Fairness and robustness in anti-causal prediction

Maggie Makar [1] **Alex D'Amour** [2]

Abstract

We discuss connections between robustness to distribution shift and fairness through a causal lens. These connections show that fairness can be motivated entirely on the basis of achieving better performance, and suggest that robustness-motivated approaches can be used to enforce the separation fairness criteria.

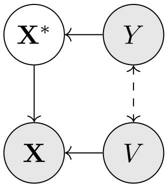

Figure 1. Causal DAG of the setting in this paper.

1. Introduction

When designing fair systems, practitioners face a number of choices: among them are the choice of fairness criterion, and how to implement it. A practitioner may choose between using an unconstrained model that maximizes overall performance, a model that makes predictions independently of the sensitive groups (independence), or a model that makes predictions independently of sensitive groups given the ground truth label (separation) [see 2, Chapter 2].

Here, we show that the causal structure of a problem can provide useful context when choosing and enforcing a fairness criterion in a given application. We present this argument by making a connection to some recent insights on the relevance of causal structure to robust machine learning. In particular, we focus on an anti-causal prediction setting, where the input to a classifier (e.g., an image) is assumed to be generated as a function of the label and the sensitive attribute (see Figure 1). In this context, we show that the connection between fairness and robustness can provide new motivations and methods for enforcing fairness criteria in practice.

As we show, for specific causal structures, some fairness criteria and distributional robustness criteria align. We discuss practical implications of this alignment for motivating and enforcing fairness criteria in this setting. Specifically, we focus on alignment between the separation fairness criterion and risk invariance as a robustness criterion—that the predictive risk of a model remain invariant across a family of distribution shifts. We show that: (1) The separation fairness criterion implies risk invariance across a family of distribution shifts that respect the causal structure in figure 1. This provides new perspective on discussions of fairness-performance tradeoffs when applying the separation criterion in practice. (2) In practice, algorithms designed to enforce risk invariance also enforce the separation criterion, in some cases more effectively than algorithms that attempt to directly incorporate the separation criterion. (3) For contrast, we show a conflict between the independence fairness criterion (often measured in terms of equalized predictions between V groups, or demographic parity) and the risk invariance property.

Related work is presented in the appendix.

2. Background and preliminaries

Our goal is to construct a predictor $f(\mathbf{X})$ that predicts a label Y (e.g., pneumonia) from an input $\mathbf{X}$ (e.g., chest X-ray). In addition, we have a protected attribute V (e.g., patient sex) available only at training time. Throughout, we will use capital letters to denote variables, and small letters to denote their value. Our training data consist of tuples $\mathcal{D} = \{(\mathbf{x}_i, y_i, v_i)\}_{i=1}^n$ drawn from a source training distribution P_s. We restrict our focus to the case where Y and V are binary and f is a classifier. Specifically, we will consider functions f of the form $f = h(\phi(\mathbf{x}))$, where ϕ is a representation mapping and h is the final classifier.

In this context, a practitioner may be interested in ensuring that the classifier $f(\mathbf{X})$ treats individuals from different groups fairly. We focus on two fairness criteria [2]. While these criteria are typically defined with respect to the predicted class (i.e., $\hat{Y} = \mathbb{1}\{f(\mathbf{X}) > \delta\}$ for some threshold δ) we consider stronger fairness notions defined with respect to the predicted probabilities $f(\mathbf{X})$ [19]. This focuses the

[1] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor [2] Google Research, Cambridge, Massachusetts. Correspondence to: Maggie Makar <mmakar@umich.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA. Copyright 2022 by the author(s).

exposition on issues relating to the quality of $f(\mathbf{X})$ independently of the choice of δ. The first criterion, separation, requires that the distribution of predictions $f(\mathbf{X})$ be the same across groups V conditional on the label Y, i.e., $f(\mathbf{X}) \perp\!\!\!\perp V \mid Y$. Separation is often evaluated in terms of equalized odds (EO) [9], defined as $\mathbb{E}_{P_s}[f(\mathbf{X}) \mid V = 0, Y = y] = \mathbb{E}_{P_s}[f(\mathbf{X}) \mid V = 1, Y = y] \quad \forall y \in \{0, 1\}$. The second criterion, independence, requires that the distribution of predictions be the same overall across groups V. Independence is often evaluated in terms of demographic parity (DP), defined as $\mathbb{E}_{P_s}[f(\mathbf{X}) \mid V = 0] = \mathbb{E}_{P_s}[f(\mathbf{X}) \mid V = 1]$. Unlike EO which requires equality conditional on Y, DP requires a marginal form of equality.

