SELECTIVE CLASSIFICATION CAN MAGNIFY DISPARITIES ACROSS GROUPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Selective classification, in which models can abstain on uncertain predictions, is a natural approach to improving accuracy in settings where errors are costly but abstentions are manageable. In this paper, we find that while selective classification can improve average accuracies, it can simultaneously magnify existing accuracy disparities between various groups within a population, especially in the presence of spurious correlations. We observe this behavior consistently across five vision and NLP datasets. Surprisingly, increasing abstentions can even *decrease* accuracies on some groups. To better understand this phenomenon, we study the margin distribution, which captures the model's confidences over all predictions. For symmetric margin distributions, we prove that whether selective classification monotonically improves or worsens accuracy is fully determined by the accuracy at full coverage (i.e., without any abstentions) and whether the distribution satisfies a property we term left-log-concavity. Our analysis also shows that selective classification tends to magnify full-coverage accuracy disparities. Motivated by our analysis, we train distributionally-robust models that achieve similar full-coverage accuracies across groups and show that selective classification uniformly improves each group on these models. Altogether, our results imply selective classification should be used with care and underscore the importance of models that perform equally well across groups at full coverage.

1 INTRODUCTION

Selective classification, in which models make predictions only when their confidence is above a threshold, is a natural approach in settings where errors are costly but abstentions are manageable. For example, in medical and criminal justice applications, model mistakes can have serious consequences, whereas abstentions can be handled by backing off to the appropriate human experts. Prior work has shown that across a broad array of applications, more confident predictions indeed tend to be more accurate (e.g., Hanczar & Dougherty (2008); Yu et al. (2011); Toplak et al. (2014); Mozannar & Sontag (2020); Kamath et al. (2020)). By varying the confidence threshold, we can select an appropriate trade-off between the abstention rate and the accuracy of the predictions made.

In this paper, we report a cautionary finding: while selective classification improves average accuracy, it can magnify existing accuracy disparities between various groups within a population, especially in the presence of spurious correlations. We observe this behavior across five vision and NLP datasets and two popular selective classification methods: softmax response (Cordella et al., 1995; Geifman & El-Yaniv, 2017) and Monte Carlo dropout (Gal & Ghahramani, 2016). Surprisingly, we find that increasing the abstention rate can even *decrease* accuracies on the groups that have lower accuracies at full coverage: on those groups, the models are not only wrong more frequently, but their confidence can actually be *anticorrelated* with whether they are correct. Even on datasets where selective classification improves accuracies across all groups, we find that it preferentially helps groups that already have high accuracies, further widening group disparities.

These group disparities are especially problematic in the same high-stakes areas where we might want to deploy selective classification, like medicine and criminal justice; there, poor performance on particular groups is already a significant issue (Chen et al., 2020; Hill, 2020). For example, we study a variant of CheXpert (Irvin et al., 2019), where the task is to predict if a patient has pleural effusion (fluid around the lung) from a chest X-ray. As these are commonly treated with chest tubes, models can latch onto this spurious correlation and fail on the group of patients with pleural effusion

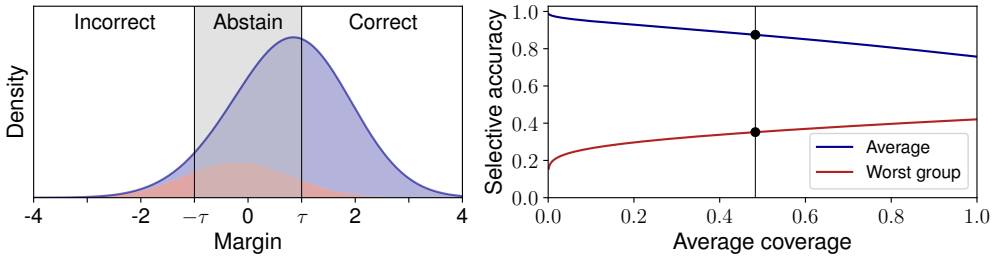

Figure 1: A selective classifier $(\hat{y}, \hat{c})$ makes the prediction $\hat{y}(x)$ on a point x if its confidence $\hat{c}(x)$ in that prediction is larger than or equal to some threshold τ. Note that it only observes the input x and not which group g it belongs to. In this schematic, we show a classifier with high accuracy on the overall data (blue) but low accuracy on a particular group (red). **Left**: The distribution of margins overall (blue) and on the red group. The margin is $\hat{c}(x)$ on correct predictions ($\hat{y}(x) = y$) and $-\hat{c}(x)$ otherwise. Since the confidence $\hat{c}(x)$ is non-negative by definition, at a threshold τ, the model makes incorrect predictions on points with margin $\leq -\tau$; abstains on points with margin between $-\tau$ and τ; and makes correct predictions on points with margin $\geq \tau$. **Right**: By varying τ, we can plot the *accuracy-coverage curve*, where the coverage is the proportion of predicted points. As coverage decreases (more abstentions), the average accuracy increases, but the worst-group accuracy *decreases*. The black dots correspond to the abstention region shaded on the left, with $\tau = 1$.

but not chest tubes or other support devices. However, this group is the most clinically relevant as it comprises potentially untreated and undiagnosed patients (Oakden-Rayner et al., 2020).

To better understand why selective classification can worsen accuracy and magnify disparities, we analyze the margin distribution, which captures the model's confidences across all predictions and determines which examples it abstains on at each threshold (Figure 1). For example, when the margin distribution is symmetric, we prove that whether selective classification monotonically improves or worsens accuracy is fully determined by the accuracy at full coverage (i.e., without any abstentions) and whether the distribution satisfies a property we term left-log-concavity. To our knowledge, this is the first work to characterize whether selective classification (monotonically) helps or hurts accuracy in terms of the margin distribution, and to compare its relative effects on different groups.

Our analysis shows that selective classification tends to magnify accuracy disparities that are present at full coverage (i.e., with no abstentions) Motivated by our analysis, we find that selective classification on group-DRO models (Sagawa et al., 2020)—which achieve similar accuracies across groups at full coverage by using group annotations during training—uniformly improves group accuracies at lower coverages, substantially mitigating the disparities observed on standard models that are instead optimized for average accuracy. These models are not a silver bullet: unlike standard models, they rely on knowing group identities during training, which is not always practical. However, they illustrate the potential of models that achieve small accuracy disparities at full coverage and can therefore maintain uniformly high accuracies under selective classification.

2 Related work

Selective classification. Abstaining from uncertain predictions is a classic idea (Chow, 1957; Hellman, 1970), and uncertainty estimation is an active area of research, from the popular approach of using softmax probabilities (Geifman & El-Yaniv, 2017) to more sophisticated methods using dropout (Gal & Ghahramani, 2016), ensembles (Lakshminarayanan et al., 2017), or training snapshots (Geifman et al., 2018). Others incorporate abstention into model training (Bartlett & Wegkamp, 2008; Geifman & El-Yaniv, 2019; Feng et al., 2019) and learn to abstain on examples human experts are more likely to get correct (Raghu et al., 2019; Mozannar & Sontag, 2020; De et al., 2020). Selective classification can also improve out-of-distribution accuracy (Pimentel et al., 2014; Hendrycks & Gimpel, 2017; Liang et al., 2018; Ovadia et al., 2019; Kamath et al., 2020). On the theoretical side, early work characterized optimal abstention rules given well-specified models (Chow, 1970; Hellman & Raviv, 1970), with more recent work on learning with perfect precision (El-Yaniv & Wiener, 2010) and guaranteed risk (Geifman & El-Yaniv, 2017). We build on this literature by establishing general conditions on the margin distribution for when selective classification helps, and importantly, by showing that it can magnify group disparities.

Group disparities. The problem of models performing poorly on some groups of data has been widely reported (e.g., Hovy & Søgaard (2015); Blodgett et al. (2016); Corbett-Davies et al. (2017); Tatman (2017); Hashimoto et al. (2018)). These disparities can arise when models latch onto spurious correlations, e.g., demographics (Buolamwini & Gebru, 2018; Borkan et al., 2019), image backgrounds (Ribeiro et al., 2016; Xiao et al., 2020), spurious clinical variables (Badgeley et al., 2019; Oakden-Rayner et al., 2020), or linguistic artifacts (Gururangan et al., 2018; McCoy et al., 2019). These disparities have implications for model robustness and equity, and mitigating them is an important open challenge, especially when models are not privy to group identities (Dwork et al., 2012; Hardt et al., 2016; Kleinberg et al., 2017; Duchi et al., 2019; Sagawa et al., 2020). Our work shows that selective classification can exacerbate this problem and must therefore be used with care.

3 Setup

A selective classifier takes in an input $x \in \mathcal{X}$ and either predicts a label $y \in \mathcal{Y}$ or abstains. We study standard confidence-based selective classifiers $(\hat{y}, \hat{c})$, where $\hat{y} : \mathcal{X} \to \mathcal{Y}$ outputs a prediction and $\hat{c} : \mathcal{X} \to \mathbb{R}_+$ outputs the model's confidence, which is always non-negative, in that prediction. It abstains on x whenever the confidence $\hat{c}(x)$ is below some threshold τ and predicts $\hat{y}(x)$ otherwise.

Data and training. We consider a data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y} \times \mathcal{G}$, where $\mathcal{G} = \{1, 2, \ldots, k\}$ corresponds to a group variable that is *unobserved* by the model. We study the common setting where the model $\hat{y}$ is trained under full coverage (i.e., without taking into account any abstentions) and the confidence function $\hat{c}$ is then derived from the trained model, as described in the next paragraph. We will primarily consider models $\hat{y}$ trained by empirical risk minimization (i.e., to minimize the average training loss); in that setting, the group $g \in \mathcal{G}$ is never observed. In Section 7, we will consider the group DRO training algorithm (Sagawa et al., 2020), which observes g at *training* time only. In both cases, g is not observed at *test* time and the model thus does not take in g. This is a common assumption: e.g., we might want to ensure that a face recognition model has equal accuracies across genders, but the model only sees the photograph (x) and not the gender (g).

Confidence. We will primarily consider *softmax response* (SR) selective classifiers, which take $\hat{c}(x)$ to be the normalized logit of the predicted class. Formally, we consider models that estimate $\hat{\mathbb{P}}(y|x)$ (e.g., through a softmax) and predict $\hat{y}(x) = \arg\max_{y \in \mathcal{Y}} \hat{\mathbb{P}}(y|x)$, with the corresponding probability estimate $\hat{\mathbb{P}}(\hat{y}(x)|x)$. For binary classifiers, we define the confidence $\hat{c}(x)$ as

$$\hat{c}(x) = \frac{1}{2} \log \left(\frac{\hat{\mathbb{P}}(\hat{y}(x)|x)}{1 - \hat{\mathbb{P}}(\hat{y}(x)|x)} \right). \tag{1}$$

This corresponds to a confidence of $\hat{c}(x) = 0$ when $\hat{\mathbb{P}}(\hat{y}(x)|x) = 0.5$, i.e., the classifier is completely unsure of its prediction. We generalize this notion to multi-class classifiers in Section A.1. Softmax response is a popular technique applicable to neural networks and has been shown to improve average accuracies on a range of applications (Geifman & El-Yaniv, 2017). Other methods offer alternative ways of computing $\hat{c}(x)$; in Appendix B.1, we also run experiments where $\hat{c}(x)$ is obtained via Monte Carlo (MC) dropout (Gal & Ghahramani, 2016), with similar results.

Metrics. The performance of a selective classifier at a threshold τ is typically measured by its *average (selective) accuracy* on its predicted points, $\mathbb{P}[\hat{y}(x) = y \mid \hat{c}(x) \geq \tau]$, and its *average coverage*, i.e., the fraction of predicted points $\mathbb{P}[\hat{c}(x) \geq \tau]$. We call the average accuracy at threshold 0 the *full-coverage accuracy*, which corresponds to the standard notion of accuracy without any abstentions. We always use the term *accuracy* w.r.t. some threshold $\tau \geq 0$; where appropriate, we emphasize this by referring to *selective accuracy*, but the terms are interchangeable.

Following convention, we evaluate models by varying τ and tracing out the *accuracy-coverage curve* (El-Yaniv & Wiener, 2010). As Figure 1 illustrates, this curve is fully determined by the distribution of the *margin*, which is $\hat{c}(x)$ on correct predictions ($\hat{y}(x) = y$) and $-\hat{c}(x)$ otherwise. We are also interested in evaluating performance on each group. For a group $g \in \mathcal{G}$, we compute its *group (selective) accuracy* by conditioning on the group, $\mathbb{P}[\hat{y}(x) = y \mid g, \hat{c}(x) \geq \tau]$, and we define *group coverage* analogously as $\mathbb{P}[\hat{c}(x) \geq \tau \mid g]$. We pay particular attention to the accuracy of the *worst group* $\operatorname{argmin}_g \mathbb{P}[\hat{y}(x) = y \mid g]$ that has the lowest accuracy at full coverage. In our setting, we only use group information to evaluate the model, which does not observe g at training or test time; we will relax this later when studying distributionally-robust models that use g at training time.

Dataset	Modality	Total examples	Prediction task	Spurious attributes $\mathcal{A}$
CelebA	Photos	202,599	Hair color	Gender
CivilComments	Text	448,000	Toxicity	Mention of Christianity
Waterbirds	Photos	11,788	Waterbird or landbird	Water or land background
CheXpert-device	X-rays	156,848	Pleural effusion	Support device
MultiNLI	Sentence pairs	412,349	Entailment	Negation words

Table 1: We study these datasets from Liu et al. (2015); Borkan et al. (2019); Sagawa et al. (2020); Irvin et al. (2019); Williams et al. (2018) respectively. For each dataset, we form a group for each combination of label $y \in \mathcal{Y}$ and spuriously-correlated attribute $a \in \mathcal{A}$, and evaluate the performance of selective classifiers on average and on each group. Dataset details in Appendix C.1.

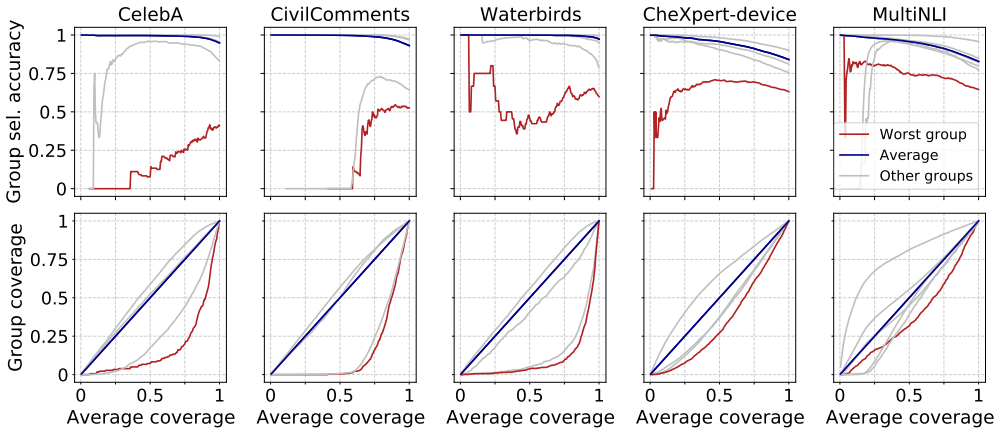

Figure 2: Accuracy (top) and coverage (bottom) for each group, as a function of the average coverage. Each average coverage corresponds to a threshold τ. The red lines represent the worst group. At low coverages, accuracy estimates are noisy as only a few predictions are made.

Datasets. We consider five datasets (Table 1) on which prior work has shown that models latch onto spurious associations, thereby performing well on average but poorly on the groups of data where the spurious association does not hold up. Following Sagawa et al. (2020), we define for each dataset the set of labels $\mathcal{Y}$ as well as a set of attributes $\mathcal{A}$ that are spuriously associated with the labels, and then form one group for each value of $(y, a) \in \mathcal{Y} \times \mathcal{A}$. For example, in the pleural effusion example from the introduction, one group would be the set of patients with pleural effusion ($y = 1$) but not support devices ($a = 0$). For simplicity, each dataset has $|\mathcal{Y}| = |\mathcal{A}| = 2$, with the exception of MultiNLI which is multi-class, with $|\mathcal{Y}| = 3$. More dataset details are in Appendix C.1.

