Disagreement-based Active Learning for Robustness Against Subpopulation Shifts

Yeat Jeng Ng[1](✉), Viktoriia Sharmanska[1], Thomas Kehrenberg[3], Anastasia Pentina, and Novi Quadrianto[1,2,3]

[1] Predictive Analytics Lab, University of Sussex
{y.ng, n.quadrianto}@sussex.ac.uk
[2] Monash University, Indonesia
[3] Basque Center for Applied Mathematics (BCAM)

Abstract. Machine learning models excel at specific tasks and are increasingly being used in critical decision-making processes. However, the data used to train these models can be biased due to spurious correlations. Spurious correlations are associations between input features and target labels that exist in the training data but do not hold in the test distribution. To address this issue, we propose a learning framework called Active Learning via Source-Target Disagreement (AL-STD), which actively explores the space of data points to mitigate spurious correlations caused by subpopulation shifts. Subpopulations can represent different demographic identities, such as race and gender, or other background attributes. Our proposed active learning (AL) method minimizes the region of disagreement between two learning hypotheses: the standard empirical risk hypothesis and a second hypothesis that uses instance reweighting to adjust for the mismatch between training and test distributions. We theoretically motivate the idea of shrinking the region of disagreement to address subpopulation shifts in the AL context. We conduct extensive experiments on four datasets, including image, tabular, and text data, demonstrating that our AL approach is more robust than comparable baselines under various subpopulation shifts.

Keywords: Active Learning · Subgroup Robustness

1 Introduction

Active learning (AL) involves selecting samples from a large unlabelled pool such that, once labelled, these samples are maximally informative for training a classifier [37,14]. This task is challenging because, without labels, it is difficult to determine how informative a sample will be. In pool-based AL [26], a small initial labelled pool is assumed to be available for training a preliminary classifier, which can then be used to evaluate the unlabelled samples.

However, this evaluation becomes unreliable if the initial labelled pool is not representative of the test set or the deployment setting [40]. Specifically, this paper investigates scenarios where there is a spurious correlation in the initial

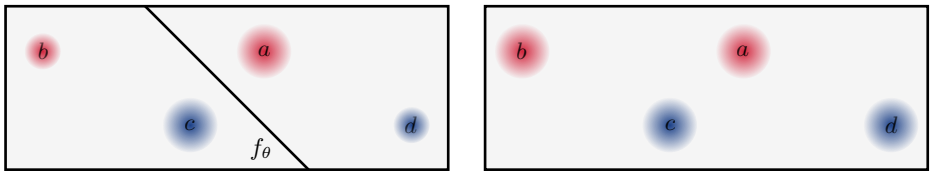

Fig. 1: Illustration of subpopulation shift in the training (source) data distribution Q (left) and test (target) distribution P (right). Features are generated from four different two-dimensional Gaussians, each representing a subgroup $(y, s) \in Y \times S$. Samples from subpopulations a and b share the target label $y = 1$ (in red) but differ in group label s; similarly, samples from subpopulations c and d have target label $y = 0$ (in blue) and differ in s. The variances indicate the proportions of subgroups in the dataset, with b and d as the minority subgroups in the training data. Coupled with a maximum-margin loss function like hinge loss, the empirical risk minimization (ERM) objective creates a decision boundary that misclassifies samples from minority subgroups, which becomes critical in the test (target) distribution.

labelled pool (but not in the test set), leading to confounded classifiers. One scenario involves the unlabelled pool itself containing these spurious correlations, which an initial labelled pool inherits if sampled uniformly from the unlabelled pool. Another scenario is that the initial pool resulted from a biased sampling process [17,24], and the high cost of obtaining labels (e.g., in medical applications [10,19]) makes it prohibitively expensive to discard the existing labelled pool and start anew. For instance, in a person-related dataset with bureaucratic and legal hurdles to obtaining more labels, an initial labelled pool might almost exclusively cover a single demographic group, providing a very uneven starting point for selecting samples to label.

In this example, the specific bias or spurious correlation studied in this paper is mediated by *subgroups*, such as demographic groups, present in the data. We refer to this bias as subpopulation shift [34,39,21], characterized by shifts in subpopulations resulting in some classes (i.e., prediction targets) being more likely than others. The shift varies between subpopulations, creating spurious correlations between the subpopulations and the prediction targets. In the presence of such spurious correlations, it is crucial to select the right samples for AL, as a naïve selection algorithm could exacerbate the problem. This paper develops an AL algorithm for selecting useful samples from an initial labelled pool experiencing subpopulation shift.

2 Related Work

Active learning. There are two prominent querying strategies in AL [37]: *uncertainty-based* and *diversity-based* approaches. In uncertainty-based strategies, a model is first trained on the available labelled data, and then samples from the unlabelled

pool are selected where the model is most uncertain. One classic method is from[33], which selects samples where the difference in predicted probabilities of the top two classes is the largest. Diversity-based approaches, on the other hand, rely on solving a coreset selection problem. Coresets are subsamples of a dataset used as proxies for the full set. For example,[36] and[13] constructed coresets by solving a k-center problem. Hybrid strategies combining uncertainty and diversity have also been proposed; for instance, [4] selected samples with gradients spanning diverse directions, where gradient magnitude indicates uncertainty. Closely related to our work is the *disagreement-based* strategy, which queries the label of a sample if it falls within a region of disagreement [14]. This method maintains a set of possible risk minimizers and queries the label of a sample x if two hypotheses h_1 and h_2 in the set yield different predictions for x. However, disagreement-based strategies typically have high label requirements. In this paper, we employ a disagreement-based strategy to handle subpopulation shifts and empirically demonstrate its competitive label complexity compared to uncertainty, diversity, and hybrid strategies.

Subpopulation shift. Subpopulation shift is a specific instance of the broader problem of train-test distribution shifts, where the training data distribution differs significantly from the test data distribution. Such shifts can notably degrade the performance of machine learning models when deployed in real-world scenarios [21]. Subpopulation shift specifically refers to changes at the level of subpopulations within the train and test data distributions, such as those based on demographics, with some subpopulations being underrepresented in the training set. The primary objective is to enhance accuracy for the *worst-off* subpopulation, often addressed through distributionally-robust optimization methods [34,39]. Unlike these studies, which explore a static setting and assume access to target label information, our work focuses on an active learning context.

Algorithmic fairness. Subpopulation shift is closely related to algorithmic fairness, particularly resonating with the concept of *Rawlsian Max-Min fairness* [32], which advocates for decisions that maximize the minimum outcome, thereby improving the worst-off situation. AL in the context of fairness has also been explored. For instance, [2] proposed a sampling strategy that queries labels to reduce uncertainty while minimally violating fairness measures like demographic parity. This approach, however, is computationally intensive as it requires computing the expected fairness measure over all possible target labels for each sample in each round, necessitating access to both target and group labels. To address this, [2] assumed that the unlabelled pool includes group information. In contrast, [38] developed a meta-learning version of the fair AL setting, avoiding reliance on manual selection strategies. Our approach, similar to the standard AL setup, assumes a completely unlabelled pool where samples lack target and group labels. Related work by [30] actively collected additional features for data points to equalize performance across different groups. Additionally, [1] proposed an active sampling strategy that selects labelled samples from the group worst off under the current model to update the model. Both strategies are tangential to the pool-

based AL setup described below, aiming to equalize performance and improve fairness through active learning.

