 3: Examples of an unblocked path and a blocked path in a causal DAG.

3. *Counterfactual fairness is equivalent to* **calibration** *if and only if all paths between Y and A, if any, are either blocked by a variable other than $X_A^\perp$ or unblocked and contain $X_A^\perp$.*

Proof. For a predictor $f(X)$ to be counterfactually fair, it must only depend on $X_A^\perp$.

1. By Definition 1, demographic parity is achieved if and only if $X_A^\perp \perp A$. In a DAG, variables are dependent if and only if there is an unblocked path between them (i.e., no unconditioned **collider** in which two arrows point directly to the same variable). For example, Figure 3a shows an unblocked path between A and $X_A^\perp$.

2. By Definition 2, equalized odds are achieved if and only if $X_A^\perp \perp A \mid Y$. If there is no path between $X_A^\perp$ and A, then $X_A^\perp$ and A are independent under any conditions. If there is a blocked path between $X_A^\perp$ and A, and it has a block that is not Y, then $X_A^\perp$ and A remain independent because a blocked path induces no dependence. If there is an unblocked path between $X_A^\perp$ and A that contains Y, then Y d-separates $X_A^\perp$ and A, so $X_A^\perp$ and A remain independent when controlling for Y. On the other hand, if there is a blocked path between $X_A^\perp$ and A and its only block is Y, then controlling for the block induces dependence. If there is an unblocked path that does not contain Y, then $X_A^\perp$ and A are dependent.

3. With Definition 3, we can apply analogous reasoning to the case of equalized odds. Calibration is achieved if and only if $Y \perp A \mid X_A^\perp$. If there is no path between Y and A, then Y and A are independent under any conditions. If there is a blocked path between Y and A, and it has a block is not $X_A^\perp$, then Y and A remain independent because a blocked path induces no dependence. If there is an unblocked path between Y and A that contains $X_A^\perp$, then $X_A^\perp$ d-separates Y and A, so Y and A remain independent when controlling for $X_A^\perp$. On the other hand, if there is a blocked path between Y and A and its only block is $X_A^\perp$, then controlling for the block induces dependence. If there is an unblocked path that does not contain $X_A^\perp$, then Y and A are dependent.

□

Corollary 2.1. *If the causal context is known and represented as a causal DAG with $X_Y^\perp$, $X_A^\perp$, Y, and A, then:*

1. *For a binary classifier, counterfactual fairness is equivalent to* **conditional demographic parity** *if and only if, when a set of legitimate factors L is held constant at level l, there is no unblocked path between $X_A^\perp$ and A.*

2. *For a binary classifier, counterfactual fairness is equivalent to* **false positive error rate balance** *if and only if, for the subset of the population with negative label (i.e., $Y = 0$), there is no path between $X_A^\perp$ and A, a path blocked by a variable other than Y, or an unblocked path that contains Y.*

3. *For a binary classifier, counterfactual fairness is equivalent to* **false negative error rate balance** *if and only if, for the subset of the population with positive label (i.e., $Y = 1$), there is no path between $X_A^\perp$ and A, a path blocked by a variable other than Y, or an unblocked path that contains Y.*

4. *For a probabilistic classifier, counterfactual fairness is equivalent to* **balance for negative class** *if and only if, for the subset of the population with negative label (i.e., $Y = 0$), there is no path between $X_A^\perp$ and A, a path blocked by a variable other than Y, or an unblocked path that contains Y.*

5. *For a probabilistic classifier, counterfactual fairness is equivalent to* **balance for positive class** *if and only if, for the subset of the population with positive label (i.e., $Y = 1$), there is no path between $X_A^\perp$ and A, a path blocked by a variable other than Y, or an unblocked path that contains Y.*

6. *For a probabilistic classifier, counterfactual fairness is equivalent to* **predictive parity** *if and only if, for the subset of the population with positive label (i.e., $D = 1$), there is no path between Y and A, a path blocked by a variable other than $X_A^\perp$, or an unblocked path that contains $X_A^\perp$.*

7. *For a probabilistic classifier, counterfactual fairness is equivalent to* **score calibration** *if and only if there is no path between Y and A, a path blocked by a variable other than $X_A^\perp$, or an unblocked path that contains $X_A^\perp$.*

Proof. Each of these seven metrics can be stated as a conditional independence statement, as shown in Table 1, and each of the seven graphical tests of those statements can be derived from one of the three graphical tests in Theorem 2. Note that the graphical test for a binary classifier is the same as that for the corresponding probabilistic classifiers because the causal graph does not change when $f(X_A^\perp)$ changes from a binary-valued (i.e., $f(X) \in \{0, 1\}$) function to a probability-valued function (i.e., $f(X) \in [0, 1]$).

