Avoiding Disparate Impact with Counterfactual Distributions

Hao Wang, Berk Ustun, and Flavio P. Calmon
Harvard SEAS
{hao_wang, berk, fcalmon}@g.harvard.edu

Abstract

When a classification model is used to make predictions on individuals, it may be undesirable or illegal for the performance of the model to change with respect to a sensitive attribute such as race or gender. In this paper, we aim to evaluate and mitigate such disparities in model performance through a distributional approach. Given a black-box classifier that performs unevenly across sensitive groups, we consider a *counterfactual distribution* of input variables that minimizes the performance gap. We characterize properties of counterfactual distributions for common fairness criteria. We then present novel machinery to efficiently recover counterfactual distributions given a sample of points from its target population. We describe how counterfactual distributions can be used to avoid discrimination between protected groups by: (i) identifying proxy variables to omit in training; and (ii) building a preprocessor that can mitigate discrimination. We validate both use cases through experiments on a real-world dataset.

1 Introduction

A machine learning model has *disparate impact* when its performance changes across groups defined by a *sensitive* attribute (e.g., race, gender). Recent work has shown that models can exhibit such variation in performance even when they do not use sensitive attributes as an input [see e.g., reports of disparate impact in facial recognition in 1, 3]. These disparities have motivated a plethora of research on how disparate impact can be measured [26, 17, 20, 12, 15, 10], and how it can be mitigated [9, 5, 25, 4].

In spite of these recent advances, disparate impact is still difficult to understand or mitigate in certain real-world applications. This is because: (i) models may be deployed on a population that does *not* reflect the patterns contained in the training data; and (ii) models are *not* be developed in-house, but procured from a third-party vendor who have the necessary technical expertise [7]. In such settings, users may only have "black-box" access to the classifier (e.g., via a prediction API), may have limited access to the training data (e.g., due to privacy or intellectual property issues), may not be able to use the training data to draw conclusions about disparate impact in their population of interest [e.g., due to dataset shift 21].

In this paper, we aim to evaluate and mitigate disparate impact in such settings using tools from information theory and robust statistics. We consider a hypothetical distribution of input variables that minimizes disparate impact in a population of interest (i.e., the *target population*). We refer to this distribution as a *counterfactual distribution* [as an analog to the counterfactual explanations of 22]. As we will show, a formal study of counterfactual distributions has much to offer. Once a classifier is fixed, disparate impact can be traced back to differences between the distributions of input and output variables across protected groups. Informally, a counterfactual distribution can be found by continuously perturbing the distribution of input features over a target population until a given discrimination metric is minimized. The resulting counterfactual distribution reveals insights on the sources of disparate impact that are tailored for the target population, but also allows us to design preprocessing methods that will reduce the disparate impact without training a new model.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

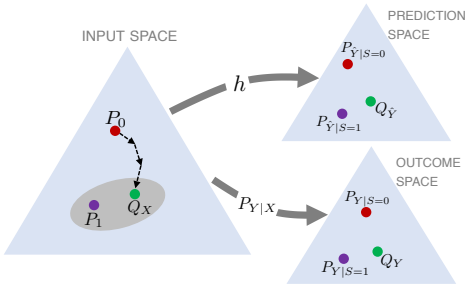

Figure 1: Illustration of disparate impact on the probability simplex for a fixed classifier h. P_0 and P_1 denote the distribution of input variables for the minority and majority groups. Disparate impact arises due to differences in the distribution of predicted outcomes ($P_{\hat{Y}|S=0}$ vs. $P_{\hat{Y}|S=1}$) and true outcomes ($P_{Y|S=0}$ vs. $P_{Y|S=1}$). A *counterfactual distribution* Q_X is a perturbation of P_0 that minimizes a measure of disparity (see Table 1). If the disparate impact persists under Q_X, there may be irreconcilable differences between the groups (i.e., $P_{Y|X,S=0} \neq P_{Y|X,S=1}$, see Prop. 1). The counterfactual distribution may not be unique, as shown by the shaded ellipse. The closest work to ours is that of [9], which proposed to map P_0 and P_1 to a common distribution via optimal transport theory in order to reduce the disparity of predicted outcomes. However, when the true outcome is involved, mapping P_0 and P_1 to the same distribution does not necessarily reduce discrimination since $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