3 Preliminaries

Active learning problem. Let $\mathcal{U} = \{x_i\}_{i\in[m]}$ denote the set of initial unlabelled samples and $\mathcal{L}_0 = \{(x_i, y_i)\}_{i\in[n]}$ denote a set of labelled samples, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$ are the input features and target of the i-th sample respectively. In this paper, we focus exclusively on classification problems, thus $\mathcal{Y} = \{0, \ldots, C-1\}$.

In pool-based AL, starting with an initial labelled pool $\mathcal{L}_0$ the agent sequentially queries an oracle for annotations of some unlabelled samples from $\mathcal{U}$. Formally, at the t-th iteration the active learner obtains a model $f_{\theta(\mathcal{L}_t)}$ using the labelled set $\mathcal{L}_t$, and subsequently queries target labels for k new samples $\mathcal{H}_t$, selected by the acquisition function $q(f_{\theta(\mathcal{L}_t)}, \mathcal{U}) = \mathcal{H}_t \subseteq \mathcal{U}$. This results in a new labelled pool $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \mathcal{H}_t$ for the next iteration. In this work, we focus on the case in which models $f_{\theta(\mathcal{L}_t)}$ are obtained using empirical risk minimization (ERM):

$$\hat{\theta} = \arg\min_{\theta\in\Theta} \hat{R}(f_\theta)\ ,\ \hat{R}(f_\theta) = \frac{1}{n_t}\sum_{i=1}^{n_t} L\left(f_\theta(x_i), y_i\right), \tag{1}$$

where $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ denotes the loss function, and n_t is the number of labelled data points in $\mathcal{L}_t$. For convenience, the notation $\mathcal{L}_t$ is omitted from $f_{\theta(\mathcal{L}_t)}$ when the context is clear. The process of querying annotations continues until a stopping criterion is reached, e.g. when the labelling budget is exhausted. The goal of the active learner is to achieve optimal metrics (e.g. maximum accuracy on the test set) with minimal label acquisition.

Subpopulation shift setup. In contrast to the standard AL setting, this work addresses the scenario where the training data ($\mathcal{U}$ and $\mathcal{L}_0$) is sampled from a distribution Q that differs from the target (or test) distribution P due to a subpopulation shift. Specifically, we assume that, in addition to the target label, the input features x are also associated with a group label $s \in \mathcal{S} = \{0, \ldots, B-1\}$, representing subpopulations. These group labels may correspond to demographic attributes such as gender or race. Subpopulation shift occurs when the distribution of subpopulations in the training data differs from that in the evaluation data. For example, in the training set (or source), there might be an equal number of positive and negative instances (indicated by color in Figure 1) and an equal number of instances across groups $s \in \mathcal{S}$ (left and right, separated by a line in Figure 1). However, the subpopulations differ: only 5% of positive (red) instances are in group $s = 0$ (left side) and only 5% of negative (blue) instances are in group s=1 (right side), while the test (or target) distribution is even across all subgroups. This spurious correlation in the training dataset causes the model to use a *shortcut* by basing its predictions on the majority subpopulation(s), leading to the misclassification of minority subpopulations. Generally, there is an implicit mapping from prediction targets y to groups s, such that if y' is

mapped to s', then $\mathbb{P}_Q(Y = y', S = s') \gg \mathbb{P}_Q(Y = y', S = s) \forall s \in \mathcal{S}, s \neq s'$. In other words, for each prediction target, there is exactly one group where that target is significantly more common than in the other groups, creating a spurious correlation.

Importance weighting. Importance weighting is a common technique to address the discrepancy between the source distribution Q and the target distribution P. The idea is to reweight the training samples to mimic learning from the target distribution. Considering P and Q as probability measures on $\mathcal{X} \times \mathcal{Y} \times \mathcal{S}$, the true classifier risk with respect to the target distribution P can be expressed in terms of source distribution Q as follows:

$$\mathbb{E}_{P(x,y,s)}[L\left(f_\theta(x), y\right)] = \mathbb{E}_{Q(x,y,s)}\left[\underbrace{\frac{P(x,y,s)}{Q(x,y,s)}}_{:=w(x,y,s)} L\left(f_\theta(x), y\right)\right],$$

provided that P and Q have the support over $\mathcal{Y} \times \mathcal{S}$. We can then express the weighted empirical loss $\hat{R}_w$ with weighting w:

$$\hat{R}_w(f_\theta) = \frac{1}{n}\sum_{i=1}^{n} w(x_i, y_i, s_i)\, L\left(f_\theta(x_i), y_i\right) \tag{2}$$

and $\hat{\theta}^w = \arg\min_{\theta \in \Theta} \hat{R}_w(f_\theta)$ the reweighed model. To account for the spurious correlation in the source distribution Q, we assume equal base rates under the target distribution P. For instance, in a setting with binary targets and binary groups, this implies that $\mathbb{P}(Y = 1|S = 0) = \mathbb{P}(Y = 1|S = 1)$. More generally, target labels Y and group labels S under P are independent (e.g. [18,7,20]):

$$w(x,y,s) = \frac{P(x,y,s)}{Q(x,y,s)} = \frac{\mathbb{P}(Y=y)\mathbb{P}(S=s)\cancel{\mathbb{P}(X=x|Y=y,S=s)}}{\mathbb{P}(Y=y,S=s)\cancel{\mathbb{P}(X=x|Y=y,S=s)}} \tag{3}$$

where we have assumed that marginal distributions within every subpopulation remain unchanged. According to the overlap assumption where $\mathbb{P}(Y, S)$ and $\mathbb{P}(Y)\mathbb{P}(S)$ are non-zero for any pair of (y, s) which implies that $0 < w < \infty$. An alternative weighting strategy is to use the inverse frequency of each subpopulation in the labelled pool, defined as $w(x, y, s) = 1/\mathbb{E}_{s' \sim Q}[\mathbb{I}(s' = s)]$. This method optimizes for a target distribution P with uniform group frequencies. By upweighting the minority groups, this approach aims to balance average and worst-group errors. However, in practice, upweighting minority groups does not always result in low training losses across all groups, as some groups may be inherently easier to fit than others [34]. Importantly, this weighting strategy is ineffective if Q already has a uniform marginal distribution over groups.

4 Active learning under subpopulation shift

Consider a scenario where the initial labelled set $\mathcal{L}_0$ consists of samples uniformly drawn from $\mathcal{U}$ which has experienced a subpopulation shift. As illustrated in

Figure 1, the presence of a spurious correlation in $\mathcal{L}_0$ leads to a biased classifier that consistently misclassifies the minority group. Consequently, the active learner employs this biased classifier to assess the informativeness of unlabelled samples, resulting in sampling bias [9]. Margin-based active learning approaches [33,5] select samples near the decision boundary, and for support vector machines, labelling points within the current margin tends to decrease the margin, thus reducing classification uncertainty. However, biased predictions in this context prompt the active learner to predominantly sample from the majority subgroup (represented by points from subpopulations a and c in Figure 1, left), exacerbating the shift towards the minority subgroup (represented by points from subpopulations b and d in Figure 1, left).