Assumptions. We assume that P_s follows an anti-causal structure shown in Figure 1, in which the inputs $\mathbf{X}$ are generated by the labels (Y, V). We assume that the labels Y and V are correlated, but not causally related. We represent this in Figure 1 with the dashed bidirectional arrow. We assume that there is a sufficient statistic $\mathbf{X}^*$ such that Y only affects $\mathbf{X}$ through $\mathbf{X}^*$, and $\mathbf{X}^*$ can be fully recovered from $\mathbf{X}$ via the function $\mathbf{X}^* := e(\mathbf{X})$, for an unknown $e(\mathbf{X})$. Finally, we make an overlap assumption on the source distribution, P_s. Specifically we assume that the support of $P_s(V)P_s(Y)$ is contained in the support of $P_s(V, Y)$. Intuitively, this assumption implies that we observe all combinations of Y and V during training time that will also appear in any target test distribution. Absent such an assumption, the behavior of f on unobserved combinations is unlearnable using the observed data.

Robustness as Risk Invariance under Shifts in Dependence. While fairness is often motivated by a principle of non-discrimination, robustness is often motivated by generalization to out-of-distribution settings. Of particular interest here are scenarios where a “shortcut” association between inputs and target label changes between training and deployment time [6]. In this paper, we focus on a particular notion of robustness called risk invariance. Following the robustness literature, we assume that the model f is trained on a source distribution P_s, and measure its predictive risk on one or many target distributions P_t.

We write the generalization risk of a function f on a distribution P as $R_P = \mathbb{E}_{\mathbf{X},Y\sim P}[\ell(f(\mathbf{X}), Y)]$, where ℓ is a predictive loss (we define this as the logistic loss in arguments that follow). We say that a model f is risk invariant across a family of distributions $\mathcal{P}$ if its predictive risk is the same for each target distribution $P_t \in \mathcal{P}$, and use $\mathcal{F}_{\text{rinv}}$ to denote the family of all risk invariant predictors. Here, we consider families of target distributions that can be generated from P_s by interventions on the causal DAG in Figure 1. Specifically, we consider interventions on the dependence between Y and V that keep the marginal distribution of Y constant. [1] Each distribution in this family can be obtained by replacing the source conditional distribution $P_s(V \mid Y)$ with a target distribution $P_t(V \mid Y)$:

$$\mathcal{P} = \{P_s(\mathbf{X} \mid \mathbf{X}^*, V)P_s(\mathbf{X}^* \mid Y)P_s(Y)P_t(V \mid Y)\}, \quad (1)$$

This family allows the marginal dependence between Y and V to change arbitrarily. For our analysis, one distribution contained in the set $\mathcal{P}$ will be important: the distribution where $Y \perp\!\!\!\perp V$, i.e., $P^\circ := P_s(\mathbf{X} \mid \mathbf{X}^*, V)P_s(\mathbf{X}^* \mid Y)P_s(Y)P^\circ(V)$. We refer to P° as the ideal distribution. In the chest x-ray example, P° is the distribution where the base rates of pneumonia are equal across sex groups.

The concept of risk invariance characterizes the stability of a given predictor but it does not characterize its predictive accuracy. For example, a predictor that returns the same prediction for all values of $\mathbf{X}$ is risk invariant in this family but is non-optimal for most non-trivial cases. We define an *optimal risk invariant* predictor f_{rinv} as an invariant predictor that also satisfies the property $f_{\text{rinv}} \in \arg\min_{f\in\mathcal{F}_{\text{rinv}}} R_{P_t}(f) \quad \forall P_t \in \mathcal{P}$. Importantly, risk invariant optimality is distinct from the in-distribution optimality, and one does not imply the other. Predictors that rely on shortcuts (and are hence not invariant) will likely have better in-distribution performance but worse out of distribution performance and vice versa.