Acronym	Performance Metric	Discrimination Metric
DA_λ	Distribution Alignment	$\text{D}_{\text{KL}}(P_{\hat{Y}\|S=0} \| P_{\hat{Y}\|S=1}) + \lambda \text{D}_{\text{KL}}(P_0 \| P_1)$
FNR	False Negative Rate	$\Pr(\hat{Y}=0\|Y=1,S=0) - \Pr(\hat{Y}=0\|Y=1,S=1)$
FPR	False Positive Rate	$\Pr(\hat{Y}=1\|Y=0,S=0) - \Pr(\hat{Y}=1\|Y=0,S=1)$

Table 1: Discrimination metrics $\mathsf{M}(P_0, P_1)$ for common fairness criteria. Distribution Alignment (DA) is a new metric related to the divergence in output distributions, which measures the statistical indistinguishability of two populations [24]. DA with $\lambda = 0$ measures the parity of predicted outcomes [similar to statistical parity in 5]. Our framework can be generalized to other metrics [see 19, 26, for a list].

Extended Version. The extended version of this paper [23] contains additional results on counterfactual distributions, our descent procedure, and experiments. We provide software to recover counterfactual distributions and reproduce our experimental results at http://github.com/ustunb/ctfdist.

2 Framework

We consider a standard classification task where the goal is to predict a label $Y \in \{0, 1\}$ using a vector of d random variables $X = (X_1, \ldots, X_d) \in \mathcal{X}$ with distribution P_X. We assume we are given a fixed black-box classifier $h : \mathcal{X} \to [0, 1]$ where $h(\boldsymbol{x}) \in \{0, 1\}$ if h outputs a predicted label (e.g., SVM) and $h(\boldsymbol{x}) \in [0, 1]$ if it outputs a predicted probability (e.g., logistic regression). We consider differences in model performance with respect to a *sensitive attribute* $S \in \{0, 1\}$. We assume S is not used as an input variable (as this would violate laws on disparate treatment [2]). We refer to individuals where $S = 0$ and $S = 1$ as the *minority* and *majority* groups, and denote the distributions of input variables as $P_0 \triangleq P_{X|S=0}$ and $P_1 \triangleq P_{X|S=1}$, respectively.

We measure the performance disparity between groups in terms of a *discrimination metric* (see Table 1 for examples for common fairness criteria).

Definition 1. *Given a classification model h and distributions $P_{Y|X,S}$ and P_S, a discrimination metric is a mapping $\mathsf{M} : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ where $\mathcal{P}$ is the set of probability distributions over $\mathcal{X}$.*

A *counterfactual distribution* is a hypothetical probability distribution of input variables for the minority group that minimizes a given discrimination metric.

Definition 2. *Given a discrimination metric $\mathsf{M}(P_0, P_1)$, a counterfactual distribution is a probability distribution of input variables X such that: $Q_X \in \operatorname{argmin}_{Q'_X \in \mathcal{P}} |\mathsf{M}(Q'_X, P_1)|$.*

The uniqueness of a counterfactual distribution depends on the choice of discrimination metric. For a metric such as DA_λ with $\lambda > 0$, there exists only one counterfactual distribution $Q_X = P_1$. In general,

however, there may be multiple counterfactual distributions. Further, the distribution of input variables for majority group P_1 is not necessarily a counterfactual distribution when $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

Counterfactual distributions provide a tool to detect irreconcilable differences in the conditional distributions across groups as shown in Proposition 1.

Proposition 1. *If* $\mathsf{M}(Q_X, P_1) > 0$ *where* Q_X *is a counterfactual distribution for a discrimination metric in Table 1, then* $P_{Y|X,S=0} \neq P_{Y|X,S=1}$.