4.1 Active Learning via Source-Target Disagreement

Under subpopulation shift, the objective of the AL is to develop a fair classifier that mitigates the adverse impacts of spurious correlations in the labelled pool. One approach could involve integrating a fairness constraint into the acquisition function. A frequently used fairness constraint is the *equalised odds disparity* [15], defined as

$$\Delta(f_\theta) = \max_{y \in Y,\ s,s' \in S} \left| R_{y,s}(f_\theta) - R_{y,s'}(f_\theta) \right|, \tag{4}$$

where $R_{y,s}(f_\theta) := \mathbb{E}_{(x,y)|y=y,s=s}\left[L(f_\theta(x), y)\right]$. It measures the maximum performance gap across all group labels $\mathcal{S}$ for a given target label $y \in \mathcal{Y}$. A high disparity suggests that the risk associated with predicting a specific class y differs significantly between groups s and s'. Approaches aiming to alleviate the disparity often operate under the assumption of fixed training data. This prompts the inquiry of whether acquiring additional training samples could satisfy such fairness criteria. Hence, the agent's goal is to select a new sample x' in a manner that minimizes the equalized odds disparity.

$$x' = \arg\min_{x \in \mathcal{U}} \Delta\left(f_{\theta(\mathcal{L}_t \cup \{(x,y)\})}\right) \tag{5}$$

A intuitive approach to optimizing this objective is to enhance the representation of minority subgroups within $\mathcal{L}$. However, achieving this objective is computationally challenging without complete knowledge of (y, s) in the unlabelled pool. Without access to labels, one can only estimate the expected objective by training new classifiers $f_{\theta(\mathcal{L} \cup \{(x,y')\})}$ for every $x \in \mathcal{U}$ across all possible $y' \in \mathcal{Y}$ [2]. This approach remains computationally infeasible for large $\mathcal{U}$ and time-consuming learning tasks. One might attempt to

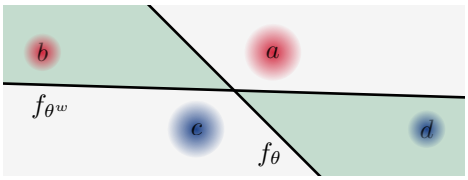

Fig. 2: Illustration of disagreement between f_θ and f_{θ^w}. The two hypotheses f_θ and f_{θ^w} are optimised using ERM and re-weighting, as in Equation (2), respectively. The shaded area delineates the region of disagreement where $f_\theta(x) \neq f_{\theta^w}(x)$.

Algorithm 1 Active Learning via Source-Target Disagreement (AL-STD)

Input: Labelled pool $\mathcal{L}_0$, Unlabelled pool $\mathcal{U}$, and Batch size k
for $t \in \{1, \dots, T\}$ **do**
  obtain $f_{\hat{\theta}}$ by Equation (1)
  obtain $f_{\hat{\theta}^{\hat{w}}}$ by Equation (2) using the estimated weights from Equation (7).
  for each $x \in \mathcal{U}$ **do**
    compute $\phi(x) = \|f_{\hat{\theta}}(x) - f_{\hat{\theta}^{\hat{w}}}(x)\|$
  end for
  label k points with the highest $\phi(x)$, $\mathcal{H} \leftarrow \{(x_i, y_i, s_i)\}_{i \in [k]}$
  update $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \mathcal{H}$, $\mathcal{U} \leftarrow \mathcal{U} \setminus \mathcal{H}$
end for
Return f_θ

relax Equation (4) by using predicted labels $(\hat{y}, \hat{s})$ through inference and unsupervised learning. However, this method may not accurately represent the original objective because the predictions will be biased, subsequently impacting the estimates of s.[4]

In this work, we introduce an alternative approach referred to as AL-STD, outlined in Algorithm 1. Suppose the labels for target (y) and group (s) are unavailable for unseen data in $\mathcal{U}$, and an oracle (annotator) furnishes (y, s) upon querying for the label of x. Let f_θ and f_{θ^w} denote two hypotheses optimized using ERM and re-weighting (as in Equation (2)), respectively. It is expected that f_θ and f_{θ^w} would concur on their predictions for the majority subgroups, yet diverge on minority subgroups (refer to Figure 2). The area of disagreement encompasses the subgroups correctly classified by f_{θ^w} but misclassified by f_θ. Samples near the disagreement region are regarded as spuriously correlated, contributing to the decline in f_θ's performance on the target distribution P. Hence, acquiring samples within the disagreement region entails giving more weight to minority subgroups, thereby alleviating the disparity between Q and P.

At each iteration t, AL-STD initially calculates biased predictions $\hat{y} = \{f_\theta(x) : x \in \mathcal{U}\}$ and unbiased predictions $\hat{y}_w = \{f_{\theta^w}(x) : x \in \mathcal{U}\}$. Subsequently, it requests labels for the sample x' that maximizes the predictive dissimilarity between f_θ and f_{θ^w}:

$$x' = \arg\max_{x \in \mathcal{U}} \phi(x) \ . \tag{6}$$

In the case of multi-class classification, $\hat{y}$ and $\hat{y}_w$ are softmax outputs and we define the dissimilarity as $\phi(x) := \|f_\theta(x) - f_{\theta^w}(x)\|$. Although w in Equation (3) is unknown in practice, we can estimate it from $\mathcal{L}$:

$$\hat{w}(x, y, s) = \frac{\widehat{P}(x, y, s)}{\widehat{Q}(x, y, s)} = \frac{\widehat{\mathbb{E}}[\mathbb{I}(y' = y)] \times \widehat{\mathbb{E}}[\mathbb{I}(s' = s)]}{\widehat{\mathbb{E}}[\mathbb{I}(y' = y \wedge s' = s)]}, \tag{7}$$

[4] In [39], the group label of a data point is estimated via unsupervised clustering conditioned on the model prediction: $\mathbb{P}(S = s|\hat{y}, x)$.

where $\widehat{\mathbb{E}}[\cdot]$ is the empirical measure over $\mathcal{L}$. The empirical loss with estimated importance weights can be derived from Equation (2) by replacing w with $\hat{w}$.

4.2 Theoretical motivations

The discrepancy between the source distribution Q and the target P originating from the subpopulation shift, renders the current problem inherently akin to that of *domain adaptation*. The following theorem quantifies the effect that the difference between the source and target distributions has on the success of domain adaptation learning:

Theorem 1 (Generalisation bounds of domain adaptation [29]). *Denote by $g : \mathcal{X} \to \mathcal{Y}$ the labelling function, for every $f_\theta \in \mathcal{H}$, the following holds:*

$$R_P(f_\theta, g) \leq R_Q(f_\theta, g) + \mathrm{disc}(P, Q) + \lambda \tag{8}$$

$$\textit{where:} \quad \lambda = \inf_{\theta \in \Theta} \left\{ R_P(f_\theta, g) + R_Q(f_\theta, g) \right\}; \qquad R_D(f, f') = \mathbb{E}_D \left[L(f(x), f'(x)) \right], \tag{9}$$

$$\mathrm{disc}(P, Q) = \sup_{\theta, \theta' \in \Theta} \left| R_P(f_\theta, f_{\theta'}) - R_Q(f_\theta, f_{\theta'}) \right|. \tag{10}$$