From demographic parity:

1. Conditional demographic parity is equivalent to demographic parity when some set of legitimate factors L is held constant at some value l.

From equalized odds:

2. False positive error rate balance is equivalent to equalized odds when considering only the population with negative label (i.e., $Y = 0$).

3. False negative error rate balance is equivalent to equalized odds when considering only the population with positive label (i.e., $Y = 1$).

4. Balance for negative class is equivalent to equalized odds for probabilistic classifiers when considering only the population with negative label (i.e., $Y = 0$).

5. Balance for positive class is equivalent to equalized odds for probabilistic classifiers when considering only the population with negative label (i.e., $Y = 1$).

From binary calibration:

6. Predictive parity is equivalent to binary calibration when considering only the population with positive label (i.e., $D = 1$).

7. Score calibration is equivalent to binary calibration for probabilistic classifiers.

□

Corollary 2.2. *Assume faithfulness.*

1. *Under the graph with measurement error as shown in Figure 2a, a predictor achieves counterfactual fairness if and only if it achieves demographic parity.*

2. *Under the graph with selection on label as shown in Figure 2b, a predictor achieves counterfactual fairness if and only if it achieves equalized odds.*

3. *Under the graph with selection on predictors as shown in Figure 2c, a predictor achieves counterfactual fairness if and only if it achieves calibration.*

Proof. By Theorem 2:

1. Observe in Figure 2a that the only path between $X_A^\perp$ and A is blocked by Y, so counterfactual fairness implies demographic parity. Because the only block in that path is Y, counterfactual fairness does not imply equalized odds. And the only path between Y and A (a parent-child relationship) is unblocked and does not contain $X_A^\perp$, so counterfactual fairness does not imply calibration.

2. Observe in Figure 2b that the only path between $X_A^\perp$ and A is unblocked (because S is necessarily included in the predictive model), so counterfactual fairness does not imply demographic parity. Because that path contains Y, counterfactual fairness implies equalized odds. And the only path between Y and A is unblocked (because S is necessarily included in the predictive model) and does not contain $X_A^\perp$, so counterfactual fairness does not imply calibration.

3. Observe in Figure 2c that the only path between $X_A^\perp$ and A is unblocked (because S is necessarily included in the predictive model), so counterfactual fairness does not imply demographic parity. Because that path does not contain Y, counterfactual fairness does not imply equalized odds. And the only path between Y and A is unblocked (because S is necessarily included in the predictive model) and contains $X_A^\perp$, so counterfactual fairness implies calibration.

□

Theorem 3. *Let X be an input dataset $X \in \mathcal{X}$ with a binary label $Y \in \mathcal{Y} = \{0, 1\}$ and protected class $A \in \{0, 1\}$. Define a predictor:*

$$f_{naive} := \operatorname*{argmin}_f \mathbb{E}[\ell(f(X, A), Y)]$$

where f is a proper scoring rule. Define another predictor:

$$f_{CF} := \mathbb{P}(A = 1) f_{naive}(X, 1) + \mathbb{P}(A = 0) f_{naive}(X, 0)$$

If the association between Y and A is purely spurious, then f_{CF} is counterfactually fair.

Proof. Notice that f_{CF} does not depend on A directly because the realization of A is not in the definition. To show that f_{CF} also does not depend on A indirectly (i.e., through X), consider that a purely spurious association means that $Y \perp X \mid X_A^\perp, A$. Therefore, the naive predictor:

$$\begin{aligned} f_{naive}(X, A) &= \mathbb{P}(Y = 1 | X, A) \\ &= \mathbb{P}(Y = 1 | X_A^\perp, A) \end{aligned}$$

Because $X_A^\perp$ is the component of X that is not causally affected by A, there is no term in f_{CF} that depends on A, which means f_{CF} is counterfactually fair. □