4 Evaluating selective classification on groups

We start by investigating how selective classification affects group (selective) accuracies across the five datasets in Table 1. For each dataset, we train a standard model with empirical risk minimization (i.e., to minimize average training loss), using ResNet50 for CelebA and Waterbirds; DenseNet121 for CheXpert; and BERT for CivilComments and MultiNLI. Details are in Appendix C. We focus on softmax response (SR) selective classifiers, but show similar results for MC-dropout in Section B.1.

Accuracy-coverage curves. Figure 2 shows group accuracy-coverage curves for each dataset, with the average in blue, worst group in red, and other groups in gray. On all datasets, average accuracies improve as coverage decreases. However, the worst-group curves fall into three categories:

1. *Decreasing.* Strikingly, on CelebA and CivilComments, worst-group accuracy *decreases* with coverage: the more confident the model is on worst-group points, the more likely it is incorrect.
2. *Mixed.* On Waterbirds and CheXpert-device, as coverage decreases, worst-group accuracy sometimes increases (though not by much, except at noisy, low coverages) and sometimes decreases.
3. *Slowly increasing*. On MultiNLI, as coverage decreases, worst-group accuracy consistently improves but more slowly than other groups: from full coverage to 50% average coverage, worst-group error goes from 35% to 25% while second-to-worst group error goes from 23% to 5%.

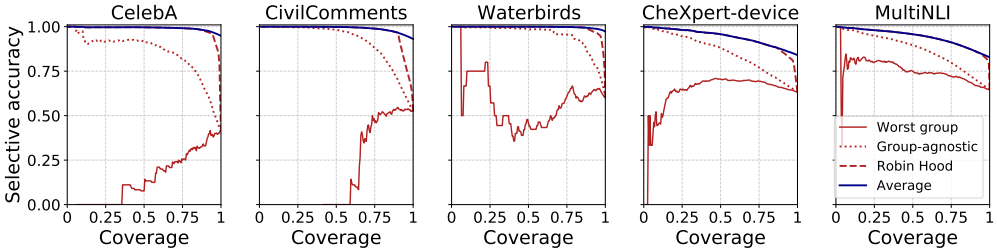

Figure 3: SR selective classifiers substantially underperform their group-agnostic baselines on the worst group, which are in turn far behind the best-case Robin Hood oracle. By construction, these share the same average accuracy-coverage curves. Similar results for MC-dropout are in Figure 7.

Algorithm 1: Group-agnostic baseline for $(\hat{y}, \hat{c})$ at threshold τ

Input: Selective classifier $(\hat{y}, \hat{c})$, threshold τ, test data $\mathcal{D}$

Output: The set of points $\mathcal{C}_\tau^{\text{ga}} \cup \mathcal{I}_\tau^{\text{ga}} \subseteq \mathcal{D}$ that the group-agnostic baseline for $(\hat{y}, \hat{c})$ makes predictions on at threshold τ.

1 Let $\mathcal{C}_\tau$ be the set of all examples that $(\hat{y}, \hat{c})$ correctly predicts at threshold τ:

$$\mathcal{C}_\tau = \{(x, y, g) \in \mathcal{D} \mid \hat{y}(x) = y \text{ and } \hat{c}(x) \geq \tau\}. \tag{2}$$

Sample a subset $\mathcal{C}_\tau^{\text{ga}}$ of size $|\mathcal{C}_\tau|$ uniformly at random from $\mathcal{C}_0$, which is the set of all examples that $\hat{y}$ would have predicted correctly at full coverage.

2 Repeat for incorrect points. Let $\mathcal{I}_\tau$ be the analogous set of incorrect predictions at τ:

$$\mathcal{I}_\tau = \{(x, y, g) \in \mathcal{D} \mid \hat{y}(x) \neq y \text{ and } \hat{c}(x) \geq \tau\}. \tag{3}$$

Sample a subset $\mathcal{I}_\tau^{\text{ga}}$ of size $|\mathcal{I}_\tau|$ uniformly at random from $\mathcal{I}_0$.

3 Return $\mathcal{C}_\tau^{\text{ga}} \cup \mathcal{I}_\tau^{\text{ga}}$. Since $|\mathcal{C}_\tau^{\text{ga}}| = |\mathcal{C}_\tau|$ and $|\mathcal{I}_\tau^{\text{ga}}| = |\mathcal{I}_\tau|$, the group-agnostic baseline makes the same numbers of correct and incorrect predictions as $(\hat{y}, \hat{c})$ at threshold τ, but distributes these correct and incorrect predictions uniformly in a group-agnostic manner.

Group-agnostic baselines and Robin Hood oracles. The results above show that even when selective classification is helping the worst group, it intuitively seems to help other groups more. We formalize this notion by comparing each classifier to a matching *group-agnostic baseline* that, at each threshold, makes the same number of correct and incorrect predictions, but distributes these predictions uniformly at random across points without knowing group identities (see Algorithm 1). By construction, the group-agnostic baseline has an identical average accuracy-coverage curve as its corresponding selective classifier, but can differ on the group accuracies. Formally, we show in Proposition 6 (Appendix A.2) that it satisfies *equalized odds* (Hardt et al., 2016) with respect to which points it decides to predict or abstain on.

The group-agnostic baseline represents the minimal level of worst-group accuracy that we should hope for. Ideally, selective classification would preferentially increase worst-group accuracy until it matches the other groups. We can capture this optimistic scenario by constructing, for a given selective classifier, a corresponding *Robin Hood oracle* which, as above, also makes the same number of correct and incorrect predictions (see Algorithm 2 in Appendix A.3). However, unlike the group-agnostic baseline, in the Robin Hood oracle the correct predictions are not chosen uniformly at random; instead, we prioritize picking them from the worst group, then the second worst group, etc. Likewise, we prioritize picking the incorrect predictions from the best group, then the second best group, etc. This results in worst-group accuracy rapidly increasing at the cost of the best group.

Both the group-agnostic baseline and the Robin Hood oracle are not implementable in practice without knowledge of all groups and labels. They act instead as references: selective classifiers that preferentially help the worst group over others would have worst-group accuracy-coverage curves that lie between those of the group-agnostic and Robin Hood curves. Unfortunately, Figure 3 shows that the SR selective classifiers substantially underperform their group-agnostic counterparts. We conclude that they are disproportionately helping groups that already have higher accuracies, further exacerbating the disparities between groups. We show similar results for MC-dropout in Section B.1.

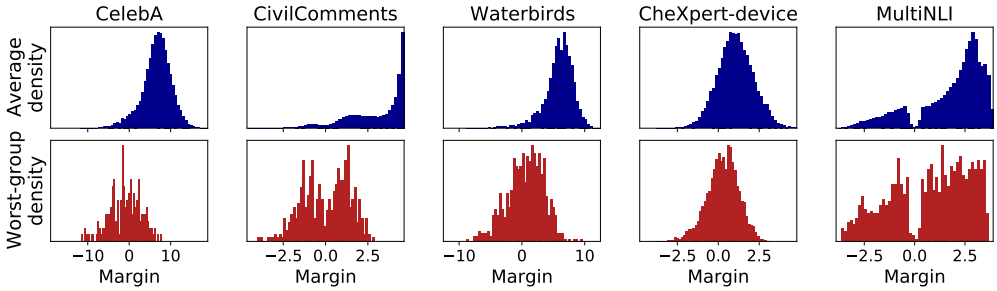

Figure 4: Margin distributions on average (top) and on the worst group (bottom). Positive (negative) margins correspond to correct (incorrect) predictions, and we abstain on points with margins closest to zero first. The worst groups have disproportionately many confident but incorrect examples.

5 Analysis: Margin distributions and accuracy-coverage curves

In the previous section, we saw that selective classification can sometimes increase and sometimes decrease worst-group accuracy, and that even when it does increase worst-group accuracy, it still exacerbates disparities by disproportionately helping other groups. We now turn to a theoretical analysis of how these behaviors arise from the margin distributions, starting in this section with the conditions under which accuracy monotonically increases or decreases with coverage.

Recall that the margin of a selective classifier $(\hat{y}, \hat{c})$ is $\hat{c}(x) \geq 0$ on correct predictions ($\hat{y}(x) = y$) and $-\hat{c}(x) \leq 0$ otherwise. The behavior of selective classification on average versus the worst group is reflected in their different margin distributions, which we show in Figure 4. For example, the worst-group distributions are consistently shifted to the left with many confident but incorrect examples, whereas most confident examples are correct on average. We will analyze a broad class of symmetric and skew-symmetric distributions that reflect most of these empirical distributions, as well as those reported in prior work (Balasubramanian et al., 2011; Lakshminarayanan et al., 2017).

Setup. We consider distributions over margins that have a differentiable cumulative distribution function (CDF) and a density, denoted by corresponding upper- and lower-case variables (e.g., F and f, respectively). The accuracy of a selective classifier is completely determined by its margin distribution F; at a threshold τ, it makes incorrect predictions on points with margin $\leq -\tau$, correct predictions on points with margin $\geq \tau$, and abstains otherwise. We denote its (selective) accuracy at τ as $A_F(\tau) = (1 - F(\tau))/(F(-\tau) + 1 - F(\tau))$. Since increasing the threshold τ monotonically decreases coverage, we focus on studying accuracy as a function of τ. All proofs are in Appendix D.

5.1 Symmetric distributions

We begin by considering symmetric distributions over margins. For symmetric distributions, monotonicity hinges on a generalization of log-concave distributions, which we term *left-log-concavity*:

Definition 1 (Left-log-concave distributions)**.** *A distribution is left-log-concave if its CDF is log-concave on $(-\infty, \mu]$, where μ is the mean of the distribution.*

Left-log-concave distributions are a strict superset of the broad family of log-concave distributions (e.g., Gaussian, beta, uniform), which require log-concave densities (instead of CDFs) on their entire support. Notably, left-log-concave distributions can be multimodal: a symmetric mixture of two Gaussians is left-log-concave, but not typically log-concave (Lemma 1 in Appendix D).

We establish a tight connection between the monotonicity of the accuracy-coverage curve and left-log-concavity, and show that for selective classifiers with left-log-concave margin distributions, the full-coverage accuracy alone determines whether accuracy will increase or decrease with τ.

Proposition 1 (Left-log-concavity and monotonicity)**.** *Let F be the CDF of a symmetric distribution. If F is left-log-concave, then $A_F(\tau)$ is monotonically increasing in τ if $A_F(0) \geq 1/2$ and monotonically decreasing otherwise. Conversely, if $A_{F_d}(\tau)$ is monotonically increasing for all translations F_d such that $F_d(\tau) = F(\tau - d)$ for all τ and $A_{F_d}(0) \geq 1/2$, then F is left-log-concave.*

Proposition 1 is consistent with the observation that selective classification improves average accuracy but hurts worst-group accuracy. As an illustration, consider a symmetric mixture of two

Gaussians whose average accuracy is >50% at full coverage, but where the worst-group accuracy (on one of the Gaussians) is <50%. Since both the overall and worst-group margin distributions are left-log-concave (Lemma 1), Proposition 1 implies that worst-group accuracy will decrease monotonically with τ while the average accuracy improves monotonically with τ.

Applied to CelebA (Figure 4), Proposition 1 is also consistent with how average accuracy improves but worst-group accuracy, which is <50% at full coverage, worsens. Beyond our group setting, Proposition 1 helps to explain why selective classification typically improves average accuracy in the literature, since average accuracies are typically high at full coverage and margin distributions often resemble Gaussians (Balasubramanian et al., 2011; Lakshminarayanan et al., 2017).

5.2 Skewed distributions

Empirically, we also observe skewed margin distributions in Figure 4. To study the effect of skew on accuracies, we consider *skew-symmetric distributions* (Azzalini & Regoli, 2012):

Definition 2. *A distribution with density $f_{\alpha,\mu}$ is skew-symmetric with skew α and center μ if*

$$f_{\alpha,\mu}(\tau) = 2h(\tau - \mu)G(\alpha(\tau - \mu)) \tag{4}$$

for all $\tau \in \mathbb{R}$, where h and G are the density and CDF of distributions that are symmetric about 0.

In other words, $f_{\alpha,\mu}$ is a skewed form of the symmetric density h, where higher α means more right skew, and setting $\alpha = 0$ yields a (translated) h. Skew-symmetric distributions are a broad family and include, for example, the common skew-normal distribution.

We first show that right skew monotonically increases accuracy. Intuitively, since points with positive margin are correctly predicted, right skew improves accuracy and left skew hurts accuracy.

Proposition 2 (Accuracy is monotone with skew)**.** *Let $F_{\alpha,\mu}$ be the CDF of a skew-symmetric distribution. For all $\tau \geq 0$, $A_{F_{\alpha,\mu}}(\tau)$ is monotonically increasing in α.*

In addition, skew in the corresponding direction preserves monotone accuracy-coverage curves:

Proposition 3 (Skew in the same direction preserves monotonicity)**.** *Let $F_{\alpha,\mu}$ be the CDF of a skew-symmetric distribution. If $\mu \geq 0$ and $A_{F_{0,\mu}}(\tau)$ is monotonically increasing in τ, then $A_{F_{\alpha,\mu}}(\tau)$ is also monotonically increasing in τ for any $\alpha > 0$. Similarly, if $\mu \leq 0$ and $A_{F_{0,\mu}}(\tau)$ is monotonically decreasing in τ, then $A_{F_{\alpha,\mu}}(\tau)$ is also monotonically decreasing in τ for any $\alpha < 0$.*

These results are consistent with the mixed worst-group accuracies in Waterbirds and CheXpert-device (Figure 3), which have full-coverage accuracies >0.50 and left-skewed margin distributions (skewness -0.30 and -0.33 respectively; Figure 4). On a margin distributions with no skew (Proposition 1) or right skew (Proposition 3), we would expect accuracy to monotonically increase if full-coverage accuracy is >0.50. However, from Proposition 2, left skew decreases accuracies, and Proposition 3 does not apply since left skew need not preserve monotonically increasing accuracy.

Discussion. The results above establish left-log-concavity as a key property of the margin distribution that determines the monotonicity of the accuracy-coverage curve on symmetric distributions and consequently on their skew-symmetric counterparts. Since accuracy-coverage curves are preserved under all odd, monotone transformations of margins (Lemma 8 in Appendix D), these results also generalize to odd, monotone transformations of these (skew-)symmetric distributions. Many margin distributions—in Figure 4 and in the broader literature (Balasubramanian et al., 2011; Lakshminarayanan et al., 2017)—resemble the distributions studied above (e.g., Gaussians and skewed Gaussians). From our analysis, we thus expect selective classification to improve average accuracy but worsen worst-group accuracy when worst-group accuracy at full coverage is low to begin with.

6 Analysis: Comparison to group-agnostic baseline

Even if selective classification improves worst-group accuracy, it can still exacerbate group disparities in the sense of underperforming the equalized-odds group-agnostic baseline on the worst group, as observed in Section 4. In this section, we continue our analysis of the margin distribution and show that while it is possible to outperform the baseline, it is challenging to do so, especially when the accuracy disparity at full coverage is large.

Setup. Throughout this section, we decompose the margin distribution into two components $F = pF_{\text{wg}} + (1-p)F_{\text{others}}$, where F_{wg} and F_{others} correspond to the margin distributions of the worst group and of all other groups combined, respectively, and the worst group has strictly worse accuracy at full coverage than the other groups (i.e., $A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0)$). Recall from Section 4 that for any selective classifier (i.e., any margin distribution), its group-agnostic baseline has the same average accuracy at each threshold τ but potentially different group accuracies. We denote the worst-group accuracy of the baseline as $\tilde{A}_{F_{\text{wg}}}(\tau)$, which can be written in terms of F_{wg}, F_{others}, and p (Appendix A.2). Other notation follows Section 5, and all proofs are in Appendix E.