Theorem 1 shows that the lower the discrepancy $\mathrm{disc}(P, Q)$ the stronger guarantees one has for the performance of the learnt hypothesis on the target distribution. In contrast to the static domain adaptation scenario, in AL, we have the privilege of altering the source distribution by adding new samples to the labelled pool. One can see the querying process of AL-STD as a way to reduce the discrepancy $\mathrm{disc}(P, Q)$. Using assumption defined in Equation (3) we can re-write the discrepancy as follows:

$$\mathrm{disc}(Q, P) = \sup_{f, f'} \left| \sum_{y,s} \left(1 - \frac{\mathbb{P}(Y = y)\mathbb{P}(S = s)}{\mathbb{P}(Y = y, S = s)} \right) \mathbb{E}_{x \sim P(x|Y=y, S=s)} L(f(x), f'(x)) \right|. \tag{11}$$

A way to maximise the right-hand side is to pick f and f' such that they maximally agree on cases where $\mathbb{P}(Y = y)\mathbb{P}(S = s) < \mathbb{P}(Y = y, S = s)$, and thus the corresponding term has a small positive weight, and maximally disagree on cases where $\mathbb{P}(Y = y)\mathbb{P}(S = s) \gg \mathbb{P}(Y = y, S = s)$, as those terms will be heavily negatively weighted. While f_θ and f_{θ^w} do not necessarily attain the exact sup, they still provide a good estimate as they are expected to agree on majority subgroups and disagree on minority subgroups. Therefore, the region of disagreement of weighted and unweighted hypotheses captures the differences between the distributions in connection to the hypothesis set used. By selecting samples with the highest disagreement, one expects to gradually reduce the discrepancy. We observe evidence supporting this intuition in our experimental evaluation (see Figure 5b). Thus, the acquisition strategy of AL-STD reduces the effects of the subpopulation shift over time, leading to improved performance of the learnt hypothesis on the target distribution.

5 Experiments

We validate the effectiveness of AL-STD through experiments conducted on four openly accessible datasets: Coloured MNIST [3] (CMNIST), CelebA [28], Adult Income [22], and CivilComments [6], which encompass image, tabular, and text datasets. Detailed information regarding the experimental setups, including dataset descriptions and configurations, can be found in Appendix A.

5.1 Experimental setup

The AL process iterates until a predetermined labelling budget is depleted. Each iteration involves training the ERM classifier on the labelled samples and subsequently employing the active learner to procure new samples. Following the same procedure as [12], we reinitialise the classifier's parameters at the start of each iteration to eliminate the dependency between consecutive acquisitions.

Baselines. Our baselines cover a wide range of AL methods including uncertainty-based, diversity-based and hybrid ones.

1) **Random**: Uniformly sample x from $\mathcal{U}$ to label; resembles passive learning.
2) **Margin [33]**: An uncertainty-based AL method that selects samples with highest uncertainty, defined by the difference of the top two class probabilities predicted by f_θ (lower implies more uncertain), e.i. $\mathbb{P}(y_1|x, f_\theta) - \mathbb{P}(y_2|x, f_\theta)$ where y_1 and y_2 correspond to the first and the second most probable classes by prediction respectively.
3) **BADGE [4]**: An AL method that incorporates uncertainty and diversity. It first computes the gradient embeddings $g(x)$ w.r.t. the penultimate layer for every $x \in \mathcal{U}$. $g(x)$ is computed by the product of Jacobian and the output of the penultimate layer. Since Jacobians are class-wise, $g(x)$ is a concatenation of gradient embeddings computed for all possible $y \in \mathcal{Y}$. After that, it selects samples by the k-MEANS++ seeding algorithm.
4) **d-Margin**: Replaces f_θ with the debiased model f_θ^w in **Margin**, using reweighing from Equation (3).
5) **CoreSet [36]**: A diversity-based AL method that solely explores informativeness in the feature space. It selects k samples by solving a k-centre problem on the output of the penultimate layer.
6) **FAL** [2]: A fair AL method that incorporates group information. It assumes that group labels in $\mathcal{U}$ are accessible prior to acquisition. Given a fairness metric M_{fair} (e.g. equalised odds), the goal of FAL is to query a sample $x \in \mathcal{U}$ that maximally reduces the unfairness: $M_{fair}\left(f_{\theta(\mathcal{L})}\right) - M_{fair}\left(f_{\theta(\mathcal{L}\cup\{(x,y)\})}\right)$.

Subpopulation shift setting. We synthesise subpopulation shifts for datasets with binary $\mathcal{Y}$ and binary $\mathcal{S}$ by downsampling the two subgroups. The sampling probability for the minority subgroup $G_0 \in \mathcal{Y} \times \mathcal{S}$ and the majority subgroup

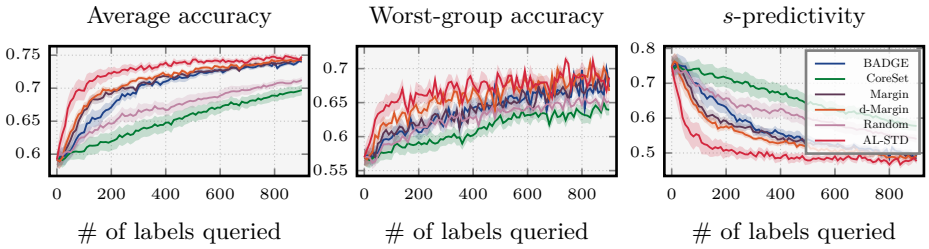

Fig. 3: Evaluation on CelebA. Our proposed algorithm, AL-STD, consistently surpasses the baselines with higher average accuracy and worst-group accuracy, alongside lower s-predictivity, within a low to medium budget (10–500 samples on the number of labels queried). Beyond this range, all methods (excluding Random and CoreSet) converge to similar performance levels. Utilizing uncertainty sampling with a debiased classifier (d-Margin) leads to marginal performance enhancements, with no improvement observed in CMNIST as depicted in Figures C.6 and C.7 in Appendix C.

$G_1 \in \mathcal{Y} \times \mathcal{S}$ will be $\mathbb{P}(G_0) = \alpha \in \left(0, \frac{1}{4}\right]$ and $\mathbb{P}(G_1) = \frac{1}{2} - \alpha$, respectively. The parameter α controls the degree to which Y and S are confounded. The confounding relationship is strengthened as $\alpha \to 0$ while it disentangles when $\alpha \to \frac{1}{4}$. The list of minority subgroups chosen for the experiments can be found in Appendix A.

Metrics. For every AL step, we retrain f_θ with the updated labelled pool and compute the following metrics: i) **average accuracy**, the average number of correct predictions over all subgroups: $\mathbb{E}_{(x,y)}\left[\mathbb{I}(f_\theta(x) = y)\right]$; ii) **worst–group accuracy** [27,41], the minimum accuracy over subgroup $\mathcal{G} = \mathcal{Y} \times \mathcal{S}$: $\min_{g' \in \mathcal{G}} \mathbb{E}\left[\mathbb{I}(f_\theta(x) = y) \mid g'\right]$; iii) **$s$-predictivity**, the mean square contingency coefficient between $\hat{y}$ and s (also known as Φ-coefficient). This coefficient quantifies the association between the predictions $\hat{y}$ and the true group label s. We present the absolute value of this coefficient (hence, $0 \leq \Phi(f_\theta, s) \leq 1$). A value close to 1 suggests that the function f_θ is predicting the group label s rather than the target label y, hence a smaller value is more preferable.