3. Fairness and robustness in anti-causal settings

In this section, we show that risk invariance and separation are aligned in the anti-causal setting. This can provide motivation for preferring separation over an unconstrained model or a model that enforces independence, beyond standard non-discrimination rationales.

Often the most difficult choice in deciding to apply fairness criteria to a classification problem is whether to constrain the model at all. Some of the oldest discussions in applying fairness to machine learning center on tradeoffs between parities enforced by fairness-constrained models and overall predictive performance of unconstrained models [see, e.g., 4, 19]. Here, we revisit this discussion in our anti-causal setting, and show that separation can be motivated on purely performance-oriented grounds if the notion of performance is expanded to include predictive risk under distribution shifts. In addition, we show that the optimal risk invariant predictor in this setting satisfies separation, suggesting that algorithms that target the optimal risk in-

[1] Our arguments can generalize to include cases where the marginal distribution of Y also changes, but would introduce some notational overhead.

variant predictor can be effective for learning models that are fair according to the separation criterion.

Our discussion here revolves around two results. First, in the anti-causal setting in Figure 1, separation implies risk invariance. Secondly, the optimal risk invariant predictor in this setting satisfies separation. We present these two results formally next. All proofs are in the appendix.

Proposition 1. *In the anti-causal setting shown in Figure 1, suppose that a a predictor f satisfies separation in the training distribution, that is, $f(\mathbf{X}) \perp\!\!\!\perp_{P_s} V \mid Y$. Then f is risk-invariant with respect to the family of target distributions $\mathcal{P}$ defined in* (1).

Proposition 2. *In the anti-causal setting shown in Figure 1, (a) the optimal risk invariant predictor with respect to $\mathcal{P}$ has the form $f^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}^*]$, and (b), this predictor satisfies separation.*

These results show that if model performance *beyond* the training distribution P_s is important in an application, risk invariance can provide a purely performance-oriented motivation for enforcing a separation criterion. This offers a counterpoint to standard discussions of fairness-performance tradeoffs, by highlighting the importance of specifying the distribution (or distributions) on which fairness and performance are defined. On one hand, in our setting, the Bayes optimal in-distribution predictor, $\mathbb{E}_{P_s}[Y \mid \mathbf{X}]$, does not in general satisfy separation (this follows because $\mathbf{X}$ is not d-separated from V by Y, so $\mathbb{E}[Y \mid \mathbf{X}]$ is not independent of V given Y). This implies that a model satisfying separation can have lower-than-optimal in-distribution accuracy. On the other hand, because Proposition 1 shows that separation implies risk invariance, the performance of a model that satisfies separation cannot degrade if the model is deployed in a causally consistent scenario included in $\mathcal{P}$. In fact, as Proposition 2 shows, the best possible risk invariant model satisfies separation, implying that there is no tradeoff between optimal *invariant* performance and this fairness criterion.

Further, Proposition 2 suggests that, in our anti-causal setting, criteria designed for targeting optimal risk invariant predictors may be effective for learning classifiers that satisfy separation. For example, consider the approach proposed in Makar et al. [18] that targets the optimal risk invariant predictor by enforcing a criterion that we will call *ideal distribution independence* (IDI). We say that a model f satisfies IDI iff $f(\mathbf{X}) \perp\!\!\!\perp_{P^\circ} V$; i.e., if its predictions $f(\mathbf{X})$ are independent of the sensitive attribute V under the ideal distribution P°, where $Y \perp\!\!\!\perp_{P^\circ} V$. While the IDI criterion does not itself imply separation, Makar et al. [18] show that enforcing IDI at training time can lead to more efficient recovery of the optimal risk invariant predictor, which does satisfy separation. This suggests that enforcing risk invariance in practice can also help to close fairness gaps. We demonstrate this empirically in section 5.

By contrast, as proposition A3 in the appendix shows, there is a conflict between the optimal risk invariant predictor and independence. Hence, if downstream performance of a model on shifted distributions is important, this result provides motivation against using independence as a fairness criterion, because it implies a tradeoff with the best achievable invariant risk.