This result illustrates how a counterfactual distribution can detect cases where a classifier has an irreconcilable performance disparity between groups (i.e., a disparity that cannot be addressed by perturbing the distribution of input variables for the minority group). The result complements recent results on inevitable trade-offs between groups [see e.g., 16], and provides a sufficient condition to inform when we should train different classifiers for different groups [see e.g., the methods of 8, 25].

3 Methodology

In what follows, we describe how influence functions provide a natural descent direction in our setting. We then use this result to design an efficient descent procedure to recover a counterfactual distribution using samples from a target population.

Computing a Descent Direction. We first consider how a discrimination metric decreases when we slightly perturb the distribution of input variables from P_0 to a perturbed distribution of the form $\widetilde{P}_0(\boldsymbol{x}) \triangleq P_0(\boldsymbol{x})(1 + \epsilon f(\boldsymbol{x}))$ where $f(\boldsymbol{x})$ represents a direction in the probability simplex while ϵ represents the magnitude of perturbation. In Proposition 2, we show that the direction of steepest descent for a discrimination metric can be computed using a normalized influence function [14, 11].

Definition 3. *The influence function* $\psi : \mathcal{X} \to \mathbb{R}$ *is:*

$$\psi(\boldsymbol{x}) \triangleq \lim_{\epsilon \to 0} \frac{\mathsf{M}\left((1-\epsilon)P_0 + \epsilon\delta_{\boldsymbol{x}}, P_1\right) - \mathsf{M}(P_0, P_1)}{\epsilon} \tag{1}$$

where $\delta_{\boldsymbol{x}}(\mathbf{z})$ *is the delta function at* $\boldsymbol{x}$.

Proposition 2. *For a given discrimination metric* $\mathsf{M}(P_0, P_1)$*, we have that*

$$\operatorname*{argmin}_{f(\boldsymbol{x})} \lim_{\epsilon \to 0} \frac{\mathsf{M}(\widetilde{P}_0, P_1) - \mathsf{M}(P_0, P_1)}{\epsilon} = \frac{-\psi(\boldsymbol{x})}{\sqrt{Var\left[\psi(X)|S=0\right]}}, \tag{2}$$

where: $f : \mathcal{X} \to \mathbb{R}$ *is a perturbation function from the class of all functions with zero mean and unit variance w.r.t.* P_0*, and* $\epsilon > 0$ *is a positive scaling constant chosen so that* $\widetilde{P}_0$ *is a valid probability distribution.*

In the extended version of this paper [23], we show that the influence functions for discrimination metrics in Table 1 can be efficiently computed via closed-form expressions given: (i) $h(\boldsymbol{x})$, the classifier that we wish to audit; (ii) a classifier that predicts group membership in the target population, $P_{S|X}(1|\boldsymbol{x})$; (iii) a classifier to predict the outcome for individuals from the minority group, $P_{Y|X,S=0}(1|\boldsymbol{x})$. Since we are given (i), we can therefore compute influence functions by training (ii) and (iii) using an *auditing dataset* $\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^n$.

Recovering a Counterfactual Distribution. In Algorithm 1, we present a descent procedure to recover a counterfactual distribution for a given discrimination metric $\mathsf{M}(\cdot)$. The procedure is analogous to a stochastic gradient descent in the space of distributions over $\mathcal{X}$. At each iteration, it computes the value of an influence function $\psi(\boldsymbol{x})$ that indicates the "direction" in which the distribution P_0 should be perturbed to reduce disparate impact for samples in the auditing dataset. Since perturbing a distribution can be achieved by a resampling operation, the procedure then samples entries for the minority population with weights $1 - \epsilon \cdot \psi(\boldsymbol{x})$. Thus, the resampled dataset mimics a dataset drawn from a perturbed distribution that achieves a lower value of $\mathsf{M}(\cdot)$. These steps are repeated until $\mathsf{M}(\cdot)$ ceases to decrease. In Figure 2, we plot the progress of Algorithm 1 for a synthetic dataset, showing that it efficiently converges to a counterfactual distribution.