5.2 Results

We present the empirical performance of AL-STD alongside other baselines across image datasets (CMNIST, CelebA), a tabular dataset (Adult Income), and a text dataset (CivilComments). In the main article, we report the results using an acquisition size of 10 ($k = 10$) and $\alpha = 0.02$, except for CivilComments and Adult Income. For further evaluations and an extensive ablation analysis covering different setups, including varying confounding factors α and acquisition sizes k, refer to Appendix C. We construct initial labelled pools by randomly selecting samples from the unlabelled pool, with 2000 samples for Coloured MNIST, 100 for CelebA and Adult Income, and 500 for CivilComments. For

additional information on datasets, model architectures, and hyperparameters, refer to Appendix A and Appendix B. All results present the mean and standard error of 10 trials.

Results for Image modality – CelebA and CMNIST. On CelebA, regarding average accuracy, AL-STD exhibits significant superiority over all baselines under subpopulation shift, as evident from Figure 3. Despite being in the same family, CoreSet and BADGE perform comparatively poorly. One explanation proposed by [4] is that the representations at the penultimate layer may lack meaningfulness, leading to CoreSet's performance potentially worse than random sampling. Similarly, comparing Margin and d-Margin, the utilization of a debiased classifier improves performance. However, d-Margin does not demonstrate such enhancements on the CMNIST dataset (see Figures C.6 and C.7 in Appendix C). Moreover, AL-STD's worst-group accuracy noticeably surpasses that of the baselines, indicating that ERM trained on samples acquired through AL-STD does not compromise any subgroup's performance. Given the initial highly biased labelled pool, all methods start with an s-predictivity of about 0.75, signifying a strong correlation between prediction and gender. All baselines, including AL-STD, succeed in reducing s-predictivity as more samples are acquired, with random sampling expected to converge to a relatively high score compared to other methods. It might appear surprising that even with random sampling, s-predictivity decreases as more labelled samples become available; after all, a spurious correlation exists in the unlabelled pool. However, there exist samples for every combination of s and y in the data, enabling ERM to eventually find the solution that minimizes the loss, favouring prediction of y over s. The challenge for a small labelled pool arises because with limited information, models tend to rely on simplistic rules (due to simplicity priors), which in our scenario translates to leaning on shortcuts (predicting s instead of y).Similar trends are observed in CMNIST, as illustrated in Figures C.6 and C.7 in Appendix C, although d-Margin does not exhibit the same improvements as observed in CelebA; in fact, it performs much worse than Margin.

Does target distribution need to be group-wise balanced? No, it is important to emphasise that group-wise balance in the target distribution isn't necessary; instead, the key is that Y and S are statistically independent, a scenario often encountered in real-world settings. We assessed AL-STD on an imbalanced CelebA test set. Figure 4a illustrates the evaluation on the test set drawn from distribution P where $\mathbb{P}_P(Y, S) = \mathbb{P}_P(Y)\mathbb{P}_P(S)$, with $\mathbb{P}_P(Y = 0) \neq \mathbb{P}_P(Y = 1)$ and $\mathbb{P}_P(S = 0) \neq \mathbb{P}_P(S = 1)$. Despite the performance degradation compared to Figure 3 due to the imbalanced nature, AL-STD still exhibits the best performance.

Does the method scale to a multi-class-multi-group setting? We evaluated AL-STD on this task on CMNIST (see Appendix A for details). The superior performance of AL-STD shown in Figure 4b suggests its scalability to any number of classes and groups. This scalability is indeed feasible since we only need to compute the importance weights for all pairs of $(y, s) \in \mathcal{Y} \times \mathcal{S}$.

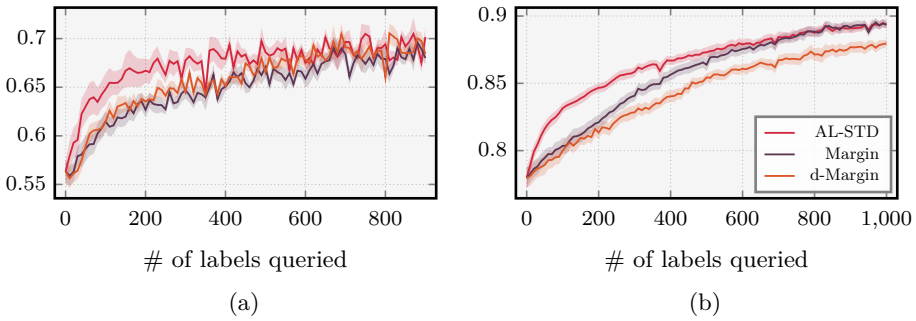

Fig. 4: Average accuracy for (a) imbalanced test set (CelebA); (b) multi-class-multi-group (CMNIST).

Results for Text modality – CivilComments. We conducted experiments with $\alpha = 0.2$ to simulate more realistic population shifts, resulting in a significantly lower s-predictivity at the initial step compared to the other two image datasets. From Figure 5a, we observe that AL-STD achieves approximately a $\times 2.7$ reduction in s-predictivity, while the best baseline (BADGE and Margin) achieves only a $\times 1.8$ reduction. It is noteworthy that d-Margin does not improve performance as significantly as observed on the CelebA dataset. We hypothesize that the benefits of using an unbiased classifier depend on the model choice and the dataset characteristics. This result also underscores the versatility of AL-STD across various tasks.

Results for Tabular data – Adult Income. We use the Adult Income dataset to evaluate performance for *inherent subpopulation shifts* and to compare AL-STD with FAL. FAL is solely assessed on this dataset due to its computational complexity for the other datasets. Following [2], we use *equalised odds* as the metric. As illustrated in Figure 6, despite FAL's explicit goal of minimizing unfairness concerning the fairness metric, we do not observe a declining trend as more acquisitions are made. Instead, it fluctuates around its initial value, while Random exacerbates unfairness due to escalating shifts. These results also highlight that AL-STD mitigates unfairness without requiring a fairness-specific objective function.

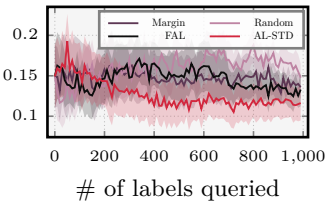

Fig. 6: Equalised odds difference on Adult Income. A high value implies less equality of odds over groups; thus, lower is better. See Figure C.10 for other metrics.