4. Enforcing Separation

The discussion in section 3 implies that there are two ways to enforce separation during training. First, it can be enforced directly by encouraging equality between representation distributions conditional on Y. Alternatively, it can be enforced indirectly my minimizing predictive risk subject to the IDI criterion, by encouraging equality between the margignal distributions of learned representations ϕ under the ideal distribution P°. In this section, we discuss these two implementation schemes.

We focus the discussion on approaches that rely on the Maximum Mean Discrepancy (MMD) to enforce statistical dependencies since it is popular in fairness [21, 17, 16], and robustness literature [18, 8]. The MMD defines a metric on probability distributions, and is equal to zero iff the two distributions are independent. Details about the MMD are included in the appendix.

The strategy for directly enforcing separation penalizes discrepancies in representation distributions conditional Y. Such a strategy is translates to minimizing the following loss: $\mathcal{L}_{\text{C-MMD}} = \min_{h,\phi}\frac{1}{n}\sum_{i=1}^{n}\ell(h(\phi(\mathbf{x}_i)), y_i) + \alpha \cdot \sum_y \widehat{\mathrm{MMD}}^2(P_{\boldsymbol{\phi}_{0,y}}, P_{\boldsymbol{\phi}_{1,y}})$, where $P_{\boldsymbol{\phi}_{v,y}} = P(\phi(\mathbf{X})|V = v, Y = y)$, α is a parameter picked through cross-validation, $\widehat{\mathrm{MMD}}^2$ is an estimate of MMD^2. This strategy has some practical limitations, especially when training using mini-batches of data in stochastic gradient descent. Within each batch, C-MMD requires first dividing the population based on Y then estimating the MMD within each subgroup. This effectively reduces the sample size used for MMD estimation, making the estimates more volatile and less reliable, especially when batch sizes are small.

Meanwhile, a strategy for enforcing separation indirectly through IDI makes use of a weighted marginal MMD discrepancy (WM-MMD) [18]. WM-MMD requires a marginal estimate of the MMD computed with respect to P° rather than the observed P_s. This strategy uses reweighting to manipulate the observed data to create a pseudo-sample from P°, using the following weights: $u(y, v) = \frac{P_s(Y=y)P_s(V=v)}{P_s(Y=y,V=v)}$, such that for each example, $u_i := u(y_i, v_i)$. The final loss to minimize for WM-MMD is: $\mathcal{L}_{\text{WM-MMD}} = \min_{h,\phi}\sum_{i=1}^{n}u_i\ell(h(\phi(\mathbf{x}_i)), y_i)\ +$

$\alpha \cdot \widehat{\mathrm{MMD}}^2(P_{\boldsymbol{\phi}_0^{\mathbf{u}}}, P_{\boldsymbol{\phi}_1^{\mathbf{u}}})$, where $\widehat{\mathrm{MMD}}^2(P_{\boldsymbol{\phi}_0^{\mathbf{u}}}, P_{\boldsymbol{\phi}_1^{\mathbf{u}}})$ is a weighted estimate of the MMD. This strategy also has practical challenges. While WM-MMD does not require this data-slicing, if base rates are too skewed within groups, the weights may be highly variable, and introduce volatility into the regularization as is typical of weighted estimators [5] Ultimately, the better strategy to employ to enforce separation depends on the context. For example, we find that WM-MMD is far more stable in our experiments. For more general use, we present a concrete heuristic to choose between the two methods in a given application.

5. Experiments

Following Jabbour et al. [12], we study the task of predicting pneumonia (Y) from chest x-rays ($\mathbf{X}$) considering sex to be a protected attribute (V). We conduct this analysis on CheXpert [11], which contains 224,316 chest x-rays, each associated with the sex of the patient and labels encoding presence or absence of pneumonia. At training time, we sample the data such that the majority of female patients have pneumonia whereas the majority of male patients do not have pneumonia i.e., $P_s(V = 1|Y = 1) = P_s(V = 0|Y = 0) = 0.9$ to create a possible shortcut for naive estimators. In addition to C-MMD, WM-MMD, we implement a deep neural network (DNN) without any robustness or fairness penalties, and M-MMD, which encourages independence rather than separation. The objective for M-MMD is similar to W-MMD with the distinction that it does not re-weight the samples, i.e., it does not utilize u_i. Details about training and cross validation are in the appendix.