A selective classifier with margin distribution F is said to outperform the baseline on the worst group if $\tilde{A}_{F_{\text{wg}}}(\tau) \leq A_{F_{\text{wg}}}(\tau)$ for *all* $\tau \geq 0$. To establish a necessary condition for outperforming the baseline, we study the neighborhood of $\tau = 0$, which corresponds to full coverage:

Proposition 4 (Necessary condition for outperforming the baseline)**.** *Assume that* $1/2 < A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0) < 1$ *and* $f_{\text{wg}}(0) > 0$*. If* $\tilde{A}_{F_{\text{wg}}}(\tau) \leq A_{F_{\text{wg}}}(\tau)$ *for all* $\tau \geq 0$*, then*

$$\frac{f_{\text{others}}(0)}{f_{\text{wg}}(0)} \leq \frac{1 - A_{F_{\text{others}}}(0)}{1 - A_{F_{\text{wg}}}(0)}. \tag{5}$$

The RHS is the ratio of full-coverage errors; the larger the disparity between the worst group and the other groups at full coverage, the harder it is to satisfy this condition. In Appendix F, we simulate mixtures of Gaussians and show that this condition is rarely fulfilled.

Motivated by the empirical margin distributions, we apply Proposition 4 to the setting where F_{wg} and F_{others} are both log-concave and are translated and scaled versions of each other. We show that if the worst group is of a larger scale than the other groups, it underperforms the baseline at some τ:

Corollary 1 (Outperforming the baseline requires smaller scaling for log-concave distributions)**.** *Assume that* $0.5 < A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0) < 1$*,* F_{wg} *is log-concave, and* $f_{\text{others}}(\tau) = v f_{\text{wg}}(v(\tau - \mu_{\text{others}}) + \mu_{\text{wg}})$ *for all* $\tau \in \mathbb{R}$*, where* v *is a scaling factor. If* $v > 1$*, then* $\tilde{A}_{F_{\text{wg}}}(\tau) > A_{F_{\text{wg}}}(\tau)$ *for some* $\tau \geq 0$*.*

This is consistent with the empirical margin distributions on Waterbirds: the worst group has higher variance, implying $v > 1$ as v is the ratio of the worst group's standard deviation to the other group's.

A further special case is when F_{wg} and F_{others} are log-concave and translations of each other, without scaling. In this setting, selective classification underperforms the baseline at *all* thresholds τ.

Proposition 5 (Translated, log-concave distributions always underperform the baseline)**.** *Assume* F_{wg} *and* F_{others} *are log-concave and* $f_{\text{others}}(\tau) = f_{\text{wg}}(\tau - d)$ *for all* $\tau \in \mathbb{R}$*. Then for all* $\tau \geq 0$*,*

$$A_{F_{\text{wg}}}(\tau) \leq \tilde{A}_{F_{\text{wg}}}(\tau). \tag{6}$$

This helps to explain our results on CheXpert-device, where the worst-group and average margin distributions are approximately translations of each other, and selective classification significantly underperforms the baseline at all confidence thresholds.

7 Selective classification on group DRO models

Our above analysis suggests that selective classification tends to exacerbate group disparities, especially when the full-coverage disparities are large. This motivates a potential solution: by reducing group disparities at full coverage, models are more likely to satisfy Proposition 4, which states a necessary condition for outperforming the group-agnostic baseline. In this section, we explore this approach by training models using group distributionally robust optimization (group DRO) (Sagawa et al., 2020), which minimizes the worst-group training loss

$$L_{\text{DRO}}(\theta) = \max_{g \in \mathcal{G}} \hat{\mathbb{E}}_{(x,y)|g} \left[\ell(\theta; (x, y)) \right]. \tag{7}$$

Unlike standard models, group DRO relies on group annotations at training time. Consistent with prior work, we found that group DRO models have much smaller full-coverage disparities (Figure 5).

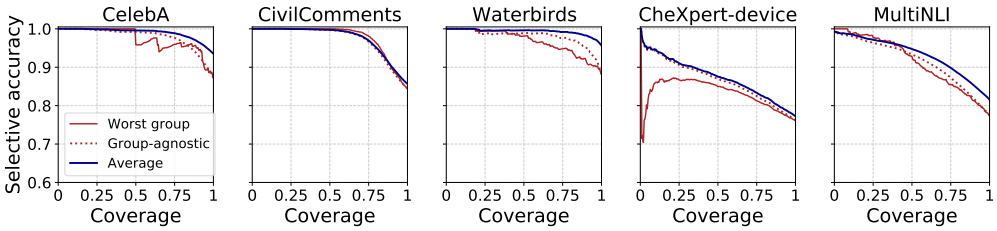

Figure 5: When applied to group DRO models, which have more similar accuracies across groups than standard models, SR selective classifiers improve average and worst-group accuracies. At low coverages, accuracy estimates are noisy as only a few predictions are made.

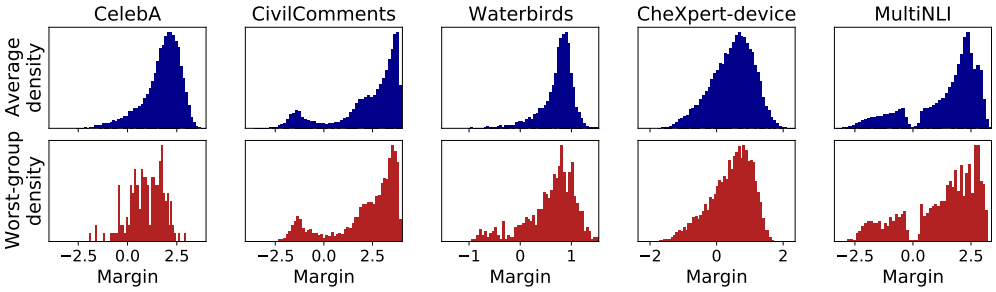

Figure 6: Density of margins for the average (top) and the worst-group (bottom) distributions over margins for the group DRO model. Positive (negative) margins correspond to correct (incorrect) predictions, and we abstain on margins closest to zero first. Unlike the selective classifiers trained with ERM, we see similar average and worst-group distributions for group DRO.

Moreover, worst-group accuracies consistently improve as coverage decreases and at a rate that is largely comparable to the group-agnostic baseline, though gaps remain on Waterbirds and CheXpert.

While our theoretical analysis motivates the above approach, whether selective classification uniformly improves group accuracies ultimately depends on the margin distributions of each group, not just on their full-coverage accuracies. Although group DRO models only optimize for similar full-coverage accuracies across groups, we found that they have strikingly similar average and worst-group margin distributions across all five datasets (Figure 6), explaining why selective classification behaves uniformly over groups in those models.

Group DRO models are not a silver bullet: they rely on group annotations for training, which can be impractical, and in some settings we might not even know in advance which groups to consider. Nevertheless, these results underscore the importance of developing models with uniform full-coverage performance across groups, especially when deployed in tandem with selective classification.

8 Discussion

We have shown that selective classification can magnify group disparities and must therefore be applied with caution. This is an insidious failure mode, since selective classification generally improves average accuracy and can appear to be working well if we do not look at group accuracies. However, we also found that selective classification can still work well on models that have equal full-coverage accuracies across groups. Training such models, especially without relying on too much additional information at training time, remains an important research direction. On the theoretical side, we characterized the behavior of selective classification in terms of the margin distributions; an open question is how different margin distributions arise from different data distributions, models, and training procedures. Another open question is how other selective classification methods beyond softmax-response and Monte Carlo dropout, such as those that jointly train the prediction and confidence functions, fare in terms of group disparities. Finally, in this paper we focused on studying selective accuracy in isolation; accounting for the cost of abstention and the equity of different of different coverages on different groups is an important direction for future work.

REFERENCES

A. Azzalini and G. Regoli. Some properties of skew-symmetric distributions. *Annals of the Institute of Statistical Mathematics*, 64(4):857–879, 2012.

M. A. Badgeley, J. R. Zech, L. Oakden-Rayner, B. S. Glicksberg, M. Liu, W. Gale, M. V. McConnell, B. Percha, T. M. Snyder, and J. T. Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digital Medicine*, 2, 2019.

M. Bagnoli and T. Bergstrom. Log-concave probability and its applications. *Economic Theory*, 26: 445–469, 2005.

K. Balasubramanian, P. Donmez, and G. Lebanon. Unsupervised supervised learning II: Margin-based classification without labels. *Journal of Machine Learning Research (JMLR)*, 12:3119–3145, 2011.

P. L. Bartlett and M. H. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research (JMLR)*, 9(0):1823–1840, 2008.

S. L. Blodgett, L. Green, and B. O'Connor. Demographic dialectal variation in social media: A case study of African-American English. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1119–1130, 2016.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web (WWW)*, pp. 491–500, 2019.

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

I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi. Ethical machine learning in health. *arXiv preprint arXiv:2009.10576*, 2020.

C. K. Chow. An optimum character recognition system using decision functions. In *IRE Transactions on Electronic Computers*, 1957.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 797–806, 2017.

L. P. Cordella, C. D. Stefano, F. Tortorella, and M. Vento. A method for improving classification reliability of multilayer perceptrons. *IEEE Transactions on Neural Networks*, 6(5):1140–1147, 1995.

M. Cule, R. Samworth, and M. Stewart. Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society*, 73:545–603, 2010.

A. De, P. Koley, N. Ganguly, and M. Gomez-Rodriguez. Regression under human assistance. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 2611–2620, 2020.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Association for Computational Linguistics (ACL)*, pp. 4171–4186, 2019.

L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 67–73, 2018.

J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses against mixture covariate shifts. `https://cs.stanford.edu/~thashim/assets/publications/condrisk.pdf`, 2019.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science (ITCS)*, pp. 214–226, 2012.

R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research (JMLR)*, 11, 2010.

J. Feng, A. Sondhi, J. Perry, and N. Simon. Selective prediction-set models with coverage guarantees. *arXiv preprint arXiv:1906.05473*, 2019.

Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Y. Geifman and R. El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning (ICML)*, 2019.

Y. Geifman, G. Uziel, and R. El-Yaniv. Bias-reduced uncertainty estimation for deep neural classifiers. In *International Conference on Learning Representations (ICLR)*, 2018.

S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *Association for Computational Linguistics (ACL)*, pp. 107–112, 2018.

B. Hanczar and E. R. Dougherty. Classification with reject option in gene expression data. *Bioinformatics*, 2008.

M. Hardt, E. Price, and N. Srebo. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3315–3323, 2016.

T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

M. Hellman and J. Raviv. Probability of error, equivocation, and the chernoff bound. *IEEE Transactions on Information Theory*, 16(4):368–372, 1970.

M. E. Hellman. The nearest neighbor classification rule with a reject option. *IEEE Transactions on Systems Science and Cybernetics*, 6(3):179–185, 1970.

D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017.

K. Hill. Wrongfully accused by an algorithm. *The New York Times*, 2020. URL `https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html`.

D. Hovy and A. Søgaard. Tagging performance correlates with age. In *Association for Computational Linguistics (ACL)*, pp. 483–488, 2015.

G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

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 *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 33, pp. 590–597, 2019.

S. Joshi, O. Koyejo, B. Kim, and J. Ghosh. xGEMs: Generating examplars to explain black-box models. *arXiv preprint arXiv:1806.08867*, 2018.

A. Kamath, R. Jia, and P. Liang. Selective question answering under domain shift. In *Association for Computational Linguistics (ACL)*, 2020.

J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, 2017.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

R. T. McCoy, E. Pavlick, and T. Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Association for Computational Linguistics (ACL)*, 2019.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. *arXiv preprint arXiv:2006.01862*, 2020.

L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pp. 151–159, 2020.

Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

J. H. Park, J. Shin, and P. Fung. Reducing gender bias in abusive language detection. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2799–2804, 2018.

M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 2014.

J. M. Porcel. Chest tube drainage of the pleural space: a concise review for pulmonologists. *Tuberculosis and Respiratory Diseases*, 81(2):106–115, 2018.

M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg. Direct uncertainty prediction for medical second opinions. In *International Conference on Machine Learning (ICML)*, pp. 5281–5290, 2019.

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the predictions of any classifier. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016.

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. In *International Conference on Learning Representations (ICLR)*, 2020.

R. Tatman. Gender and dialect bias in YouTube's automatic captions. In *Workshop on Ethics in Natural Langauge Processing*, volume 1, pp. 53–59, 2017.

M. Toplak, R. Močnik, M. Polajnar, Z. Bosnić, L. Carlsson, C. Hasselgren, J. Demšar, S. Boyer, B. Zupan, and J. Stålring. Assessment of machine learning reliability methods for quantifying the applicability domain of QSAR regression models. *Journal of Chemical Information and Modeling*, 54, 2014.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Association for Computational Linguistics (ACL)*, pp. 1112–1122, 2018.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

K. Xiao, L. Engstrom, A. Ilyas, and A. Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

D. Yu, J. Li, and L. Deng. Calibration of confidence measures in speech recognition. *Trans. Audio, Speech and Lang. Proc.*, 19(8):2461–2473, 2011.

B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2017.

A SETUP

A.1 SOFTMAX RESPONSE SELECTIVE CLASSIFIERS

In this section, we describe our implementation of softmax response (SR) selective classifiers (Geifman & El-Yaniv, 2017). Recall from Section 3 that a selective classifier is a pair $(\hat{y}, \hat{c})$, where $\hat{y} : \mathcal{X} \to \mathcal{Y}$ outputs a prediction and $\hat{c} : \mathcal{X} \to \mathbb{R}_+$ outputs the model's confidence, which is always non-negative, in that prediction. SR classifiers are defined for neural networks (for classification), which generally have a last softmax layer over the k possible classes. For an input point x, we denote its maximum softmax probability, which corresponds to its predicted class $\hat{y}(x)$, as $\hat{\mathbb{P}}(\hat{y}(x)|x)$. We defined the confidence $\hat{c}(x)$ for binary classifiers as

$$\hat{c}(x) = \frac{1}{2} \log \left(\frac{\hat{\mathbb{P}}(\hat{y}(x)|x)}{1 - \hat{\mathbb{P}}(\hat{y}(x)|x)} \right). \tag{8}$$

Since the maximum softmax probability $\hat{\mathbb{P}}(\hat{y}(x)|x)$ is at least 0.5 for binary classification, $\hat{c}(x)$ is nonnegative for each x, and is thus a valid confidence. For $k > 2$ classes, however, $\hat{\mathbb{P}}(\hat{y}(x)|x)$ can be less than 0.5, in which case $\hat{c}(x)$ would be negative. To ensure that confidence is always nonnegative, we define $\hat{c}(x)$ for k classes to be

$$\hat{c}(x) = \frac{1}{2} \log \left(\frac{\hat{\mathbb{P}}(\hat{y}(x)|x)}{1 - \hat{\mathbb{P}}(\hat{y}(x)|x)} \right) + \frac{1}{2} \log(k-1). \tag{9}$$

With k classes, the maximum softmax probability $\hat{\mathbb{P}}(\hat{y}(x)|x) \geq 1/k$, and therefore we can verify that $\hat{c}(x) \geq 0$ as desired. Moreover, when $k = 2$, Equation (9) reduces to our original binary confidence; we can therefore interpret the general form in Equation (9) as a normalized logit.

Note that $\hat{c}(x)$ is a monotone transformation of the maximum softmax probability $\hat{\mathbb{P}}(\hat{y}(x)|x)$. Since the accuracy-coverage curve of a selective classifier only depends on the relative ranking of $\hat{c}(x)$ across points, we could have equivalently set $\hat{c}(x)$ to be $\hat{\mathbb{P}}(\hat{y}(x)|x)$. However, following prior work, we choose the logit-transformed version to make the corresponding distribution of confidences easier to visualize (Balasubramanian et al., 2011; Lakshminarayanan et al., 2017).

Finally, we remark on one consequence of SR on the margin distribution for multi-class classification. Recall that we define the *margin* of an example to be $\hat{c}(x)$ on correct predictions ($\hat{y}(x) = y$) and $-\hat{c}(x)$ otherwise, as described in Section 3. In Figure 4, we plot the margin distributions of SR selective classifiers on all five datasets. We observe that on MultiNLI, which is the only multi-class dataset (with $k = 3$), there is a gap (region of lower density) in the margin distribution around 0. We attribute this gap in part to the comparative rarity of seeing a maximum softmax probability of $\frac{1}{3}$ when $k = 3$ versus seeing $\frac{1}{2}$ when $k = 2$; in the former, all three logits must be the same, while for the latter only two logits must be the same.