Algorithm 1: Distributional Descent

Input

$h : \mathcal{X} \to [0, 1]$ ▷ classification model
$\mathsf{M}(\cdot)$ ▷ discrimination metric
$\epsilon > 0$ ▷ step size
$P_{S|X}(1|\boldsymbol{x})$ ▷ group membership model
$P_{Y|X,S=0}(1|\boldsymbol{x})$ ▷ outcome model for minority
$\mathcal{D}^{\text{audit}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=1}^{n}$ ▷ auditing dataset
$\mathcal{D}^{\text{holdout}} = \{(\boldsymbol{x}_i, y_i, s_i)\}_{i=n+1}^{m}$ ▷ holdout dataset

Initialize

$c(\boldsymbol{x}) \leftarrow 1$ ▷ sampling weights
$\mathcal{D} \leftarrow \mathcal{D}^{\text{audit}}$
$\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathcal{D}^{\text{holdout}}\right)$

repeat

$\psi(\boldsymbol{x}) \leftarrow$ compute influence function for all $\boldsymbol{x} \in \mathcal{D}$
$c(\boldsymbol{x}) \leftarrow (1 - \epsilon \cdot \psi(\boldsymbol{x}))c(\boldsymbol{x})$
$\mathcal{D} \leftarrow \mathsf{Resample}(\mathcal{D}, 1 - \epsilon\psi(\boldsymbol{x}))$ ▷ resample points
$\mathsf{M}_{\text{old}} \leftarrow \mathsf{M}_{\text{new}}$
$\mathsf{M}_{\text{new}} \leftarrow \mathsf{M}\left(\mathsf{Resample}\left(\mathcal{D}^{\text{holdout}}, c(\boldsymbol{x})\right)\right)$

until $\mathsf{M}_{\text{new}} \geq \mathsf{M}_{\text{old}}$
return: $c(\boldsymbol{x})$ ▷ $c(\boldsymbol{x})$ is an aggregate perturbation

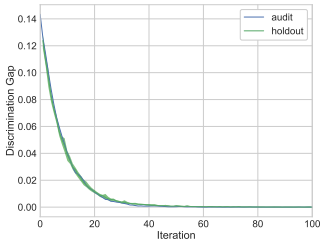

Figure 2: Values of DA across iterations of distributional descent for the auditing dataset (blue) and holdout dataset (green). Here, the procedure converges to a counterfactual distribution in 40 iterations. We show additional steps for the sake of illustration.

4 Demonstrations

We present next an experiment that demonstrates how counterfactual distributions can be used to understand and mitigate disparate impact (cf. extended version [23] for additional experiments). We recover counterfactual distributions for a classifier trained on the `adult` dataset. We use 30% of samples to train the classifier, and 20% to evaluate the performance of our classifier. The remaining 50% of samples are used to train two models: (i) $P_{S|X}(1|\boldsymbol{x})$, a model to predict the sensitive attribute; and (ii) $P_{Y|X,S=0}(1|\boldsymbol{x})$, a model to estimate the true outcome for the minority group $S = 0$. Using these models, we can recover a counterfactual distribution via the descent procedure in Algorithm 1.

Interpretation. Counterfactual distributions provide a means to understand discrimination via contrastive analyses. The difference between the observed and counterfactual distributions (visualized in Figure 3) can be used to either identify prototypical samples [see e.g, 13], or to score features in terms of their ability to discriminate by proxy in the target population [6].

Mitigation. Given a counterfactual distribution Q_X, we can mitigate the disparate impact of a classifier in a target population by building a *preprocessor* that maps features from P_0 to Q_X. The preprocessor can be built by solving an optimal transport problem [via linear programming or other techniques, 18]. In Figure 3, we show the impact of building a randomized preprocessor using counterfactual distributions recovered with in Algorithm 1. As shown, the preprocessor can reduce disparate impact in the minority group without a major effect on accuracy across the ROC curve.