Results. We examine the extent to which different fairness criteria (separation/independence) and their implementation align with robustness across distributions. To do so, we generated 6 test distributions that are compatible with the DAG in Figure 1, where the base rates of pneumonia were systematically skewed between sex groups by selective sampling. We denote these test distributions $\mathcal{P}_t = \{P_{0.1}, P_{0.3}, P_{0.5}, P_{0.7}, P_{0.9}, P_{0.95}\}$, where P_μ is generated such that $P_t(V = 1 \mid Y = 1) = P_t(V = 0 \mid Y = 0) = \mu$ with $P_t(Y = 1)$ held constant. We compute the area under the receiver operating curve (AUROC) of each of the models on the 6 test sets. To estimate the uncertainty in performance, we create 1000 bootstraps of the test set, and compute the means and standard deviations of the AUROCs.

Figure 2 shows the results. The $x-$axis shows $P(Y = 1 \mid V = 1) = P(Y = 0 \mid V = 0)$ at test time, while the $y-$axis shows the corresponding mean AUROC. The vertical dashed line shows the conditional probability at training time. DNN achieves the best in-distribution performance but its performance quickly deteriorates on test distributions dissimilar to the training distribution. C-MMD, which enforces the EO criterion achieves better risk invariance, which conforms with Propositions 1. While C-MMD should, in theory, achieve optimal risk invariant performance, its performance changes across distributions signaling that it has some dependence on sex, albeit less severe than that of the DNN. We revisit this later. Meanwhile, M-MMD uses the shortcut to satisfy the fairness criterion, which leads to encoding an opposite dependence on sex and poor in-distribution performance, confirming that there is no alignment between independence (enforced by DP) and robustness (consistent with proposition A3). Finally, while WM-MMD achieves worse in-distribution performance compared to DNN and C-MMD, it has has better out of distribution performance for distributions most dissimilar to the training distribution. This is consistent with proposition 2. Indeed, WM-MMD achieves a maximum EO violation equal to 0.02, compared to C-MMD's 0.15.

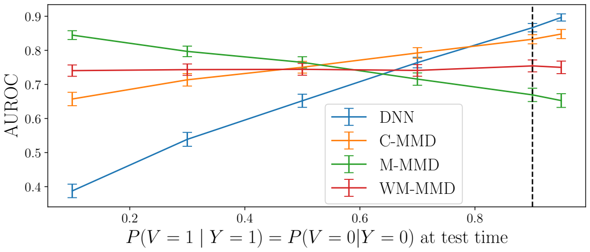

Figure 2. Test distribution on the x-axis, AUROC on the y-axis. Dashed line shows training distribution. WM-MMD is the most risk invariant sacrificing a little on in-distribution performance.

We now investigate the discrepancy in performance between WM-MMD and C-MMD: while in theory they both target risk invariance, empirical performance suggests that WM-MMD is more effective. This discrepancy can be explained, in part, by how well each optimized optimized MMD criterion generalizes to test data. Figure 3 shows the estimated MMD on the training, validation and testing data broken down by the target label. Here the testing data is sampled from the same distribution as the training data to highlight estimation error rather than errors due to data shift. The plot shows several important findings. First, the estimated MMD at training time is a more reliable estimate of the test-set MMD when weighted marginal MMD penalties are used at training and validation time, signaling that data slicing in C-MMD leads to unreliable estimates. Second, the difference between the MMD among the group defined by $Y = 1$ compared to the group defined by $Y = 0$ is smaller when using WM-MMD. Smaller difference in the MMD between the two groups is important to ensure that the outcomes for both groups defined by the target label do not vary based on the protected attribute. Usefully, this analysis can be repeated in practice to choose between the two penalties: the estimated MMD on the validation data is a reliable proxy for the estimate's generalizability. In practice, if there is a large discrepancy between the training and validation C-MMD estimates, WM-MMD might be a better option and vice versa.