A.2 GROUP-AGNOSTIC BASELINE

Here, we describe the group-agnostic baseline described in Section 4 in more detail. We elaborate on the construction from the main text and then define the baseline formally. Finally, we show that the group-agnostic baseline satisfies equalized odds.

A.2.1 DEFINITION

We begin by recalling on the construction described in the main text for a finite test set $\mathcal{D}$. In Algorithm 1 we relied on two important sets; the set of correctly classified points at threshold τ, $\mathcal{C}_\tau$:

$$\mathcal{C}_\tau = \{(x, y, g) \in \mathcal{D} \mid \hat{y}(x) = y \text{ and } \hat{c}(x) \geq \tau\}, \tag{10}$$

and similarly, the set of incorrectly classified points at threshold τ, $\mathcal{I}_\tau$:

$$\mathcal{I}_\tau = \{(x, y, g) \in \mathcal{D} \mid \hat{y}(x) \neq y \text{ and } \hat{c}(x) \geq \tau\}. \tag{11}$$

To compute the accuracy of the group-agnostic baseline, we sample a subset $\mathcal{C}_\tau^{\text{ga}}$ of size $|\mathcal{C}_\tau|$ uniformly at random from $\mathcal{C}_0$, and similarly a subset $\mathcal{I}_\tau^{\text{ga}}$ of size $|\mathcal{I}_\tau|$ uniformly at random from $\mathcal{I}_0$. The

group-agnostic baseline makes predictions when examples are in $\mathcal{C}_\tau^{\text{ga}} \cup \mathcal{I}_\tau^{\text{ga}}$ and abstains otherwise. We compute group-accuracies over this set of predicted examples.

Note that the group accuracies, as defined, are randomized due to the sampling. For the remainder of our analysis, we take the expectation over this randomness to compute the group accuracies. We now generalize the above construction by considering data distributions $\mathcal{D}$. We first define the number of correctly and incorrectly classified points:

Definition 3 (Correctly and incorrectly classified points)**.** *Consider a selective classifier $(\hat{y}, \hat{c})$. For each threshold τ, we define the fractions of points that are predicted (not abstained on), and correctly or incorrectly classified, as*

$$C(\tau) = \hat{\mathbb{P}}(\hat{y}(x) = y \wedge \hat{c}(x) \geq \tau), \tag{12}$$

$$I(\tau) = \hat{\mathbb{P}}(\hat{y}(x) \neq y \wedge \hat{c}(x) \geq \tau). \tag{13}$$

We define analogous metrics for each group g as

$$C_g(\tau) = \hat{\mathbb{P}}(\hat{y}(x) = y \wedge \hat{c}(x) \geq \tau \mid g), \tag{14}$$

$$I_g(\tau) = \hat{\mathbb{P}}(\hat{y}(x) \neq y \wedge \hat{c}(x) \geq \tau \mid g). \tag{15}$$

For each threshold τ, we will make predictions on a $C(\tau)/C(0)$ fraction of the $C(0)$ total (probability mass of) correctly classified points. Since each group g has $C_g(0)$ correctly classified points, at threshold τ, the group-agnostic baseline will make predictions on $C_g(0)C(\tau)/C(0)$ correctly classified points in group g. We can reason similarly over the incorrectly classified points. Putting it all together, we can define the group-agnostic baseline as satisfying the following:

Definition 4 (Group-agnostic baseline)**.** *Consider a selective classifier $(\hat{y}, \hat{c})$ and let $\tilde{C}$, $\tilde{I}$, $\tilde{C}_g$, $\tilde{I}_g$ denote the analogous quantities to Definition 3 for its matching group-agnostic baseline. For each threshold τ, these satisfy*

$$\tilde{C}(\tau) = C(\tau) \tag{16}$$

$$\tilde{I}(\tau) = I(\tau), \tag{17}$$

and for each threshold τ and group g,

$$\tilde{C}_g(\tau) = C_g(0)C(\tau)/C(0) \tag{18}$$

$$\tilde{I}_g(\tau) = I_g(0)I(\tau)/I(0). \tag{19}$$

The group-agnostic baseline thus has the following accuracy on group g:

$$\tilde{A}_g(\tau) = \frac{\tilde{C}_g(\tau)}{\tilde{C}_g(\tau) + \tilde{I}_g(\tau)} \tag{20}$$

$$= \frac{C_g(0)C(\tau)/C(0)}{C_g(0)C(\tau)/C(0) + I_g(0)I(\tau)/I(0)}. \tag{21}$$

A.2.2 Connection to equalized odds

We now show that the group-agnostic baseline satisfies equalized odds with respect to which points it predicts or abstains on. The goal of a selective classifier is to make predictions on points it would get correct (i.e., $\hat{y}(x) = y$) while abstaining on points that it would have gotten incorrect (i.e., $\hat{y}(x) \neq y$). We can view this as a meta classification problem, where a true positive is when the selective classifier decides to make a prediction on a point x and gets it correct ($\hat{y}(x) = y$), and a false positive is when the selective classifier decides to make a prediction on a point x and gets it incorrect ($\hat{y}(x) \neq y$). As such, we can define the *true positive rate* $R^{\mathsf{TP}}(\tau)$ and *false positive rate* $R^{\mathsf{FP}}(\tau)$ of a selective classifier:

Definition 5. *The true positive rate of a selective classifier at threshold τ is*

$$R^{TP}(\tau) = \frac{C(\tau)}{C(0)}, \tag{22}$$

and the false positive rate at threshold τ is

$$R^{\mathsf{FP}}(\tau) = \frac{I(\tau)}{I(0)}. \tag{23}$$

Analogously, the true positive and false positive rates on a group g are

$$R_g^{\mathsf{TP}}(\tau) = \frac{C_g(\tau)}{C_g(0)}, \tag{24}$$

$$R_g^{\mathsf{FP}}(\tau) = \frac{I_g(\tau)}{I_g(0)}. \tag{25}$$

The group-agnostic baseline satisfies equalized odds (Hardt et al., 2016) with respect to this definition:

Proposition 6. *The group-agnostic baseline defined in Definition 4 has equal true positive and false positive rates for all groups $g \in \mathcal{G}$ and satisfies equalized odds.*

Proof. By construction of the group-agnostic baseline (Definition 4), we have that

$$\tilde{C}_g(\tau) = C_g(0)C(\tau)/C(0) \tag{26}$$

$$= \tilde{C}_g(0)\tilde{C}(\tau)/\tilde{C}(0), \tag{27}$$

and therefore, for each group g, we can show that the true-positive rate of the group-agnostic baseline $\tilde{R}_g^{\mathsf{TP}}$ on g is equal to its average true-positive rate $\tilde{R}^{\mathsf{TP}}(\tau)$.

$$\tilde{R}_g^{\mathsf{TP}}(\tau) = \frac{\tilde{C}_g(\tau)}{\tilde{C}_g(0)} \tag{28}$$

$$= \tilde{C}(\tau)/\tilde{C}(0) \tag{29}$$

$$= \tilde{R}^{\mathsf{TP}}(\tau). \tag{30}$$

Each group thus has the same true positive rate with the group-agnostic baseline. Using similar reasoning, each group also has the same false positive rate. By the definition of equalized odds, the group-agnostic baseline thus satisfies equalized odds. □

A.3 Robin Hood oracle

In this section, we define the *Robin Hood oracle*, which preferentially increases the worst-group accuracy through abstentions until it matches the other groups. Like the group-agnostic baseline, we constrain the Robin Hood oracle to make the same number of correct and incorrect predictions as a given selective classifier.

Algorithm 2: Robin Hood oracle at threshold τ

Input: Selective classifier $(\hat{y}, \hat{c})$, threshold τ, test data $\mathcal{D}$

Output: The set of points $\mathcal{P} \subseteq \mathcal{D}$ that the Robin Hood oracle for $(\hat{y}, \hat{c})$ makes predictions on at threshold τ.

1 In Algorithm 1, we defined $\mathcal{C}_\tau$ and $\mathcal{I}_\tau$ to be the sets of correct and incorrect points predicted on and abstained on at threshold τ respectively. Define $\mathcal{Q}$, the set of subsets of $\mathcal{D}$ that have the same number of correct and incorrect predictions as $(\hat{y}, \hat{c})$ at threshold τ as follows:

$$\mathcal{Q} = \{\mathcal{S} \subseteq \mathcal{D} \mid |\mathcal{P} \cap \mathcal{C}_0| = |\mathcal{C}_\tau|, |\mathcal{P} \cap \mathcal{I}_0| = |\mathcal{I}_\tau|\}. \tag{31}$$

2 Let $\mathbf{acc}_g(\mathcal{S})$ be the accuracy of $\hat{y}$ on points in $\mathcal{S}$ that belong to group g. We return the set $\mathcal{P}$ that minimizes the difference between the best-group and worst-group accuracies.

$$\mathcal{P} = \operatorname*{argmin}_{\mathcal{S} \in \mathcal{Q}} \left(\max_{g \in \mathcal{G}} \mathbf{acc}_g(\mathcal{S}) - \min_{g \in \mathcal{G}} \mathbf{acc}_g(\mathcal{S}) \right). \tag{32}$$

We formalize the definition of the Robin Hood oracle in Algorithm 2. The Robin Hood oracle makes predictions on the subset of examples from $\mathcal{D}$ that has the smallest discrepency between best-group and worst-group accuracies, while still matching the number of correct and incorrect predictions of a given selective classifier. Since enumerating over all possible subsets of the test data $\mathcal{D}$ is intractable, we compute the accuracies of the Robin Hood oracle iteratively. Starting from full coverage, we abstain on examples from lowest to highest confidence. Whenever we abstain on an incorrect example, we assume it comes from the current lowest accuracy group, and similarly whenever we abstain on a correctly classified example we assume it comes from the current highest accuracy group. This reduces disparities to the maximum extent possible as the threshold τ increases.

B SUPPLEMENTAL EXPERIMENTS

B.1 MONTE-CARLO DROPOUT

In the main text, we observed that SR selective classifiers monotonically improve average accuracy as coverage decreases, but exacerbate accuracy disparities across groups on all five datasets. To demonstrate that these observations are not specific to SR selective classifiers, we now present our empirical results on another standard selective classification method: Monte-Carlo (MC) dropout (Gal & Ghahramani, 2016; Geifman & El-Yaniv, 2017). We find that MC-dropout selective classifiers exhibit similar empirical trends as SR selective classifiers.

MC-dropout selective classifiers. MC-dropout is an alternate way of assigning confidences to points. Taking a model with a dropout layer, the selective classifier first predicts $\hat{y}(x)$ simply by taking the model output without dropout. To estimate the confidence, it then samples n softmax probabilities corresponding to the label $\hat{y}(x)$ over the randomness of the dropout layer. The confidence is computed as $\hat{c}(x) = 1/s$, where s^2 is the variance of the sampled probabilities. We implement MC-dropout by using the existing dropout layers for BERT and adding dropout to the final fully-connected layer for ResNet and DenseNet, with a dropout probability 0.1. We present empirical results for $n = 10$; Gal & Ghahramani (2016) observed that this was sufficient to produce good confidence estimates.

Results. The MC-dropout selective classifiers exhibit similar trends to those that observed in Section 4, demonstrating that the observed empirical trends are not specific to SR response; even though the average accuracy improves monotonically as coverage decreases across all five datasets, the worst-group accuracy tends to decrease for CelebA and CivilComments, fails to increase significantly for CheXpert and Waterbirds, and increases consistently but slowly for MultiNLI. Comparing SR and MC-dropout selective classifiers, we observe that the MC-dropout selective classifiers performs slightly worse. For example, we see a more prominent drop in worst-group accuracy for Waterbirds and much smaller improvements in worst-group accuracy for CheXpert.

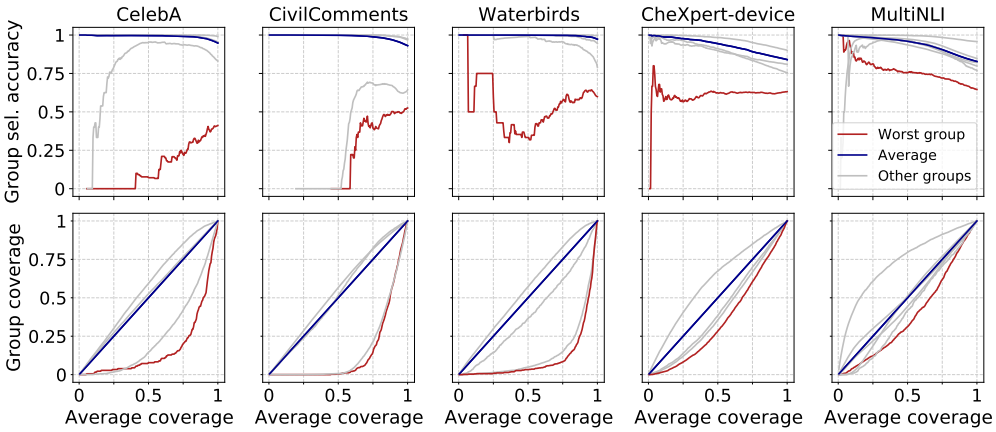

Figure 7: Accuracy (top) and coverage (bottom) for each group, as a function of the average coverage for the MC-dropout selective classifier. Each average coverage corresponds to a threshold τ. The red lines represent the worst group.

B.2 GROUP DRO

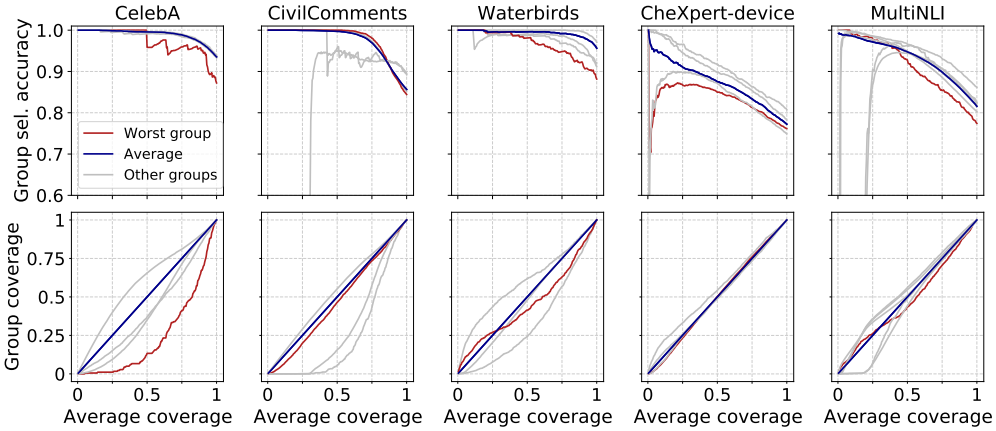

Figure 8: Accuracy (top) and coverage (bottom) for each group as a function of the average coverage for the softmax response selective classifier with model optimized with group DRO. Each average coverage corresponds to a specific threshold τ. The red lines represent the worst group (i.e. the one with the lowest accuracy at full coverage,) and the grey lines represent the other groups.

We showed in Section 7 that SR selective classifiers trained with the group DRO objective successfully improve worst-group accuracies as coverage decreases, and perform comparably to the group-agnostic baseline. We now present additional empirical results for these selective classifiers.

Accuracy-coverage curves. We first present the accuracy-coverage curves for all groups, along with the group coverage trends, in Figure 8. Group accuracies tend to improve monotonically, in stark contrast with our results on standard (ERM) selective classifiers. These trends hold generally across groups and datasets, with a few exceptions; accuracies drop on two groups in MultiNLI at roughly 20% average coverage, and accuracies improve slowly on two groups in CivilComments. Below, we look into these anomalies and offer potential explanations.