Analysis of AL-STD. AL-STD is motivated by the intuition that its acquisition function boosts the representation of minority subgroups in $\mathcal{L}_t$, thereby reducing the gap between the source Q and target P distributions. To understand the mechanism behind AL-STD, we visualize the evolution of $\hat{w}(G_0)$, the importance

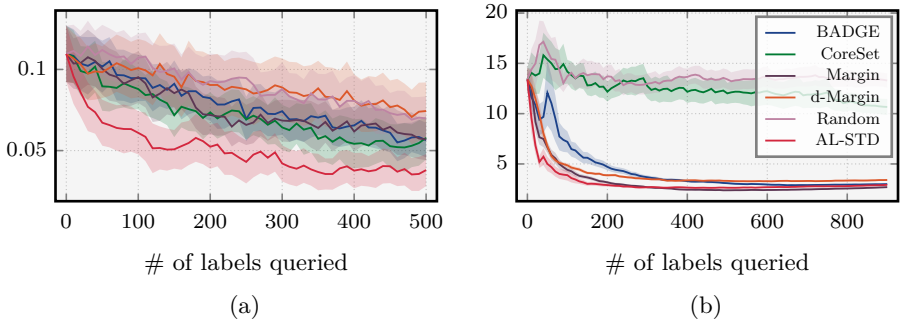

Fig. 5: (a) s-predictivity on CivilComments dataset. See Figure C.9 for other metrics. (b) The estimated importance weights aggregated over the minority subgroups: $\mathbb{E}_{G' \in G_0}[\widehat{P}(G')/\widehat{Q}(G')]$. Initially due to the high mismatch between P and Q, the minority subgroup is given an excessive weighting and it decreases exponentially as label acquisition is made, except for Random and CoreSet, while AL-STD shows the fastest decrease.

weight of the minority subgroup, over time, in Figure 5b. This illustrates that the weight $\hat{w}(G_0)$, especially in the initial stages, is notably higher than 1, indicating a substantial distributional shift between P and Q. As learning progresses, the weight gradually converges to a smaller value, suggesting a reduction in the discrepancy between P and Q: $Q \to P$. However, our empirical evaluation does not show that the weight converges to a value close to one. This phenomenon may be attributed to the fact that, ideally, acquiring new samples with Equation (6) should alter Q, the joint distribution of Y and S, while keeping P, the marginal distribution, unchanged. However, in practice, P also shifts due to prediction errors and noise in the data. This occurrence arises when samples from the majority subgroup are incorrectly classified as being in the region of disagreement, leading a well-trained model to effectively reduce the likelihood of such instances.

6 Conclusion

We have explored the concept of subpopulation shift in the AL framework, demonstrating the viability of constructing an AL algorithm – AL-STD – centred on the notion of disagreement regions to handle subpopulation shift. Compared to the baselines, AL-STD has twice the memory and time complexity: i) additional memory resources are required for f_{θ^w} ii) extra training time iii) extra inference time on $\mathcal{U}$. The time complexity can be reduced by running f_θ and f_{θ^w} in parallel. However, this will burden memory consumption. Thus, it is not possible to reduce both simultaneously. Dataset users should take extra care to perform a cost-benefit analysis for selecting particular datasets for their machine learning (ML) tasks. We should consider whether to start over with our initial labelled pool or even our unlabelled pool. We note the general agreement of our ML community

that any bias intervention – including the algorithmic solution presented here – will only be effective in tandem with broader awareness and thoughtfulness in building applications. Corrective actions such as bias interventions or, conversely, explicit *inaction* should be recorded. As future work, it might be feasible to make use of the size of the disagreement region in order to identify a good point to terminate the AL procedure. Furthermore, this work focused on subpopulation shift, but it might be possible to extend the method to also include a shift in marginal distributions within every subpopulation, which presumably involves exploring different weighting strategies.

Acknowledgments. This research was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. This work is supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme Grant Agreement no. 851538 - BayesianGDPR, Horizon Europe research and innovation programme Grant Agreement no. 101120763 - TANGO. Novi Quadrianto is also supported by BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIN/AEI/10.13039/501100011033. Viktoriia Sharmanska is currently at Epic Games.

References

1. Abernethy, J.D., Awasthi, P., Kleindessner, M., Morgenstern, J., Russell, C., Zhang, J.: Active Sampling for Min-Max Fairness. In: Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 53–65. PMLR (2022-07-17/2022-07-23)
2. Anahideh, H., Asudeh, A., Thirumuruganathan, S.: Fair active learning. Expert Systems with Applications **199**, 116981 (Aug 2022). https://doi.org/10.1016/j.eswa.2022.116981
3. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant Risk Minimization (Mar 2020). https://doi.org/10.48550/arXiv.1907.02893
4. Ash, J.T., Zhang, C., Krishnamurthy, A., Langford, J., Agarwal, A.: Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In: International Conference on Learning Representations (2020)
5. Balcan, M.F., Broder, A., Zhang, T.: Margin Based Active Learning. In: Learning Theory. vol. 4539, pp. 35–50. Springer Berlin Heidelberg, Berlin, Heidelberg (2007). https://doi.org/10.1007/978-3-540-72927-3_5
6. Borkan, D., Dixon, L., Sorensen, J., Thain, N., Vasserman, L.: Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification. In: Companion Proceedings of The 2019 World Wide Web Conference. pp. 491–500. ACM, San Francisco USA (May 2019). https://doi.org/10.1145/3308560.3317593
7. Chouldechova, A.: Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data **5**(2), 153–163 (Jun 2017). https://doi.org/10.1089/big.2016.0047

8. Creager, E., Jacobsen, J.H., Zemel, R.: Environment Inference for Invariant Learning. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 2189–2200. PMLR (2021-07-18/2021-07-24)
9. Dasgupta, S., Hsu, D.: Hierarchical Sampling for Active Learning. In: Proceedings of the 25th International Conference on Machine Learning. pp. 208–215. ICML '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1390156.1390183
10. El-Hasnony, I.M., Elzeki, O.M., Alshehri, A., Salem, H.: Multi-Label Active Learning-Based Machine Learning Model for Heart Disease Prediction. Sensors **22**(3), 1184 (Feb 2022). https://doi.org/10.3390/s22031184
11. Federici, M., Tomioka, R., Forré, P.: An Information-theoretic Approach to Distribution Shifts. In: Advances in Neural Information Processing Systems. vol. 34, pp. 17628–17641. Curran Associates, Inc. (2021)
12. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian Active Learning with Image Data. In: International Conference on Machine Learning. pp. 1183–1192. PMLR (Jul 2017)
13. Geifman, Y., El-Yaniv, R.: Deep Active Learning over the Long Tail (Nov 2017). https://doi.org/10.48550/arXiv.1711.00941
14. Hanneke, S.: Theory of Disagreement-Based Active Learning. Foundations and Trends® in Machine Learning **7**(2-3), 131–309 (2014). https://doi.org/10.1561/2200000037
15. Hardt, M., Price, E., Srebro, N.: Equality of Opportunity in Supervised Learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. pp. 3323–3331. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)
16. Hutchinson, B., Denton, E., Mitchell, M., Gebru, T.: Detecting Bias with Generative Counterfactual Face Attribute Augmentation. In: CVPR 2019 Workshop on Fairness Accountability Transparency and Ethics in Computer Vision (2019)
17. Kallus, N., Zhou, A.: Residual Unfairness in Fair Machine Learning from Prejudiced Data. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 2439–2448. PMLR (2018)
18. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems **33**(1), 1–33 (Oct 2012). https://doi.org/10.1007/s10115-011-0463-8
19. Kim, T., Lee, K.H., Ham, S., Park, B., Lee, S., Hong, D., Kim, G.B., Kyung, Y.S., Kim, C.S., Kim, N.: Active learning for accuracy enhancement of semantic segmentation with CNN-corrected label curations: Evaluation on kidney segmentation in abdominal CT. Scientific Reports **10**(1), 366 (Jan 2020). https://doi.org/10.1038/s41598-019-57242-9
20. Kleinberg, J., Mullainathan, S., Raghavan, M.: Inherent Trade-Offs in the Fair Determination of Risk Scores. LIPIcs, Volume 67, ITCS 2017 **67**, 43:1–43:23 (2017). https://doi.org/10.4230/LIPICS.ITCS.2017.43
21. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B.A., Haque, I.S., Beery, S., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: WILDS: A Benchmark of in-the-Wild Distribution Shifts. In: International Conference on Machine Learning (ICML) (2021)
22. Kohavi, R., Becker, B.: UCI Adult Data Set. UCI Meachine Learning Repository **5** (1996)

23. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big Transfer (BiT): General Visual Representation Learning. In: Computer Vision – ECCV 2020. pp. 491–507. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58558-7_29
24. Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., Mullainathan, S.: The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 275–284. ACM, Halifax NS Canada (Aug 2017). https://doi.org/10.1145/3097983.3098066
25. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE **86**(11), 2278–2324 (Nov/1998). https://doi.org/10.1109/5.726791
26. Lewis, D.D., Gale, W.A.: A Sequential Algorithm for Training Text Classifiers. In: SIGIR '94, pp. 3–12. Springer London, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_1
27. Liu, E.Z., Haghgoo, B., Chen, A.S., Raghunathan, A., Koh, P.W., Sagawa, S., Liang, P., Finn, C.: Just Train Twice: Improving Group Robustness without Training Group Information. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 6781–6792. PMLR (Jul 2021)
28. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep Learning Face Attributes in the Wild. In: Proceedings of International Conference on Computer Vision (ICCV) (Dec 2015)
29. Mansour, Y., Mohri, M., Rostamizadeh, A.: Domain Adaptation: Learning Bounds and Algorithms. In: COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009 (2009)
30. Noriega-Campero, A., Bakker, M.A., Garcia-Bulle, B., Pentland, A.S.: Active Fairness in Algorithmic Decision Making. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 77–83. ACM, Honolulu HI USA (Jan 2019). https://doi.org/10.1145/3306618.3314277
31. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
32. Rawls, J.: Justice as Fairness: A Restatement. Harvard University Press (May 2001). https://doi.org/10.2307/j.ctv31xf5v0
33. Roth, D., Small, K.: Margin-Based Active Learning for Structured Output Spaces. In: Machine Learning: ECML 2006. pp. 413–424. Springer Berlin Heidelberg, Berlin, Heidelberg (2006)
34. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. In: International Conference on Learning Representations (2020)
35. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter (Feb 2020)
36. Sener, O., Savarese, S.: Active Learning for Convolutional Neural Networks: A Core-Set Approach. In: International Conference on Learning Representations (2018)
37. Settles, B.: Active Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, Springer International Publishing, Cham (2012). https://doi.org/10.1007/978-3-031-01560-1

38. Sharaf, A., Daume Iii, H., Ni, R.: Promoting Fairness in Learned Models by Learning to Active Learn under Parity Constraints. In: 2022 ACM Conference on Fairness, Accountability, and Transparency. pp. 2149–2156. ACM, Seoul Republic of Korea (Jun 2022). https://doi.org/10.1145/3531146.3534632
39. Sohoni, N., Dunnmon, J., Angus, G., Gu, A., Ré, C.: No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In: Advances in Neural Information Processing Systems. vol. 33, pp. 19339–19352 (2020)
40. Zhao, E., Liu, A., Anandkumar, A., Yue, Y.: Active Learning under Label Shift. In: Proceedings of The 24th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 130, pp. 3412–3420. PMLR (2021-04-13/2021-04-15)
41. Zhou, C., Ma, X., Michel, P., Neubig, G.: Examining and Combating Spurious Features under Distribution Shift. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 12857–12867. PMLR (2021-07-18/2021-07-24)

A Datasets

CelebA[5] *[28].* We set the groups based on gender: $\mathcal{S} = \{\text{female}, \text{male}\}$ and the prediction targets based on whether the photographed person is smiling: $\mathcal{Y} = \{\text{not smiling}, \text{smiling}\}$. This choice of $\mathcal{S}$ and $\mathcal{Y}$ has been used in previous work by [16] (see Figure A.1 for sample images). We conduct experiments on two configurations containing two majority subgroups each: Setup (A) $\{(\text{not smiling}, \text{male}), (\text{smiling}, \text{female})\}$; Setup (B) $\{(\text{not smiling}, \text{female}), (\text{smiling}, \text{male})\}$.

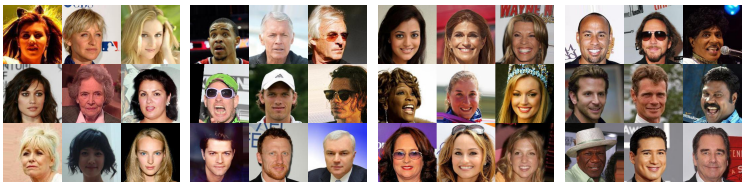

Fig. A.1: Sample images of CelebA. From left to right: female not smiling, male not smiling, female smiling and male smiling.

Coloured MNIST. Coloured MNIST [3] is based on the MNIST[6] dataset [25]. This dataset has been used in previous studies related to algorithmic fairness [8,11]. Two colours (red and blue) are applied randomly to the images, such that we get 2 groups: $\mathcal{S} = \{\text{red}, \text{blue}\}$. Furthermore, instead of the 10 digit classes that MNIST has as the prediction targets, we merely use two classes: label 0 for digits 0-4 (all digits < 5) and label 1 for digits 5-9 (all digits ≥ 5). Similarly, the two configurations are defined as follows: Setup (A) $\{(0, \text{blue}), (1, \text{red})\}$; Setup (B) $\{(0, \text{red}), (1, \text{blue})\}$. For multi-class-multi-group CMNIST, digits one to nine are split into triplets and form three classes $\mathcal{Y} = \{0, 1, 2\}$ and the groups are defined by three colours $\mathcal{S} = \{\text{red}, \text{blue}, \text{green}\}$. The minority subgroups are $\{(0, \text{red}), (1, \text{blue}), (2, \text{green})\}$. See Figure A.2 for sample images.

Fig. A.2: Sample images of Coloured MNIST.

CivilComments[7] *[6].* In CivilComments-WILDS [6,21], the prediction target is whether an online comment is toxic or non-toxic: $\mathcal{Y} = \{\text{toxic, non-toxic}\}$. This

[5] Available for non-commercial research purposes only.
[6] Creative Commons Attribution-Share Alike 3.0 license.
[7] CC0 1.0 Public Domain license

target label is spuriously correlated with mentions of certain demographic identities (male, female, White, Black, LGBTQ, Muslim, Christian, and other religions). We follow the experimental procedure of [21], which defines 16 overlapping groups for each of the above 8 demographic identities. Therefore, the group label $s = 1$ when a demographic identity is mentioned in the comment; otherwise $s = 0$. We experiment with the following configuration: $\{(\text{non-toxic}, \text{not indentified}), (\text{toxic}, \text{indentified})\}$.

Adult Income [22]. The groups are defined based on race: $\mathcal{S} = \{\text{black}, \text{white}\}$ and the prediction targets based on whether or not an individual's income exceeds \$50K per year: $\mathcal{Y} = \{\text{income} < \$50\text{k}, \text{income} \geq \$50\text{k}\}$. The features include attributes such as workclass, occupation, gender, education and etc. With this dataset, no synthesised subpopulation shifts were introduced.