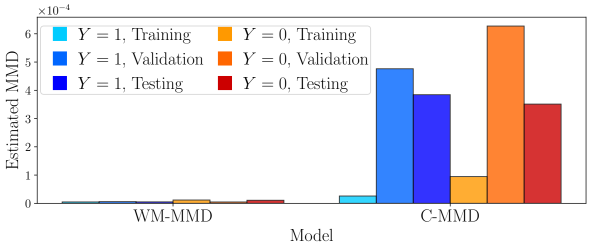

Figure 3. Estimated conditional MMD on the training, validation and testing data for WM-MMD and C-MMD. Estimates of the MMD are more stable for the WM-MMD objective than C-MMD.

Conclusion. We showed that by taking into account the causal structure of a prediction problem, we are able to draw connections between robustness and fairness. These connections provide performance-oriented motivation for applying the separation criterion in an important class of problems, and provides a new set of tools, borrowed from the robustness literature, to enforce the criterion in practice.

Acknowledgements

We thank the anonymous reviewers for their feedback. We also thank Victor Veitch for his insightful comments. This work was partially supported by the National Science Foundation under Grant No. 2153083.

References

[1] R. Adragna, E. Creager, D. Madras, and R. Zemel. Fairness and robustness in invariant learning: A case study in toxicity classification. *arXiv preprint arXiv:2011.06485*, 2020.

[2] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

[3] S. Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

[4] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. 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, 2017.

[5] C. Cortes, Y. Mansour, and M. Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

[6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *arXiv preprint arXiv:2004.07780*, 2020.

[7] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13 (1):723–773, 2012.

[8] R. Guo, P. Zhang, H. Liu, and E. Kiciman. Out-of-distribution prediction with invariant risk minimization: The limitation and an effective fix. *arXiv preprint arXiv:2101.07732*, 2021.

[9] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[10] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[12] S. Jabbour, D. Fouhey, E. Kazerooni, M. W. Sjoding, and J. Wiens. Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference*, pages 750–782. PMLR, 2020.

[13] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 656–666, 2017.

[14] I. Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163, 2018.

[15] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[16] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[17] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

[18] M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated

shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[19] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, 8:141–163, 2021.

[20] R. Nabi and I. Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] F. Prost and H. Qian. H chi, jilin chen, and alex beutel. 2019. In *Toward a better trade-off between performance and fairness with kernelbased distribution matching."ML with Guarantees" workshop at 33rd Conference on Neural Information Processing Systems*, 2019.

[22] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, and K.-W. Chang. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *arXiv preprint arXiv:2106.10826*, 2021.

[23] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[24] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests. *arXiv preprint arXiv:2106.00545*, 2021.

[25] J. Zhang and E. Bareinboim. Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

A. Related work

Robustness to distribution shift and fairness are closely related, and many lines of work have aimed to highlight formal and empirical connections between them. For example, Sagawa et al. [23] explored applying distributionally robust optimization to address worst-subgroup performance for under-represented groups. Adragna et al. [1] show that using methods meant to induce robustness leads to "more fair" classifiers for internet comment toxicity. Along similar lines, Pruksachatkun et al. [22] found that certified robustness approaches designed to ensure robustness of NLP methods against word substitution attacks can be used to reduce violations to the equalized odds criterion.

A key perspective in our work is that many fairness-robustness relationships are mediated by causal structure. In this sense, our work is most similar and complementary to Veitch et al. [24], in which the authors derive implications of counterfactual invariance to certain input perturbations, and show that these implications depend strongly on the causal structure of the problem at hand. Here, we focus on a narrower setting, with a greater focus on implications for fairness. Specifically, we focus on direct connections between distributional criteria that are often used as metrics in practice, without the conceptual overhead of defining counterfactuals with respect to sensitive attributes.

Our use of causal ideas is distinct from another body of work that defines fairness criteria directly in terms of a causal model. These include definitions of fairness in terms of direct causal effects of sensitive attributes on outcomes [13, 20, 25], or discrepancies between counterfactual outcomes [15, 3]. Instead, we focus on "oblivious" fairness criteria that are not themselves a function of causal structure [9], but show that causal structure informs how they should be applied.