For MultiNLI, we first note that the drops are observed at very low *group* coverages—lower than 1%—at which point the accuracies are computed based on very few examples and thus are noisy. In addition, much of the anomaly can be explained by label noise; we manually inspect the 20 examples from the above two groups with the highest softmax confidences, and find that 17 of them are labeled incorrectly. The observations on CivilComments can also be potentially attributed to label noise. Removing examples with high inter-annotator disagreement (with the fraction of toxic annotations between 0.5 and 0.6) yields an accuracy-coverage curve that fits the broader empirical trends.

C EXPERIMENT DETAILS

C.1 DATASETS

CelebA. Models have been shown to latch onto spurious correlations between labels and demographic attributes such as race and gender (Buolamwini & Gebru, 2018; Joshi et al., 2018), and we study this on the CelebA dataset (Liu et al., 2015). Following Sagawa et al. (2020), we consider the task of classifying hair color, which is spuriously correlated with the gender. Concretely, inputs are celebrity face images, labels are hair color $\mathcal{Y} = \{\text{blond}, \text{non-blond}\}$, and spurious attributes are gender, $\mathcal{A} = \{\text{male}, \text{female}\}$, with blondness associated with being female. Of the four groups, blond males are the smallest group, with only 1,387 examples out of 162,770 training examples, and they tend to be the worst group empirically. We use the official train-val-split of the dataset.

Waterbirds. Object recognition models are prone to using image backgrounds as a proxy for the label (Ribeiro et al., 2016; Xiao et al., 2020). We study this on the Waterbirds dataset (Sagawa

et al., 2020), constructed using images of birds from the Caltech-UCSD Birds dataset (Wah et al., 2011) placed on backgrounds from the Places dataset (Zhou et al., 2017). The task is to classify a photograph of a bird as one of $\mathcal{Y} = \{\text{waterbird}, \text{landbird}\}$, and the label is spuriously correlated with the background $\mathcal{A} = \{\text{water background}, \text{land background}\}$. Of the four groups, waterbirds on land backgrounds make up the smallest group with only 56 examples out of 4,795 training examples, and they tend to be the worst group empirically. We use the train-val-split provided by Sagawa et al. (2020), and also follow their protocol for computing average metrics; to compute average accuracies and coverages, we first compute the metrics for each group and obtain a weighted average according to group proportions in the training set, in order to account for the discrepancy in group proportions across the splits.

CheXpert-device. Models can latch onto spurious correlations even in high-stakes applications such as medical imaging. When models are trained to classify whether a patient has certain pathologies from chest X-rays, models have been shown to spuriously detect the presence of a support device, in particular a chest drain, instead (Oakden-Rayner et al., 2020). We study this phenomenon in a modified version of the CheXpert dataset (Irvin et al., 2019), which we call CheXpert-device. Concretely, the inputs are chest X-rays, labels are $\mathcal{Y} = \{\text{pleural effusion}, \text{no pleural effusion}\}$, and spurious attributes indicate the the presence of a support device, $\mathcal{A} = \{\text{support device}, \text{no support device}\}$. We note that chest drain is one type of support device, and is used to treat suspected pleural effusion (Porcel, 2018).

CheXpert-device is a subsampled version of the full CheXpert dataset that manifests the spurious correlation more strongly. To create CheXpert-device, we first create a new 80/10/10 train/val/test split of examples from the publicly available CheXpert train and validation sets, randomly assigning patients to splits so that all X-rays of the same patient fall in the same split. We then subsample the training set; in particular, we enforce that in 90% of examples, the label of support device matches the pleural effusion label. Of the four groups, cases of pleural effusion without a support device make up the smallest group, with 5,467 examples out of 112,100 training examples, and they tend to be the worst group empirically.

To compute the average accuracies and coverages, we weight groups according to group proportions in the training set, similarly to Waterbirds. Another complication with CheXpert is that some patients can have multiple X-rays from one visit. Following Irvin et al. (2019), we treat these images as separate training examples at training time, but output one prediction for each patient-study pair at evaluation time. Concretely, we predict pleural effusion if the model detects the condition in *any* of the X-ray images belonging to the patient-study pair, as pathologies may only appear clearly in some X-rays.

CivilComments. In toxicity comment detection, models have been shown to latch onto spurious correlations between the toxicity and mention of certain demographic groups (Park et al., 2018; Dixon et al., 2018). We study this in the CivilComments dataset (Borkan et al., 2019). The task is to classify the toxicity of comments on online articles with labels $\mathcal{Y} = \{\text{toxic}, \text{non-toxic}\}$. As spurious attributes, we consider whether each comment mentions a Christian identity, $\mathcal{A} = \{\text{mention of Christian identity}, \text{no mention of Christian identity}\}$; non-toxicity is associated with the mention of Christian identity, often resulting in high false negative rate on comments with such mentions. Of the four groups, toxic comments with a mention of Christian identity make up the smallest group, with only 3,033 examples out of 359,950 training examples, and they tend to be the worst group empirically. We use the official development and test sets from the associated Kaggle competition,[1] and further split the development set into a training set and a validation set by randomly splitting on articles and associating all comments with each article to either set. The training, validation, and test set all comprise comments from disjoint sets of articles. The original dataset also contains many additional examples that have toxicity annotations but not identity annotations; we do not use these in our experiments.

In the original CivilComments dataset, each comment is given a probabilistic labels for both the toxicity and the mention of a Christian identity, where a probabilistic label is the average of binary labels across annotators. Following the associated Kaggle competition, we use binarized labels obtained by thresholding the probabilistic labels at 0.5.

[1] `www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/`

MultiNLI. Lastly, we consider natural langage inference (NLI), where the task is to predict whether a hypothesis is entailed, contradicted by, or neutral to an associated premise, $\mathcal{Y} = \{\text{entailed}, \text{contradictory}, \text{neutral}\}$. NLI models have been shown to exploit annotation artifacts, for example predicting contradictory whenever negation words such as *never* or *nobody* are present (Gururangan et al., 2018). We study this on the MultiNLI dataset (Williams et al., 2018). To annotate examples' spurious attributes $\mathcal{A} = \{\text{negation words}, \text{no negation words}\}$, we consider the following negation words following Gururangan et al. (2018): "*nobody*", "*no*", "*never*", and "*nothing*". We use the splits used in Sagawa et al. (2020), whose training set includes 206,175 examples with 1,521 examples from the smallest group (entailment with negations). The worst group for standard models tends to be neutral examples with negation words.

C.2 Models

We train ResNet (He et al., 2016) for CelebA and Waterbirds (images), DenseNet (Huang et al., 2017) for CheXpert (X-rays), and BERT (Devlin et al., 2019) for CivilComments and MultiNLI (text). For tasks studied in Sagawa et al. (2020) (CelebA, Waterbirds, MultiNLI), we use the hyperparameters from Sagawa et al. (2020). For others (CivilComments and CheXpert-device), we test the same number of hyperparameter sets for each of ERM and DRO, and report the best set below. Across all image and X-ray tasks, inputs are downsampled to resolution 224 x 224.

CelebA. To train a model on CelebA, we initialize to pretrained ResNet-50. For ERM we optimize with learning rate 1e-4, weight decay 1e-4, batch size 128, and train for 50 epochs. For DRO we use learning rate 1e-5, weight decay 1e-1, and use generalization adjustment 1 (described in Sagawa et al. (2020)). The batch size is 128, and we train for 50 epochs.

Waterbirds. For Waterbirds, as with CelebA, we use pretrained ResNet-50 as an initialization. For ERM we use learning rate 1e-3, weight decay 1e-4, batch size 128, and train for 300 epochs. For DRO we use learning rate 1e-5, weight decay 1, geneneralization adjustment 1, batch size 128, and train for 300 epochs.

CheXpert-device. For CheXpert-device, we fine-tune pretrained DenseNet-121 for three epochs. For ERM we use learning rate 1e-3, no weight decay, batch size 16, and choose the model (out of the first three epochs) with highest average accuracy (epoch 2). For DRO, we use learning rate 1e-4, weight decay 1e-1, and batch size 16, and choose the model with highest worst-group accuracy (epoch 1).

CivilComments. To train a model for CivilComments, we fine-tune bert-base-uncased using the implementation from Wolf et al. (2019). For both ERM and DRO we use learning rate 1e-5, weight decay 1e-2, and batch size 16. We train the ERM models and DRO models for three epochs (early stopping,) then choose the model with highest average accuracy for ERM (epoch 2), and highest worst-group accuracy for DRO (epoch 1).

MultiNLI. For MultiNLI we again fine-tune bert-base-uncased using the implementation from Wolf et al. (2019). For ERM, we fine-tune for three epochs with learning rate 2e-5, weight decay 0, and batch size 32. For DRO, we also use learning rate 2e-5, and weight decay 0, and batch size 32, but use generalization adjustment 1. For both ERM and DRO the model after the third epoch is best in terms of average accuracy for ERM and worst-group accuracy for DRO.

D Proofs: Margin distributions and accuracy-coverage curves

D.1 Left-log-concavity

Recall the definition of left-log-concave from Section 5:

Definition 1 (Left-log-concave distributions)**.** *A distribution is left-log-concave if its CDF is log-concave on* $(-\infty, \mu]$*, where* μ *is the mean of the distribution.*

We first prove that a symmetric mixture of gaussians is left-log-concave, but not necessarily log-concave.

D.1.1 SYMMETRIC MIXTURE OF GAUSSIANS IS LEFT-LOG-CONCAVE

Lemma 1 (Symmetric mixture of two gaussians is left-log-concave)**.** *Consider a symmetric mixture of Gaussians with density $f = 0.5f_\mu + 0.5f_{-\mu}$, where f_μ is the density of a $\mathcal{N}(\mu, \sigma^2)$ random variable and likewise $f_{-\mu}$ is the density of a $\mathcal{N}(-\mu, \sigma^2)$ random variable. Then the mixture is left-log-concave for all values of $\mu \in \mathbb{R}, \sigma > 0$, but only log-concave if $|\mu| \leq \sigma$.*

Proof. Without loss of generality, we can take $\sigma = 1$, since (left-)log-concavity is invariant to scaling, and also assume that μ is positive. First consider the case where $\mu \leq 1$. Then the mixture is log-concave, and therefore left-log-concave (Cule et al., 2010).

Now, consider the case where $\mu > 1$. Cule et al. (2010) show the mixture is no longer log-concave.[2] However, we claim that it is still left-log-concave. We start by studying the gradient of $\log f$,

$$\frac{f'(x)}{f(x)} = \frac{-\exp\left(-\frac{(x-\mu)^2}{2}\right)(x-\mu) - \exp\left(-\frac{(x+\mu)^2}{2}\right)(x+\mu)}{\exp\left(-\frac{(x-\mu)^2}{2}\right) + \exp\left(-\frac{(x+\mu)^2}{2}\right)} \tag{33}$$

$$= -x + \mu\left[\frac{1-\exp(-2x\mu)}{1+\exp(-2x\mu)}\right]. \tag{34}$$

We claim that $\frac{f'(x)}{f(x)}$ has a local minimum at $x = -a < 0$, and as it is an odd function, a corresponding local maximum at $x = a > 0$. To show this, we first differentiate to obtain

$$\frac{d}{dx}\frac{f'(x)}{f(x)} = -1 + \frac{4\mu^2\exp(-2x\mu)}{(1+\exp(-2x\mu))^2}. \tag{35}$$

Setting the derivative to 0 gives us the quadratic equation

$$0 = -1 + \frac{4\mu^2\exp(-2x\mu)}{(1+\exp(-2x\mu))^2} \tag{36}$$

$$\iff \quad (1+\exp(-2x\mu))^2 = 4\mu^2\exp(-2x\mu) \tag{37}$$

$$\iff \quad [\exp(-2x\mu)]^2 + (2-4\mu^2)\exp(-2x\mu) + 1 = 0 \tag{38}$$

$$\iff \quad \exp(-2x\mu) = 2\mu^2 - 1 \pm 2\mu\sqrt{\mu^2-1}. \tag{39}$$

Since $\mu > 1$, there are two distinct roots of this quadratic, and two corresponding critical points of f'/f. Let $v(\mu) = 2\mu^2 - 1 + 2\mu\sqrt{\mu^2-1}$ be the larger root. Then $v(\mu)$ is a strictly increasing function for $\mu \geq 1$, and since $v(1) = 1$, we have that $v(\mu) > 1$ for all $\mu > 1$. Let $x = -a$ satisfy $\exp(2a\mu) = v(\mu)$. Then we have that $-a$, the smaller of the critical points of f'/f, is

$$-a = -\frac{\log v(\mu)}{2\mu} \tag{40}$$

$$= -\frac{\log(2\mu^2 - 1 + 2\mu\sqrt{\mu^2-1})}{2\mu} \tag{41}$$

$$< 0. \tag{42}$$

To show that $f'(a)/f(a)$ is a local minimum, we take the second derivative

$$\frac{d^2}{dx^2}\frac{f'(x)}{f(x)} = \frac{d}{dx}\left[-1 + \frac{4\mu^2\exp(-2x\mu)}{(1+\exp(-2x\mu))^2}\right] \tag{43}$$

$$= \frac{8\mu^3\exp(-2x\mu)(\exp(-2x\mu)-1)}{(\exp(-2x\mu)+1)^3}, \tag{44}$$

which at $x = -a$ gives

$$\frac{d^2}{dx^2}\frac{f'(x)}{f(x)}|_{x=-a} = \frac{8\mu^3 v(\mu)(v(\mu)-1)}{(v(\mu)+1)^3} \tag{45}$$

$$> 0, \tag{46}$$

[2]This is a technical sleight-of-hand: the citation just says that f is not LC, but from experiments F seems to be LC until roughly $\mu = 1.4$.

since $v(\mu) > 1$. Since $-a$ is the only critical point of f'/f that is less than 0 and it is a local minimum, f'/f must be decreasing on $(-\infty, -a]$, which in turn implies that f, and therefore F, is log-concave on $(-\infty, -a]$.

It remains to show that F is also log-concave on $[-a, 0]$. We make use of two facts. First, since $-a$ is a local minimum and the only critical point less than 0, we have that $\frac{f'(-a)}{f(-a)} \leq \frac{f'(x)}{f(x)} \leq \frac{f'(0)}{f(0)} = 0$ for all $x \in [-a, 0]$. Second, since $f(x)$ and $F(x)$ are non-negative for all x, $f(x)/F(x)$ is also non-negative for all x. Thus, for all $x \in [-a, 0]$

$$\frac{d}{dx}\frac{f(x)}{F(x)} = \frac{F(x)f'(x) - f(x)^2}{F(x)^2} \tag{47}$$

$$= \underbrace{\frac{f(x)}{F(x)}}_{\geq 0}\left(\underbrace{\frac{f'(x)}{f(x)}}_{\leq 0} - \underbrace{\frac{f(x)}{F(x)}}_{\geq 0}\right) \tag{48}$$

$$\leq 0, \tag{49}$$

and therefore F is also log-concave on $[-a, 0]$. □

Remark 1. *Note that if f is (left-)log-concave, then F is also (left-)log-concave (Bagnoli & Bergstrom, 2005). However, the reverse direction does not hold.*

D.2 Symmetric distributions

In this section we prove Proposition 1. We first prove the following helpful lemma:

Lemma 2 (Conditions for monotonicity of selective accuracy.)**.** *$A_F(\tau)$ is monotone increasing in τ if and only if*

$$\frac{f(-\tau)}{F(-\tau)} \geq \frac{f(\tau)}{1 - F(\tau)} \tag{50}$$

for all $\tau \geq 0$. Conversely, $A_F(\tau)$ is monotone decreasing in τ if and only if the above inequality is flipped for all $\tau \geq 0$.