	Majority	Minority	Counterfactual Distribution	
	P_1	P_0	$\text{DA}_{0.0}$	FNR
Married	61.6	17.5	31.9	15.7
Immigrant	10.4	10.5	10.0	10.9
HighestDegree is HS	32.4	32.1	26.4	31.1
HighestDegree is BS	17.2	13.7	18.7	14.0
HighestDegree is MSorPhD	7.2	5.4	8.5	4.3
AnyCapitalLoss	4.5	3.2	6.4	3.3
Age≤30	31.1	40.1	33.6	40.6
WorkHrsPerWeek<40	17.6	38.5	34.8	38.0
JobType is WhiteCollar	18.5	33.6	35.4	34.2
JobType is BlueCollar	33.8	4.6	3.7	4.5
JobType is Specialized	21.9	22.9	26.8	22.5
JobType is ArmedOrProtective	2.9	0.9	1.1	1.0
Industry is Private	69.2	70.2	66.3	71.0
Industry is is Government	12.2	15.8	18.3	15.2
Industry is SelfEmployed	13.9	5.2	7.5	5.3

		No Preprocessing			With Preprocessor		Change in Performance	
Metric	Minority Group	Majority Value	Minority Value	**Disc. Gap**	Minority Value	**Disc. Gap New**	Minority AUC Before	Minority AUC After
FPR	Male	0.016	0.105	**0.089**	0.019	**0.002**	0.826	0.724
FNR	Female	0.508	0.653	**0.144**	0.500	**-0.008**	0.893	0.857
$\text{DA}_{0.1}$	Male	-	-	**0.206**	-	**0.000**	0.826	0.683

Figure 3: Top: Marginals of the counterfactual distributions recovered with Algorithm 1 for a classifier on `adult`. Bottom: change in disparate impact and classifier performance when using a randomized preprocessor built using a counterfactual distribution.

References

[1] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. *ProPublica*.

[2] Barocas, S. and Selbst, A. (2016). Big data's disparate impact.

[3] Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91.

[4] Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001.

[5] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM.

[6] Datta, A., Sen, S., and Zick, Y. (2016). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy*, pages 598–617. IEEE.

[7] Diakopoulos, N. (2014). Algorithmic-accountability: the investigation of black boxes. *Tow Center for Digital Journalism*.

[8] Dwork, C., Immorlica, N., Kalai, A. T., and Leiserson, M. D. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133.

[9] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

[10] Galhotra, S., Brun, Y., and Meliou, A. (2017). Fairness testing: testing software for discrimination. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, pages 498–510. ACM.

[11] Huber, P. J. (2011). Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer.

[12] Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666.

[13] Kim, B., Khanna, R., and Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288.

[14] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894.

[15] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4069–4079.

[16] Lipton, Z. C., Chouldechova, A., and McAuley, J. (2018). Does mitigating ml's impact disparity require treatment disparity? *arXiv preprint arXiv:1711.07076*.

[17] Pierson, E., Corbett-Davies, S., and Goel, S. (2017). Fast threshold tests for detecting discrimination. *arXiv preprint arXiv:1702.08536*.

[18] Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume I: Theory*, volume 1. Springer Science & Business Media.

[19] Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(5):582–638.

[20] Simoiu, C., Corbett-Davies, S., Goel, S., et al. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216.

[21] Sugiyama, M., Lawrence, N. D., Schwaighofer, A., et al. (2017). *Dataset shift in machine learning*. The MIT Press.

[22] Wachter, S. and Mittelstadt, B. (2018). A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai.

[23] Wang, H., Ustun, B., and Calmon, F. P. (2018a). Avoiding discrimination with counterfactual distributions. *Harvard SEAS Working Paper, https://scholar.harvard.edu/files/hao/files/ctf.pdf*.

[24] Wang, H., Ustun, B., and Calmon, F. P. (2018b). On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning. In *2018 IEEE International Symposium on Information Theory*, pages 1216–1220. IEEE.

[25] Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A. (2017). From parity to preference-based notions of fairness in classification. In *Advances in Neural Information Processing Systems*, pages 229–239.

[26] Žliobaitė, I. (2017). Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089.