B. Proofs for section 3

Proposition A1. *(Restated proposition 1)In the anti-causal setting shown in Figure 1, suppose that a a predictor f satisfies separation in the training distribution, that is, $f(\mathbf{X}) \perp\!\!\!\perp_{P_s} V \mid Y$. Then f is risk-invariant with respect to the family of target distributions $\mathcal{P}$ defined in (1).*

Proof. We show this by decomposing the risk of the model f on any target distribution $P_t \in \mathcal{P}$ in terms of the conditional risk given V and Y. Within the family $\mathcal{P}$, $P_t(\ell(f(\mathbf{X}), Y) \mid Y = y, V = v)$ is the same for all $P_t \in \mathcal{P}$; thus, we can write the risk conditional on Y and V independently of the target distribution. Let $R_{vy} := \mathbb{E}_{P_t}[\ell(f(\mathbf{X}), Y) \mid V = v, Y = y]$ for any $P_t \in \mathcal{P}$. The overall risk on a target distribution P_t can be written as the weighted average of these subgroup risks

$$R_{P_t} = \sum_{y\in\{0,1\}} P_s(Y = y) \left[P_t(V = 0 \mid Y = y)R_{0y} + P_t(V = 1 \mid Y = y)R_{1y}\right]. \tag{2}$$

Now, the separation criterion states that $f(\mathbf{X}) \perp\!\!\!\perp V \mid Y$, which implies $\ell(f(\mathbf{X}), Y) \perp\!\!\!\perp V \mid Y$, which implies that $R_{0y} = R_{1y}$ for $y \in \{0, 1\}$. Thus, the terms in the summation (2) are the same regardless of the $P_t(V = v \mid Y = y)$ factors, and thus the risk is invariant for all $P_t \in \mathcal{P}$. □

Proposition A2. *(Restated proposition 2). In the anti-causal setting shown in Figure 1, (a) the optimal risk invariant predictor with respect to $\mathcal{P}$ has the form $f^*(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}^*]$,* [2] *and (b), this predictor satisfies separation.*

Proof. Part (a) is shown as Proposition 1 in Makar et al. [18]. The key points of the proof are that (1) $\mathbb{E}[Y \mid \mathbf{X}^*]$ is representable by a function $f^*(\mathbf{X})$ under the assumptions made in Section 2, under causal assumptions (namely, the assumption that $\mathbf{X}^*$ can be written as $e(\mathbf{X})$; (2) under the uncorrelated distribution P°, $\mathbb{E}[Y \mid \mathbf{X}^*]$ is Bayes optimal; and (3) $\mathbb{E}[Y \mid \mathbf{X}^*]$ is risk invariant. (2) and (3) imply that no other risk invariant predictor can have lower risk across $\mathcal{P}$.

Part (b) follows from the fact that $\mathbf{X}^*$ is d-separated from V conditional on Y in the DAG in Figure 1. Thus, $\mathbb{E}[Y \mid \mathbf{X}^*] \perp\!\!\!\perp_{P_t} V \mid Y$ for all $P_t \in \mathcal{P}$. □

Remark 1. Veitch et al. [24] also provide a number of related results in their study of the implications of counterfactual invariance in anti-causal settings. Counterfactual invariance requires that the predictions of a model be invariant across label-preserving counterfactual versions of the input, such as the counterfactual that we would observe if the sensitive attribute were different. Our findings concern a narrower case where there exists a sufficient statistic $\mathbf{X}^*$, which Veitch et al. [24] refer to this as the "purely spurious" case. In their findings, Veitch et al. [24] show that counterfactual invariance implies the separation criterion generally in anticausal settings, and that the optimal counterfactually invariant predictor

[2] Note that $\mathbb{E}_{P_t}[Y \mid \mathbf{X}^*]$ is the same for all $P_t \in \mathcal{P}$, so we omit the distributional subscript on this expectation.

in also minimax optimal in the purely spurious case. Here, our results speak more directly to the implications of fairness criteria in the purely spurious setting, and, by focusing on the weaker risk invariance property, we make the connection without the conceptual ambiguity of defining counterfactuals with respect to sensitive attributes [see, e.g. 14, for discussion of this point].