Proof. $A_F(\tau)$ is monotone increasing in τ if and only if $\frac{dA_F}{d\tau} \geq 0$ for all $\tau \geq 0$. We obtain $\frac{dA_F}{d\tau}$ by differentiating A_F and simplifying:

$$\frac{dA_F}{d\tau} = \frac{d}{d\tau}\left(\frac{1 - F(\tau)}{1 - F(\tau) + F(-\tau)}\right) \tag{51}$$

$$= \frac{\big[1 - F(\tau) + F(-\tau)\big]\big[-f(\tau)\big] - \big[1 - F(\tau)\big]\big[-f(\tau) - f(-\tau)\big]}{(1 - F(\tau) + F(-\tau))^2} \tag{52}$$

$$= \frac{-f(\tau) + f(\tau)F(\tau) - f(\tau)F(-\tau) + f(\tau) + f(-\tau) - f(\tau)F(\tau) - f(-\tau)F(\tau)}{(1 - F(\tau) + F(-\tau))^2}$$

$$= \frac{f(-\tau) - f(\tau)F(-\tau) - f(-\tau)F(\tau)}{(1 - F(\tau) + F(-\tau))^2} \tag{53}$$

$$= \frac{f(-\tau)[1 - F(\tau)] - f(\tau)F(-\tau)}{(1 - F(\tau) + F(-\tau))^2}. \tag{54}$$

Since the denominator is always positive, we have that $\frac{dA_F}{d\tau} \geq 0$ if and only if the numerator $f(-\tau)[1 - F(\tau)] - f(\tau)F(-\tau) \geq 0$, which in turn is equivalent to

$$\frac{f(-\tau)}{F(-\tau)} \geq \frac{f(\tau)}{1 - F(\tau)}, \tag{55}$$

as desired. The case for monotone decreasing $A_F(\tau)$ is analogous. □

In the next two lemmas, we prove the necessary and sufficient conditions for Proposition 1 respectively.

Lemma 3 (Left-log-concavity and symmetry imply monotonicity.)**.** *Let f be symmetric about μ and let F be left-log-concave. If $A_F(0) \geq 0.5$, then $A_F(\tau)$ is monotone increasing. Conversely, if $A_F(0) \leq 0.5$, then $A_F(\tau)$ is monotone decreasing.*

Proof. Consider the case where $A_F(0) \geq 0.5$; the case where $A_F(0) \leq 0.5$ is analogous. From Lemma 2 and the symmetry of f, we have that $A_F(\tau)$ is monotone increasing if (and only if)

$$\frac{f(-\tau)}{F(-\tau)} \geq \frac{f(\tau)}{1 - F(\tau)} = \frac{f(2\mu - \tau)}{F(2\mu - \tau)} \tag{56}$$

holds for all $\tau \geq 0$.

To show that this inequality holds for all $\tau \geq 0$, we first note that since $A_F(0) \geq 0.5$, we have that $F(0) \leq 0.5$, which together with the symmetry of F implies that $\mu \geq 0$. Thus, $-\tau \leq 2\mu - \tau$ for all $\tau \geq 0$.

Now, if $\tau \geq \mu$, then $2\mu - \tau \leq \mu$, so the desired inequality (56) follows from the log-concavity of F on $(-\infty, \mu]$ (Remark 1 of Bagnoli & Bergstrom (2005).)

If instead $0 \leq \tau \leq \mu$, we have

$$\frac{f(-\tau)}{F(-\tau)} \geq \frac{f(\tau)}{F(\tau)} \qquad \text{[by log-concavity of } F \text{ on } (-\infty, \mu]\text{]} \tag{57}$$

$$\geq \frac{f(\tau)}{1 - F(\tau)} \qquad \text{[since } F(\tau) \leq 0.5\text{]} \tag{58}$$

and thus (56) also holds. □

Lemma 4 (Monotonicity and symmetry imply left-log-concavity.)**.** *Let f be symmetric with mean 0, and let f_μ be a translated version of f that has mean μ. If A_{f_μ} is monotone increasing for all $\mu \geq 0$, then f is left-log-concave.*

Proof. First consider $\mu \geq 0$. Since A_{f_μ} is monotone increasing for all $\mu \geq 0$, from Lemma 2 and the symmetry of f_μ,

$$\frac{f_\mu(-\tau)}{F_\mu(-\tau)} \geq \frac{f_\mu(\tau)}{1 - F_\mu(\tau)} = \frac{f_\mu(2\mu - \tau)}{F_\mu(2\mu - \tau)} \tag{59}$$

must hold for all $\tau \geq 0$. Since $f_\mu(x) = f(x - \mu)$ by construction, we can equivalently write

$$\frac{f(-\mu - \tau)}{F(-\mu - \tau)} \geq \frac{f(\mu - \tau)}{F(\mu - \tau)}, \tag{60}$$

which holds for all $\tau \geq 0$ as well, and by assumption, for all $\mu \geq 0$ as well. By letting $a = -\mu - \tau$ and $b = \mu - \tau$, we can equivalently write the above inequality as

$$\frac{f(a)}{F(a)} \geq \frac{f(b)}{F(b)} \tag{61}$$

for all $a, b \in \mathbb{R}$ where $a \leq b, a + b \leq 0$.

By the definition of log-concavity, this means that F is log-concave on $(-\infty, 0]$. □

D.2.1 LEFT-LOG-CONCAVITY AND MONOTONICITY.

Proposition 1 (Left-log-concavity and monotonicity)**.** *Let F be the CDF of a symmetric distribution. If F is left-log-concave, then $A_F(\tau)$ is monotonically increasing in τ if $A_F(0) \geq 1/2$ and monotonically decreasing otherwise. Conversely, if $A_{F_d}(\tau)$ is monotonically increasing for all translations F_d such that $F_d(\tau) = F(\tau - d)$ for all τ and $A_{F_d}(0) \geq 1/2$, then F is left-log-concave.*

Proof. If F is left-log-concave, by Lemma 3 $A_F(\tau)$ is monotonically increasing in τ if $A_{F_\mu}(0) \geq \frac{1}{2}$ and monotonically decreasing otherwise. Conversely, if A_F is monotonically increasing for all translations F_d such that $F_d(\tau) = F(\tau - d)$ for all τ by Lemma 4, F is log-concave, completing the proof. □

D.3 SKEWED DISTRIBUTIONS

We now turn to how skew affects selective classification. Recall in Section 5 we defined skew-symmmetric distributions as follows:

Definition 2. *A distribution with density $f_{\alpha,\mu}$ is skew-symmetric with skew α and center μ if*

$$f_{\alpha,\mu}(\tau) = 2h(\tau - \mu)G(\alpha(\tau - \mu)) \tag{4}$$

for all $\tau \in \mathbb{R}$, where h and G are the density and CDF of distributions that are symmetric about 0.

When $\alpha = 0$, $f = h$ and there's no skew. Increasing α increases rightward skew, and decreasing α increases leftward skew. Note that in general, the mean of f depends on α as well. We will also consider the translated distribution:

$$f_{\alpha,\mu}(x) = f_\alpha(x - \mu). \tag{62}$$

The CDF of $f_{\alpha,\mu}$ is $F_{\alpha,\mu}$ and the CDF of f_α is F_α. We will use the following properties of these distributions:

Lemma 5 (Skew-symmetry about μ)**.** *These properties hold when we flip the skew from α to $-\alpha$:*

$$f_{\alpha,\mu}(x) = f_{-\alpha,c}(2\mu - x) \tag{63}$$
$$F_{\alpha,\mu}(x) = 1 - F_{-\alpha,c}(2\mu - x) \tag{64}$$

Proof.

$$\begin{aligned}
f_{\alpha,\mu}(x) &= 2h(x - \mu)G(\alpha(x - \mu)) && (65)\\
&= 2h(\mu - x)[G((-\alpha)(\mu - x))] && (66)\\
&= f_{-\alpha,c}(2\mu - x). && (67)\\
F_{\alpha,\mu}(x) &= \int_{-\infty}^{x} f_{\alpha,\mu}(t)dt && (68)\\
&= \int_{-\infty}^{x} f_{-\alpha,c}(2c - t)dt && (69)\\
&= \int_{2c-x}^{\infty} f_{-\alpha,c}(t)dt && (70)\\
&= 1 - F_{-\alpha,c}(2c - x). && (71)
\end{aligned}$$

□

Lemma 6 (Stochastic ordering with α)**.** *Let $\alpha_1 \leq \alpha_2$. Then $F_{\alpha_1,\mu} \geq F_{\alpha_2,\mu}$.*

Proof. Since the ordering is invariant to translation, without loss of generality we can take $\mu = 0$ First consider $x \leq 0$. Then

$$\begin{aligned}
F_{\alpha_1}(x) &= \int_{-\infty}^{x} 2h(t)G(\alpha_1 t)dt && (72)\\
&\geq \int_{-\infty}^{x} 2h(t)G(\alpha_2 t)dt && (73)\\
&= F_{\alpha_2}(x). && (74)
\end{aligned}$$

Now consider $x \geq 0$. We have

$$F_{\alpha_1}(x) = \int_{-\infty}^{x} 2h(t)G(\alpha_1 t)dt \tag{75}$$

$$= \int_{-\infty}^{x} 2h(t)(1 - G(-\alpha_1 t))dt \qquad \text{[by symmetry of G]} \tag{76}$$

$$= 2H(x) - \int_{-\infty}^{x} 2h(t)G(-\alpha_1 t)dt \tag{77}$$

$$= 2H(x) - \int_{-\infty}^{x} 2h(-t)G(-\alpha_1 t)dt \qquad \text{[by symmetry of h]} \tag{78}$$

$$= 2H(x) - \int_{-x}^{\infty} 2h(t)G(\alpha_1 t)dt \qquad \text{[change of variables]} \tag{79}$$

$$= 2H(x) - 1 + F_{\alpha_1}(-x), \tag{80}$$

which reduces to the case where $x \leq 0$. [This is Proposition 4 from Azzalini & Regoli (2012)] □

Lemma 7 (Log gradient ordering by skew)**.** *For all $\alpha \geq 0$ and $\tau \geq 0$,*

$$\frac{f_{\alpha,\mu}(\tau)}{F_{\alpha,\mu}(\tau)} \geq \frac{f_{0,\mu}(\tau)}{F_{0,\mu}(\tau)} \geq \frac{f_{-\alpha,c}(\tau)}{F_{-\alpha,c}(\tau)}. \tag{81}$$

Proof. Since the ordering is invariant to translation, without loss of generality we can take $c = 0$. We have:

$$\frac{f_\alpha(\tau)}{F_\alpha(\tau)} = \frac{h(\tau)G(\alpha\tau)}{\int_{-\infty}^{\tau} h(t)G(\alpha t)dt} \tag{82}$$

$$= \frac{h(\tau)}{\int_{-\infty}^{\tau} h(t)\frac{G(\alpha t)}{G(\alpha\tau)}dt} \tag{83}$$

$$\geq \frac{h(\tau)}{\int_{-\infty}^{\tau} h(t)dt} = \frac{f_0(\tau)}{f_0(\tau)} \tag{84}$$

$$\geq \frac{h(\tau)}{\int_{-\infty}^{\tau} h(t)\frac{G(-\alpha t)}{G(-\alpha\tau)}dt} \tag{85}$$

$$= \frac{h(\tau)G(-\alpha\tau)}{\int_{-\infty}^{\tau} h(t)G(-\alpha t)dt} \tag{86}$$

$$= \frac{f_{-\alpha}(\tau)}{F_{-\alpha}(\tau)}. \tag{87}$$

To get the inequalities, note that when $\alpha \geq 0$ and $t \leq \tau$, we have $\alpha \geq 0$, $G(\alpha t)/G(\alpha\tau) \leq 1$ since G is an increasing function. Similarly, $G(-\alpha t)/G(-\alpha\tau) \geq 1$. Since h is non-negative, the inequalities then follow from the monotonicity of the integral. □

We now prove the main results of this section.

D.3.1 Accuracy is monotone with skew

Proposition 2 (Accuracy is monotone with skew)**.** *Let $F_{\alpha,\mu}$ be the CDF of a skew-symmetric distribution. For all $\tau \geq 0$, $A_{F_{\alpha,\mu}}(\tau)$ is monotonically increasing in α.*

Proof. We use the skew-symmetry of $f_{\alpha,\mu}$ to write the selective accuracy as

$$A_{f_{\alpha,\mu}}(\tau) = \frac{1}{1 + \frac{F_{\alpha,\mu}(-\tau)}{1 - F_{\alpha,\mu}(\tau)}} \tag{88}$$

$$= \frac{1}{1 + \frac{F_{\alpha,\mu}(-\tau)}{F_{\alpha,\mu}(2\mu - \tau)}}. \tag{89}$$

We see that $A_{f_{\alpha,\mu}}(\tau)$ is a monotone decreasing function of $\frac{F_{\alpha,\mu}(-\tau)}{F_{-\alpha,c}(2\mu-\tau)}$. From Lemma 6, the numerator decreases with increasing α while the denominator increases. Thus, this fraction decreases, which in turn implies that $A_{f_{\alpha,\mu}}(\tau)$ is monotone increasing with α as desired. □

D.3.2 SKEW IN THE SAME DIRECTION PRESERVES MONOTONICITY

Proposition 3 (Skew in the same direction preserves monotonicity)**.** *Let $F_{\alpha,\mu}$ be the CDF of a skew-symmetric distribution. If $\mu \geq 0$ and $A_{F_{0,\mu}}(\tau)$ is monotonically increasing in τ, then $A_{F_{\alpha,\mu}}(\tau)$ is also monotonically increasing in τ for any $\alpha > 0$. Similarly, if $\mu \leq 0$ and $A_{F_{0,\mu}}(\tau)$ is monotonically decreasing in τ, then $A_{F_{\alpha,\mu}}(\tau)$ is also monotonically decreasing in τ for any $\alpha < 0$.*

Proof. The idea is to use Lemma 7 to reduce the statement to the case where $\alpha = 0$, so that we can apply monotonicity. First consider the case where $c \geq 0$. We thus have

$$\begin{aligned}
\frac{f_{\alpha,\mu}(-\tau)}{F_{\alpha,\mu}(-\tau)} &\geq \frac{f_{0,\mu}(-\tau)}{F_{0,\mu}(-\tau)} && \text{[Lemma 7]} && (90)\\
&= \frac{h_{\mu}(-\tau)}{H_{\mu}(-\tau)} && \text{[definition of } f\text{]} && (91)\\
&\geq \frac{h_{\mu}(\tau)}{1 - H_{\mu}(\tau)} && \text{[from monotonicity of } A\text{]} && (92)\\
&= \frac{h_{\mu}(2\mu - \tau)}{H_{\mu}(2\mu - \tau)} && \text{[symmetry of } h\text{]} && (93)\\
&= \frac{f_{0,\mu}(2\mu - \tau)}{F_{0,\mu}(2\mu - \tau)} && \text{[definition of } f\text{]} && (94)\\
&\geq \frac{f_{-\alpha,c}(2\mu - \tau)}{F_{-\alpha,c}(2\mu - \tau)} && \text{[Lemma 7]} && (95)\\
&\geq \frac{f_{\alpha,\mu}(\tau)}{1 - F_{\alpha,\mu}(\tau)}. && \text{[Lemma 5]} && (96)
\end{aligned}$$

Applying Lemma 2 completes the proof. The case where $\mu \leq 0$ is analogous. □

D.4 MONOTONE ODD TRANSFORMATIONS PRESERVE ACCURACY-COVERAGE CURVES.

We now show that our results are unchanged by strictly monotone and odd transformations to the PDF.

Lemma 8 (Odd and strictly monotonically increasing transformations preserve selective accuracy)**.** *Let X be a real-valued random variable with CDF F_X, T be a strictly monotonically increasing function, and define random variable $Y = T(X)$ with CDF F_Y. Then, $A_{F_X}(\tau) = A_{F_Y}(T(\tau))$ for each τ.*

Proof. For each τ, since T is a strictly monotonically increasing function, we apply the change of variables formula for transformations of univariate random variables to get the following:

$$\begin{aligned}
F_X(\tau) &= F_Y(T(\tau)) && && (97)\\
F_X(-\tau) &= F_Y(T(-\tau)) = F_Y(-T(\tau)). && [T \text{ is odd}] && (98)
\end{aligned}$$

We now solve for selective accuracy:

$$\begin{aligned}
A_{F_X}(\tau) &= \frac{1 - F_X(\tau)}{1 - F_X(\tau) + F_X(-\tau)} && (99)\\
&= \frac{1 - F_Y(T(\tau))}{1 - F_Y(T(\tau)) + F_Y(-T(\tau))} && (100)\\
&= A_{F_Y}(T(\tau)). && (101)
\end{aligned}$$

□

With Lemma 8, our results extend all random variables $T(X)$ where X corresponds to a left-log-concave distribution and T is odd and strictly monotonically increasing. In particular, X and $T(X)$ have the same accuracy-coverage curves.