The cost is that our results are restricted to an anti-causal setting where, in the terminology of Veitch et al. [24], the impact of the sensitive attribute is “purely spurious”. However, by focusing on this narrower setting, we aim to make the core insight that causal structure plays a key role in both the motivation and implementation of fairness criteria as concrete as possible.

The following proposition formally characterizes the conflict between the optimal risk invariant predictor and independence.

Proposition A3. *Let $f^*(\mathbf{X}) = \mathbb{E}[Y \mid e(\mathbf{X})] = \mathbb{E}[Y \mid \mathbf{X}^*]$ be the optimal risk invariant predictor with respect to $\mathcal{P}$. In addition, assume that $\mathbb{E}[f^*(\mathbf{X}) \mid Y = y]$ is a non-trivial function of y, i.e., that the value of Y actually affects the expectation of the sufficient statistic $\mathbf{X}^*$. Then for any distribution $P_t \neq P^\circ$ in $\mathcal{P}$, $f^*(\mathbf{X}) \not\perp\!\!\!\perp V$ and independence is not satisfied.*

Proof. Note that for each v, $\mathbb{E}_{P_t}[f^*(\mathbf{X}) \mid V = v] = \sum_{y\in\{0,1\}} \mathbb{E}[f^*(\mathbf{X}) \mid Y = y]P_t(Y = y \mid V = v)$. For any P_t where $Y \not\perp\!\!\!\perp V$, the weights $P_t(Y = y \mid V = v)$ in this summation will differ as a function of v. By assumption, the expectations $\mathbb{E}[f^*(\mathbf{X}) \mid Y = y]$ differ for different values of y, so changing their weights in the summation will change the sum. Thus, for $v \neq v'$, $\mathbb{E}_{P_t}[f^*(\mathbf{X}) \mid V = v] \neq \mathbb{E}_{P_t}[f^*(\mathbf{X}) \mid V = v']$, which implies $f^*(\mathbf{X}) \not\perp\!\!\!\perp V$. □

C. Methods

The MMD defined as follows:

Definition 1. *Let $Z \sim P_Z$, and $Z' \sim P_{Z'}$, be two arbitrary variables. And let Ω be a class of functions $\omega : \mathcal{Z} \to \mathbb{R}$, $\mathrm{MMD}(\Omega, P_Z, P_{Z'}) = \sup_{\omega\in\Omega} \left(\mathbb{E}_{P_Z}\omega(Z) - \mathbb{E}_{P_{Z'}}\omega(Z')\right)$.*

When Ω is set to be a general reproducing kernel Hilbert space (RKHS), the MMD defines a metric on probability distributions, and is equal to zero if and only if $P_Z = P_{Z'}$. We take Ω to be the RKHS. MMD-based regularization methods enforce statistical independences at training time by penalizing discrepancies between distributions that would be equal if the independence held. The MMD penalty can be applied to the final layer f or to the learned representation ϕ. Both methods induce the required invariances. We follow the literature in imposing the penalty on the representation ϕ.

In all our experiments, we use the V-statistic estimator for the MMD presented in [7]. This estimator relies on the radial basis function (RBF), which requires a bandwidth parameter γ picked through cross-validation.

D. Experiment details

Training details. We split the dataset into 70% examples used for training and validation, while the rest is held out for testing. We use DensNet-121 [10], pretrained on ImageNet, and fine tuned for our specific task. We train all models using a batch size of 64, and image sizes 256 × 256 for 50 epochs.

Cross-validation details. For the three MMD based models, we follow the two step cross-validation procedure outlined in Makar et al. [18] to pick α and γ. For C-MMD, M-MMD and WM-MMD, we use cross validation to get the optimal value for α and γ. For α, we pick from the values $[1e3, 1e5, 1e7]$ and for γ, we pick from the values $[10, 100, 1000]$. For the DNN, we pick the $L2$ penalty on the model weights from $[0.0, 0.0001, 0.001]$.

Training each model takes roughly 2.5 hours using a Tesla T4 GPU. We train 30 models to generate the results presented in the main paper for a total of roughly 75 hours of compute time.