E PROOFS: COMPARISON TO GROUP-AGNOSTIC BASELINE

In this section, we present the proofs from Section 6, which outline conditions under which selective classifiers outperform the group-agnostic baseline.

E.1 DEFINITIONS

We first write certain metrics on selective classifiers and their group-agnostic baseline in terms of margin distributions.

Definition 6. *Consider a margin distribution with CDF*

$$F = \sum_{g \in \mathcal{G}} p_g F_g, \tag{102}$$

where p_g is the mixture weight and F_g is the CDF for group g. We write the fraction of predictions in corectly and incorrectly classified examples on group g and on average as follows:

$$C_g(\tau) = 1 - F_g(\tau) \tag{103}$$

$$I_g(\tau) = F_g(-\tau) \tag{104}$$

$$C(\tau) = \sum_g p_g C_g(\tau) \tag{105}$$

$$I(\tau) = \sum_g p_g I_g(\tau) \tag{106}$$

We write the true-positive rate $R^{\mathsf{TP}}(\tau)$ and false-positive rate $R^{\mathsf{FP}}(\tau)$ for a given threshold τ as

$$R^{\mathsf{TP}}(\tau) = \frac{C(\tau)}{C(0)}, \tag{107}$$

$$R^{\mathsf{FP}}(\tau) = \frac{I(\tau)}{I(0)}. \tag{108}$$

Finally, we write the accuracy of the group-agnostic baseline on group g as

$$\tilde{A}_{F_g}(\tau) = \frac{A_{F_g}(0) R^{\mathsf{TP}}(\tau)}{A_{F_g}(0) R^{\mathsf{TP}}(\tau) + (1 - A_{F_g}(0)) R^{\mathsf{FP}}(\tau)} \tag{109}$$

$$= \frac{C_g(0) C(\tau)/C(0)}{C_g(0) C(\tau)/C(0) + I_g(0) I(\tau)/I(0)} \tag{110}$$

For convenience in our proofs below, we also define the fraction of each group g in correctly and incorrectly classified predictions at each threshold τ.

Definition 7. *We define the fraction of group g out of correctly and incorrectly classified predictions at threshold τ, $\mathbf{CF}_g(\tau)$ and $\mathbf{IF}_g(\tau)$ respectively, as follows:*

$$\mathbf{CF}_g(\tau) = \frac{p_g C_g(\tau)}{C(\tau)} \tag{111}$$

$$\mathbf{IF}_g(\tau) = \frac{p_g I_g(\tau)}{I(\tau)}. \tag{112}$$

E.2 GENERAL NECESSARY CONDITIONS TO OUTPERFORM THE BASELINE

We now present the proofs for Proposition 4 and Corollary 1, starting with the supporting lemmas.

We first write the accuracy of the group-agnostic baseline in terms of $\mathbf{IF}_{\mathsf{wg}}(0)$ and $\mathbf{CF}_{\mathsf{wg}}(0)$.

Lemma 9.

$$\tilde{A}_{F_{\mathsf{wg}}}(\tau) = \frac{1}{1 + \frac{\mathbf{IF}_{\mathsf{wg}}(0)}{\mathbf{CF}_{\mathsf{wg}}(0)} \cdot \frac{I(\tau)}{C(\tau)}}. \tag{113}$$

Proof. We have:

$$\begin{aligned}
\tilde{A}_{F_{\mathsf{wg}}}(\tau) &= \frac{A_{F_{\mathsf{wg}}}(0) R^{\mathsf{TP}}(\tau)}{A_{F_{\mathsf{wg}}}(0) R^{\mathsf{TP}}(\tau) + (1 - A_{F_{\mathsf{wg}}}(0)) R^{\mathsf{FP}}(\tau)} && (114)\\
&= \frac{p C_{\mathsf{wg}}(0) R^{\mathsf{TP}}(\tau)}{p C_{\mathsf{wg}}(0) R^{\mathsf{TP}}(\tau) + p I_{\mathsf{wg}}(0) R^{\mathsf{FP}}(\tau)} && (115)\\
&= \frac{p C_{\mathsf{wg}}(0) \frac{C(\tau)}{C(0)}}{p C_{\mathsf{wg}}(0) \frac{C(\tau)}{C(0)} + p I_{\mathsf{wg}}(0) \frac{I(\tau)}{I(0)}} && (116)\\
&= \frac{\mathbf{CF}_{\mathsf{wg}}(0) C(\tau)}{\mathbf{CF}_{\mathsf{wg}}(0) C(\tau) + \mathbf{IF}_{\mathsf{wg}}(0) I(\tau)} && (117)\\
&= \frac{1}{1 + \frac{\mathbf{IF}_{\mathsf{wg}}(0)}{\mathbf{CF}_{\mathsf{wg}}(0)} \cdot \frac{I(\tau)}{C(\tau)}}. && (118)
\end{aligned}$$

□

Lemma 10 (Bounding the derivative of $1/\mathbf{CF}_{\mathsf{wg}}(\tau)$ at $\tau = 0$)**.** *If* $A_{F_{\mathsf{others}}}(0) \geq A_{F_{\mathsf{wg}}}(0)$,

$$\frac{d}{d\tau}\left(\frac{1}{\mathbf{CF}_{\mathsf{wg}}(\tau)}\right)\bigg|_{\tau=0} \geq \left(\frac{1-p}{p}\right)\left(\frac{F_{\mathsf{wg}}(0)}{1 - F_{\mathsf{wg}}(0)}\right)\left(\frac{f_{\mathsf{wg}}(0) F_{\mathsf{others}}(0) - f_{\mathsf{others}}(0) F_{\mathsf{wg}}(0)}{F_{\mathsf{wg}}(0)^2}\right) \tag{119}$$

Proof.

$$\begin{aligned}
&\frac{d}{d\tau}\left(\frac{1}{\mathbf{CF}_{\mathsf{wg}}(\tau)}\right)\bigg|_{\tau=0} && (120)\\
=&\frac{d}{d\tau}\left(\frac{p C_{\mathsf{wg}}(\tau) + (1-p) C_{\mathsf{others}}(\tau)}{p C_{\mathsf{wg}}(\tau)}\right)\bigg|_{\tau=0} && (121)\\
=&\frac{d}{d\tau}\left(1 + \left(\frac{1-p}{p}\right)\left(\frac{1 - F_{\mathsf{others}}(\tau)}{1 - F_{\mathsf{wg}}(\tau)}\right)\right)\bigg|_{\tau=0} && (122)\\
=&\frac{1-p}{p}\frac{d}{d\tau}\left(\frac{1 - F_{\mathsf{others}}(\tau)}{1 - F_{\mathsf{wg}}(\tau)}\right)\bigg|_{\tau=0} && (123)\\
=&\frac{1-p}{p}\left(\frac{f_{\mathsf{wg}}(\tau)(1 - F_{\mathsf{others}}(\tau)) - f_{\mathsf{others}}(\tau)(1 - F_{\mathsf{wg}}(\tau))}{(1 - F_{\mathsf{wg}}(\tau))^2}\right)\bigg|_{\tau=0} && (124)\\
=&\frac{1-p}{p}\frac{1}{1 - F_{\mathsf{wg}}(\tau)}\left(f_{\mathsf{wg}}(\tau)\left(\frac{1 - F_{\mathsf{others}}(\tau)}{1 - F_{\mathsf{wg}}(\tau)}\right) - f_{\mathsf{others}}(\tau)\right)\bigg|_{\tau=0} && (125)\\
=&\frac{1-p}{p}\frac{1}{1 - F_{\mathsf{wg}}(0)}\left(f_{\mathsf{wg}}(0)\left(\frac{1 - F_{\mathsf{others}}(0)}{1 - F_{\mathsf{wg}}(0)}\right) - f_{\mathsf{others}}(0)\right) && (126)\\
\geq&\frac{1-p}{p}\frac{1}{1 - F_{\mathsf{wg}}(0)}\left(f_{\mathsf{wg}}(0)\frac{F_{\mathsf{others}}(0)}{F_{\mathsf{wg}}(0)} - f_{\mathsf{others}}(0)\right) && [0 < F_{\mathsf{others}}(0) \leq F_{\mathsf{wg}}(0) < 1] \quad (127)\\
=&\frac{1-p}{p}\frac{1}{1 - F_{\mathsf{wg}}(0)}\frac{f_{\mathsf{wg}}(0) F_{\mathsf{others}}(0) - f_{\mathsf{others}}(0) F_{\mathsf{wg}}(0)}{F_{\mathsf{wg}}(0)} && (128)\\
=&\frac{1-p}{p}\frac{F_{\mathsf{wg}}(0)}{1 - F_{\mathsf{wg}}(0)}\frac{f_{\mathsf{wg}}(0) F_{\mathsf{others}}(0) - f_{\mathsf{others}}(0) F_{\mathsf{wg}}(0)}{F_{\mathsf{wg}}(0)^2} && (129)
\end{aligned}$$

□

Lemma 11. *If* $A_{F_{\text{others}}}(0) \geq A_{F_{\text{wg}}}(0) > 0.5$*, then*

$$\frac{d}{d\tau}\frac{\mathbf{IF}_{\text{wg}}(\tau)}{\mathbf{CF}_{\text{wg}}(\tau)}\bigg|_{\tau=0} \geq C(f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0)) \tag{130}$$

for some positive constant C*.*

Proof.

$$\frac{d}{d\tau}\frac{\mathbf{IF}_{\text{wg}}(\tau)}{\mathbf{CF}_{\text{wg}}(\tau)}\bigg|_{\tau=0} \tag{131}$$

$$= \left(\frac{d}{d\tau}\mathbf{IF}_{\text{wg}}(\tau)\bigg|_{\tau=0}\right)\frac{1}{\mathbf{CF}_{\text{wg}}(0)} + \mathbf{IF}_{\text{wg}}(0)\left(\frac{d}{d\tau}\frac{1}{\mathbf{CF}_{\text{wg}}(\tau)}\bigg|_{\tau=0}\right) \tag{132}$$

$$\geq \left(\frac{d}{d\tau}\mathbf{IF}_{\text{wg}}(\tau)\bigg|_{\tau=0}\right)\frac{1}{\mathbf{CF}_{\text{wg}}(0)} \tag{133}$$

$$+ \mathbf{IF}_{\text{wg}}(0)\left(\frac{1-p}{p}\right)\left(\frac{F_{\text{wg}}(0)}{1-F_{\text{wg}}(0)}\right)\left(\frac{f_{\text{wg}}(0)F_{\text{others}}(0) - f_{\text{others}}(0)F_{\text{wg}}(0)}{F_{\text{wg}}(0)^2}\right) \tag{134}$$

[Lemma 10

$$= \left(\frac{1}{1+\frac{1-p}{p}\frac{F_{\text{others}}(0)}{F_{\text{wg}}(0)}}\right)^2\left(\frac{1-p}{p}\right)\left(\frac{f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0)}{F_{\text{wg}}(0)^2}\right)\left(1+\left(\frac{1-p}{p}\right)\left(\frac{1-F_{\text{others}}(0)}{1-F_{\text{wg}}(0)}\right)\right) \tag{135}$$

$$+ \left(\frac{1}{1+\frac{1-p}{p}\frac{F_{\text{others}}(0)}{F_{\text{wg}}(0)}}\right)\left(\frac{1-p}{p}\right)\left(\frac{F_{\text{wg}}(0)}{1-F_{\text{wg}}(0)}\right)\left(\frac{f_{\text{wg}}(0)F_{\text{others}}(0) - f_{\text{others}}(0)F_{\text{wg}}(0)}{F_{\text{wg}}(0)^2}\right)$$

$$= \underbrace{\left(\frac{1}{1+\frac{1-p}{p}\frac{F_{\text{others}}(0)}{F_{\text{wg}}(0)}}\right)}_{>0}\underbrace{\left(\frac{1-p}{p}\right)}_{>0}(f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0))\underbrace{\left(\frac{1}{F_{\text{wg}}(0)^2}\right)}_{>0} \tag{136}$$

$$* \left(\underbrace{\frac{1+\frac{1-p}{p}\frac{1-F_{\text{others}}(0)}{1-F_{\text{wg}}(0)}}{1+\frac{1-p}{p}\frac{F_{\text{others}}(0)}{F_{\text{wg}}(0)}}}_{\geq 1 \text{ because } A_{F_{\text{others}}}(0) \geq A_{F_{\text{wg}}}(0)} - \underbrace{\frac{F_{\text{wg}}(0)}{1-F_{\text{wg}}(0)}}_{<1 \text{ because } A_{F_{\text{wg}}}(0) > 0.5}\right)$$

$$= C(f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0)) \tag{137}$$

□

E.2.1 Necessary condition for outperforming the baseline

Proposition 4 (Necessary condition for outperforming the baseline)**.** *Assume that* $1/2 < A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0) < 1$ *and* $f_{\text{wg}}(0) > 0$*. If* $\tilde{A}_{F_{\text{wg}}}(\tau) \leq A_{F_{\text{wg}}}(\tau)$ *for all* $\tau \geq 0$*, then*

$$\frac{f_{\text{others}}(0)}{f_{\text{wg}}(0)} \leq \frac{1-A_{F_{\text{others}}}(0)}{1-A_{F_{\text{wg}}}(0)}. \tag{5}$$

Proof. Recall that

$$A_{F_{\text{wg}}}(\tau) = \frac{1}{1+\frac{\mathbf{IF}_{\text{wg}}(\tau)}{\mathbf{CF}_{\text{wg}}(\tau)}\cdot\frac{I(\tau)}{C(\tau)}} \tag{138}$$

$$\tilde{A}_{F_{\text{wg}}}(\tau) = \frac{1}{1+\frac{\mathbf{IF}_{\text{wg}}(0)}{\mathbf{CF}_{\text{wg}}(0)}\cdot\frac{I(\tau)}{C(\tau)}}. \tag{139}$$

If $A_{F_{\text{wg}}}(\tau) \geq \tilde{A}_{F_{\text{wg}}}(\tau)$ for all $\tau \geq 0$, then

$$\frac{d}{d\tau} \frac{\mathbf{IF}_{\text{wg}}(\tau)}{\mathbf{CF}_{\text{wg}}(\tau)}\bigg|_{\tau=0} \leq 0. \tag{140}$$

From Lemma 11,

$$C(f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0)) \leq \frac{d}{d\tau} \frac{\mathbf{IF}_{\text{wg}}(\tau)}{\mathbf{CF}_{\text{wg}}(\tau)}\bigg|_{\tau=0} \tag{141}$$

for some positive constant C. Combined,

$$f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0) \leq 0. \tag{142}$$

□

E.2.2 Outperforming the baseline requires smaller scaling for log-concave distributions

Corollary 1 (Outperforming the baseline requires smaller scaling for log-concave distributions)**.** *Assume that* $0.5 < A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0) < 1$*,* F_{wg} *is log-concave, and* $f_{\text{others}}(\tau) = v f_{\text{wg}}(v(\tau - \mu_{\text{others}}) + \mu_{\text{wg}})$ *for all* $\tau \in \mathbb{R}$*, where* v *is a scaling factor. If* $v > 1$*, then* $\tilde{A}_{F_{\text{wg}}}(\tau) > A_{F_{\text{wg}}}(\tau)$ *for some* $\tau \geq 0$*.*

Proof. By Proposition 4, if $A_{F_{\text{wg}}}(\tau) \geq \tilde{A}_{F_{\text{wg}}}(\tau)$ for all $\tau \geq 0$, then:

$$f_{\text{others}}(0)F_{\text{wg}}(0) - f_{\text{wg}}(0)F_{\text{others}}(0) \leq 0, \tag{143}$$

and equivalently,

$$\frac{f_{\text{others}}(0)}{F_{\text{others}}(0)} \leq \frac{f_{\text{wg}}(0)}{F_{\text{wg}}(0)}. \tag{144}$$

From the definition of f_{others},

$$v \leq \frac{f_{\text{wg}}(0)/F_{\text{wg}}(0)}{f_{\text{wg}}(-\mu_{\text{bg}}v + \mu_{\text{wg}})/F_{\text{wg}}(-\mu_{\text{bg}}v + \mu_{\text{wg}})} \tag{145}$$

Because $A_{F_{\text{others}}}(0) > A_{F_{\text{wg}}}(0)$, $-\mu_{\text{bg}}v + \mu_{\text{wg}} < 0$. Applying log-concavity yields

$$v < 1. \tag{146}$$

Thus, when $v > 1$, there exists some threshold $\tau \geq 0$ where $\tilde{A}_{F_{\text{wg}}}(\tau) > A_{F_{\text{wg}}}(\tau)$. □

E.3 Translated, log-concave distributions always underperform the baseline

We now present a proof of Proposition 5 Assume f_{wg} and f_{others} are log concave and symmetric, group 1 has mean μ_{wg}, group 2 has mean μ_{bg}, and $f_{\text{others}}(x) = f_{\text{wg}}(x - (\mu_{\text{bg}} - \mu_{\text{wg}}))$, (the densities are translations of each other.) For convenience, define $d = \mu_{\text{bg}} - \mu_{\text{wg}}$. This implies:

$$\begin{aligned} F_{\text{others}}(x) &= F_{\text{wg}}(x - (\mu_{\text{bg}} - \mu_{\text{wg}})) && (147)\\ &= F_{\text{wg}}(x - d). && (148) \end{aligned}$$

We will first show that wg is indeed the worst group at full coverage; i.e. $wgacc(0) < A_{F_{\text{others}}}(0)$:

Lemma 12. *If* f_{wg} *is a PDF and* f_{others} *is obtained by translating* f_{wg} *to the right (i.e.* $f_{\text{others}}(x) = f_{\text{wg}}(x - d)$ *for* $d > 0$*), and* F_{wg} *and* F_{others} *are their associated CDFs,* $A_{F_{\text{wg}}}(0) < A_{F_{\text{others}}}(0)$*.*

Proof. We have:

$$\begin{aligned} A_{F_{\text{wg}}}(0) &= 1 - F_{\text{wg}}(0) && (149)\\ &\leq 1 - F_{\text{wg}}(-d) && (150)\\ &= 1 - F_{\text{others}}(0) && (151)\\ &= A_{F_{\text{others}}}(0). && (152) \end{aligned}$$

□

We next show that $\mathbf{CF}_{\mathsf{wg}}(\tau)$ is monotonically decreasing in τ:

Lemma 13. *If f_{wg}, F_{wg}, f_{others}, F_{others}, p, and $\mathbf{CF}_{\mathsf{wg}}$ are as described, $\mathbf{CF}_{\mathsf{wg}}$ is monotonically decreasing in τ.*

Proof. First, defining $\mathbf{CF}_{\mathsf{wg}}$ in terms of F_{wg}, we have:

$$\mathbf{CF}_{\mathsf{wg}}(\tau) = \frac{p(1-F_{\mathsf{wg}}(\tau))}{p(1-F_{\mathsf{wg}}(\tau)) + (1-p)(1-F_{\mathsf{others}}(\tau))} \tag{153}$$

$$= \frac{p(1-F_{\mathsf{wg}}(\tau))}{p(1-F_{\mathsf{wg}}(\tau)) + (1-p)(1-F_{\mathsf{wg}}(\tau-d))} \tag{154}$$

$$= \frac{pF_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}{pF_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau) + (1-p)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)} \tag{155}$$

$$= \frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)}{pF_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}}. \tag{156}$$

Taking the derivative of $\mathbf{CF}_{\mathsf{wg}}$, we have:

$$\frac{d}{d\tau}\mathbf{CF}_{\mathsf{wg}}(\tau) = \frac{d}{d\tau}\left(\frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)}{pF_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}}\right) \tag{157}$$

$$= -\left(\frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)}{pF_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}}\right)^2 \left(\frac{1-p}{p}\right)\frac{d}{d\tau}\left(\frac{F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)}{F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}\right) \tag{158}$$

$$= -c\left(-f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau) + f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)\right), \tag{159}$$

where c is a positive constant (since p and $1-p$ are positive and all of the terms are squares.) Thus, we can say that $\mathbf{CF}_{\mathsf{wg}}(\tau)$ is monotonically decreasing for all τ if:

$$f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau) \leq f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d) \tag{160}$$

$$\frac{f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)}{F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau+d)} \leq \frac{f_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}{F_{\mathsf{wg}}(2\mu_{\mathsf{wg}}-\tau)}. \tag{161}$$

Now, since $d > 0$, this follows from the log concavity of f_{wg}. Therefore, the fraction of correct examples that come from group 1 is a decreasing function of τ. □

Next, we show that the $\mathbf{IF}_{\mathsf{wg}}(\tau)$ is monotonically increasing in τ:

Lemma 14. *If f_{wg}, F_{wg}, f_{others}, F_{others}, p and $\mathbf{IF}_{\mathsf{wg}}$ are as described, $\mathbf{IF}_{\mathsf{wg}}$ is monotonically increasing in τ.*

Proof. First, we note that:

$$\mathbf{IF}_{\mathsf{wg}}(\tau) = \frac{pF_{\mathsf{wg}}(-\tau)}{pF_{\mathsf{wg}}(-\tau) + (1-p)F_{\mathsf{wg}}(-\tau-d)} \tag{162}$$

$$= \frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(-\tau-d)}{pF_{\mathsf{wg}}(-\tau)}}. \tag{163}$$

We will show that the derivative of $\mathbf{IF}_{\mathsf{wg}}(\tau)$ with respect to τ is positive. Doing so gives us:

$$\frac{d}{d\tau}\mathbf{IF}_{\mathsf{wg}}(\tau) = \frac{d}{d\tau}\frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(-\tau-d)}{pF_{\mathsf{wg}}(-\tau)}} \tag{164}$$

$$= -\left(\frac{1}{1+\frac{(1-p)F_{\mathsf{wg}}(-\tau-d)}{pF_{\mathsf{wg}}(-\tau)}}\right)^2 \left(\frac{1-p}{p}\right)\frac{d}{d\tau}\left(\frac{F_{\mathsf{wg}}(-\tau-d)}{F_{\mathsf{wg}}(-\tau)}\right) \tag{165}$$

$$= -c\left(-f_{\mathsf{wg}}(-\tau-d)F_{\mathsf{wg}}(-\tau) + f_{\mathsf{wg}}(-\tau)F_{\mathsf{wg}}(-\tau-d)\right), \tag{166}$$

where c is positive since it is the product of squared terms and $(1-p)/p$, which is also positive. Thus, $\frac{d}{d\tau}\mathbf{IF}_{\mathsf{wg}}(\tau) \geq 0$ is equivalent to:

$$-\left(-f_{\mathsf{wg}}(-\tau-d)F_{\mathsf{wg}}(-\tau) + f_{\mathsf{wg}}(-\tau)F_{\mathsf{wg}}(-\tau-d)\right) \geq 0 \tag{167}$$

$$f_{\mathsf{wg}}(-\tau-d)F_{\mathsf{wg}}(-\tau) \geq f_{\mathsf{wg}}(-\tau)F_{\mathsf{wg}}(-\tau-d) \tag{168}$$

$$\frac{f_{\mathsf{wg}}(-\tau-d)}{F_{\mathsf{wg}}(-\tau-d)} \geq \frac{f_{\mathsf{wg}}(-\tau)}{F_{\mathsf{wg}}(-\tau)}. \tag{169}$$

Since f_{wg} is log-concave and d is positive, the last inequality is true, so $\mathbf{IF}_{\mathsf{wg}}(\tau)$ is monotonically increasing in τ, as desired. □

Lastly, we show that Lemma 13 and Lemma 14 imply that the ratio $\frac{\mathbf{IF}_{\mathsf{wg}}(\tau)}{\mathbf{CF}_{\mathsf{wg}}(\tau)}$ is monotonically increasing in τ.

Lemma 15. $\frac{\mathbf{IF}_{\mathsf{wg}}(\tau)}{\mathbf{CF}_{\mathsf{wg}}(\tau)}$ *is monotonically increasing in* τ

Proof. We simply follow the quotient rule:

$$\frac{d}{d\tau}\frac{\mathbf{IF}_{\mathsf{wg}}(\tau)}{\mathbf{CF}_{\mathsf{wg}}(\tau)} = \mathbf{CF}_{\mathsf{wg}}(\tau)IF_1'(\tau) - \mathbf{IF}_{\mathsf{wg}}(\tau)CF_1'(\tau) \tag{170}$$

$$\geq 0, \tag{171}$$

where we note that $\mathbf{CF}_{\mathsf{wg}}(\tau), \mathbf{IF}_{\mathsf{wg}}(\tau) \geq 0$, $CF_1'(\tau) < 0$ from Lemma 13, and $IF_1'(\tau) > 0$ from Lemma 14. □

We will now prove the main result of this section:

E.3.1 Translated, log-concave distributions always underperform the baseline

Proposition 5 (Translated, log-concave distributions always underperform the baseline)**.** *Assume* F_{wg} *and* F_{others} *are log-concave and* $f_{\mathsf{others}}(\tau) = f_{\mathsf{wg}}(\tau - d)$ *for all* $\tau \in \mathbb{R}$*. Then for all* $\tau \geq 0$*,*

$$A_{F_{\mathsf{wg}}}(\tau) \leq \tilde{A}_{F_{\mathsf{wg}}}(\tau). \tag{6}$$

Proof. We'll consider the case of underperforming the baseline: the case of overperforming the baseline is analogous. Assume group 1 is the lower accuracy group (so d is positive.) Using the definition of selective accuracy, we have:

$$A_{F_{\mathsf{wg}}}(\tau) = \frac{C_{\mathsf{wg}}(\tau)}{C_{\mathsf{wg}}(\tau) + I_1(\tau)} \tag{172}$$

$$= \frac{\mathbf{CF}_{\mathsf{wg}}(\tau)C(\tau)}{\mathbf{CF}_{\mathsf{wg}}(\tau)C(\tau) + \mathbf{IF}_{\mathsf{wg}}(\tau)I(\tau)} \tag{173}$$

$$= \frac{1}{1 + \frac{\mathbf{IF}_{\mathsf{wg}}(\tau)}{\mathbf{CF}_{\mathsf{wg}}(\tau)} * \frac{I(\tau)}{C(\tau)}} \tag{174}$$

$$\leq \frac{1}{1 + \frac{\mathbf{IF}_{\mathsf{wg}}(0)}{\mathbf{CF}_{\mathsf{wg}}(0)} * \frac{I(\tau)}{C(\tau)}} \qquad \text{[by Lemma 15]} \tag{175}$$

$$= \frac{\mathbf{CF}_{\mathsf{wg}}(0)C(\tau)}{\mathbf{CF}_{\mathsf{wg}}(0)C(\tau) + \mathbf{IF}_{\mathsf{wg}}(0)I(\tau)} \tag{176}$$

$$= \tilde{A}_{F_{\mathsf{wg}}}(\tau). \tag{177}$$

Thus, group 1 (the worst group) underperforms the baseline, as desired. □

F SIMULATIONS

To demonstrate that it is possible but challenging to outperform the group-agnostic baseline, we present simulation results on margin distributions that are mixtures of two Gaussians. Following the setup from Section 6, we consider a best-group margin distribution $\mathcal{N}(1, 1)$ and a worst-group margin distribution $\mathcal{N}(\mu, \sigma^2)$ varying the parameters μ, σ. We set the fraction of mass from the worst group, p to be 0.5. We evaluate whether selective classification outperforms the group-agnostic baseline, plotting the results in Figure 9. We observe that selective classification outperforms the group-agnostic baseline under some parameter settings, notably when the necessary condition in Proposition 4 is met and when the variance is smaller as implied by Corollary 1, but it underperforms the baseline under most parameter settings.

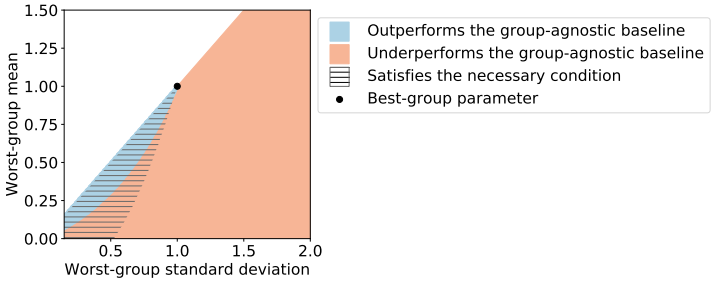

Figure 9: Simulation results on margin distributions that are mixtures of two Gaussians, where we vary the mean and the variance of the worst-group margin distributions. For parameters corresponding to the blue and orange regions, the worst group outperforms and underperforms the group-agnostic baseline, respectively. We do not consider parameters in the white region to maintain that they yield full-coverage accuracies that are worse than the other group. We shade parameters that satisfy the necessary condition in Proposition 4.

In addition to whether the worst group underperforms or outperforms the baseline, we study the magnitude of the difference in the worst-group accuracy with respect to the group-agnostic baseline through the same simulations (Figure 10). Concretely, we compute the difference in worst-group accuracy with respect to the group-agnostic baseline, taking the worst-case difference across thresholds: $\max_\tau \tilde{A}_{F_{\text{wg}}}(\tau) - A_{F_{\text{wg}}}(\tau)$. We observe significant disparities between the observed worst-group accuracy and the group-agnostic baseline.

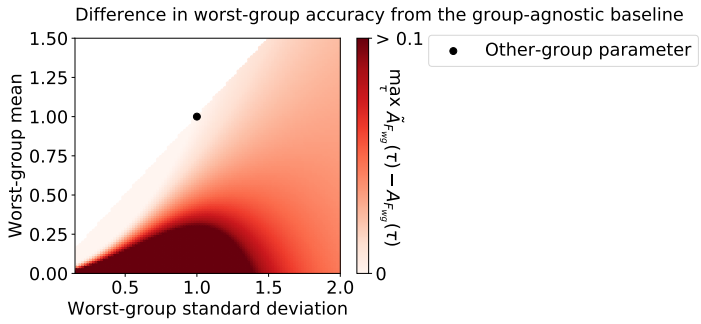

Figure 10: Simulation results on margin distributions that are mixtures of two Gaussians, where we vary the mean and the variance of the worst-group margin distribution, and keep the other group's margin distribution fixed with mean and variance 1. We compute the difference in worst-group accuracy with respect to the group-agnostic baseline, taking the worst-case difference across thresholds: $\max_\tau \tilde{A}_{F_{\text{wg}}}(\tau) - A_{F_{\text{wg}}}(\tau)$.

Lastly, we simulate the effects of varying the worst-group and other-group means while keeping the variance fixed at $\sigma^2 = 1$ for both groups, following the same simulation protocol otherwise.

We similarly observe substantial gaps between the observed worst-group accuracy and the group-agnostic baseline (Figure 11).

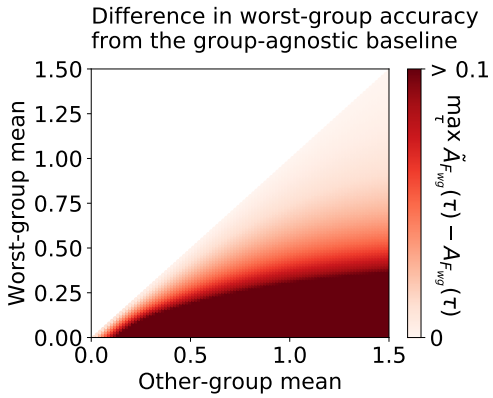

Figure 11: Simulation results on margin distributions that are mixtures of two Gaussians, where we vary the means of the two margin distributions, keeping their variances fixed at 1. We compute the difference in worst-group accuracy with respect to the group-agnostic baseline, taking the worst-case difference across thresholds: $\max_\tau \tilde{A}_{F_{\text{wg}}}(\tau) - A_{F_{\text{wg}}}(\tau)$.