B Experimental details

The training-test split ratio is 70:30 for all datasets. We use the training set as the unlabelled pool, but we apply subsampling to introduce distributional shifts. Our experiments utilised some existing codebases. We used the implementation from [4] for CoreSet and BADGE. For CivilComments and Adult Income, we used the codebase from [38] and the codebase from [27] for data preprocessing.

CelebA. The dataset contains more than 200,000 images. We subsample 50,000 images from the dataset for our experiments. Our CelebA model is a pre-trained ResNet-50 model [23]. We fine-tuned the head of the model with an SGD optimiser with a fixed learning rate of 0.001 for 80 epochs. During backpropagation, gradients are clipped between $[-0.5, 0.5]$, which we found improves stability. To improve performance, all input images are resized to 224×224 pixels.

Coloured MNIST. The foreground of every image from MNIST is coloured in red, blue and green. The Coloured MNIST model consists of 2 blocks of {`Conv2d`, `BatchNorm2d`, `ReLU`, `MaxPool2d`} as the backbone and one linear layer as the classifier. We trained the model with an SGD optimiser for 120 epochs, fixing the learning rate at 0.001.

CivilComments. Similar to CelebA dataset, we subsampled 50,000 sentences from the whole dataset. We fine-tuned the uncased DistilBERT [35] base model, following the standard procedure including using an AdamW optimiser, a learning-rate warmup schedular, an optimiser which implements gradient bias correction as well as weight decay, and applying weight decay to all parameters other than bias and layer normalisation terms. We used a learning rate of 1×10^{-6}, a weight decay coefficient of 0.1 and a batch size of 32. Our experiments are run on a GPU with 24GB memory. We report the results over 3 trials using the acquisition size of 5 ($k = 5$), $\alpha = 0.02$, and an initial labelled pool of 100

Adult Income. It is computationally expensive to evaluate FAL on the full dataset, so we only subsampled 10,000 data points for our experiment. Data points are standardised prior to training and inference. We use the `SGDClassifier` from Scikit-learn [31] with logistic loss, L2 regularisation and the maximum iteration of 2000.

C Additional results

Here, we present additional results for all datasets. In addition, we also study the performance of active learning algorithms with respect to acquisition size k and the confounding factor α. At a high level, the results are very similar for the different k and α. We observe that AL-STD enhances the average and the worst-case performance in the presence of spurious correlation in $\mathcal{U}$. see Figures C.3, C.4, C.6 and C.7. Apart from that, we also demonstrate the effectiveness of AL-STD in realistic scenarios in which the dataset contains a moderate spurious correlation ($\alpha = 0.2$, see Figure C.9) as well as an inherent correlation (see Figure C.10). Furthermore, AL-STD reduces s-predictivity better than other baselines even when α is high (less affected by spurious correlation). Figures C.5 and C.8 show that AL-STD is robust to the acquisition size. Figure C.11 provides the worst-group performance on imbalanced test set on CelebA and the multi-class-multi-group on CMNIST.

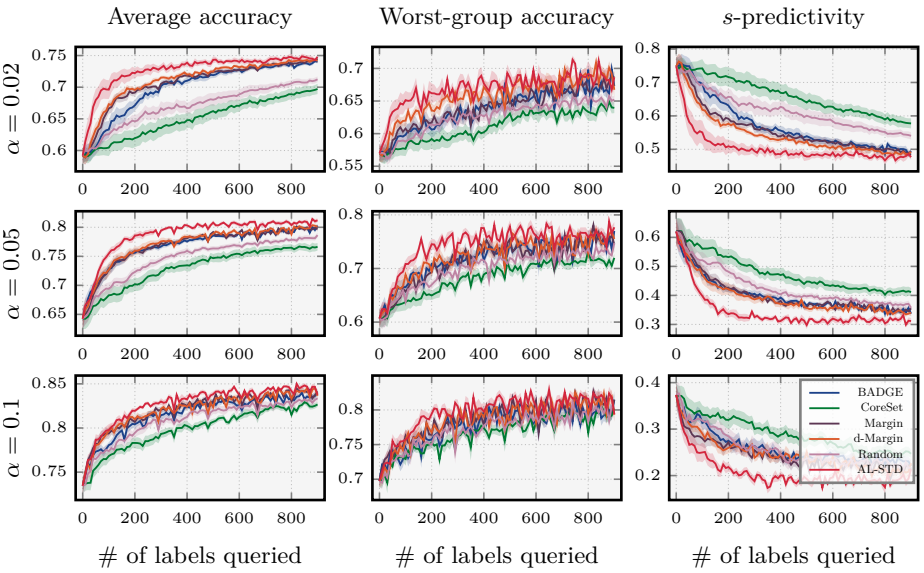

Fig. C.3: Performance on CelebA (Setup A) for varying confounding factor α. We observe that AL-STD consistently outperforms the baselines throughout the run in this configuration for varying α. The advantage of using d-Margin over Margin is less evident for large α. The performance of BADGE improves relative to that of other baselines when α increases.

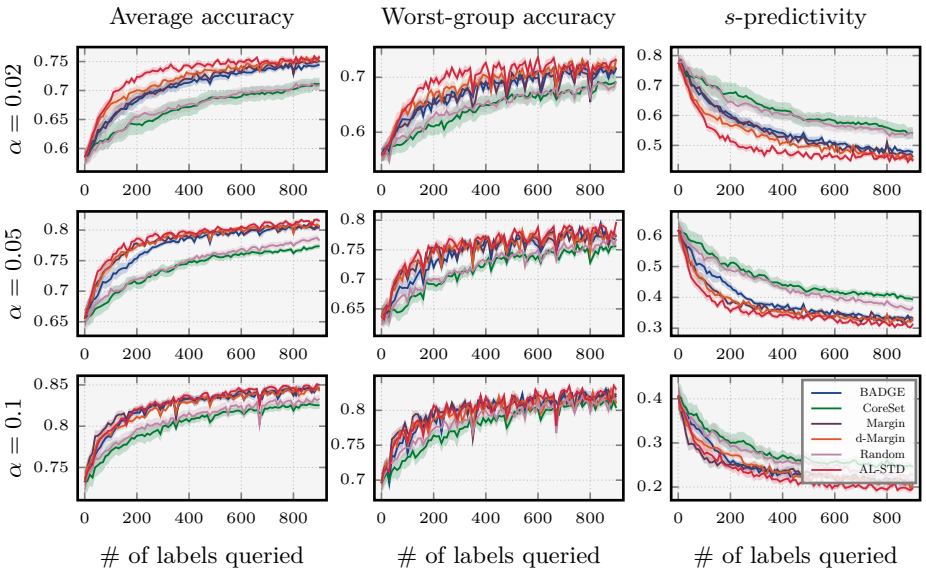

Fig. C.4: Performance on CelebA (Setup B) for varying confounding factor α. The performance of all methods improves as α increases (less affected by spurious correlation) in terms of average accuracy, worst-group accuracy and s-predictivity. Especially when $\alpha = 0.1$, AL-STD seems to be comparable with baselines (except Random and CoreSet) for # of labels queried ≤ 200. After that, it outperforms them until the labelling budget exhausted. CoreSet appears to work better for larger α but it still performs similarly to Random. Besides that, the advantage of d-Margin is less recognisable for larger α.

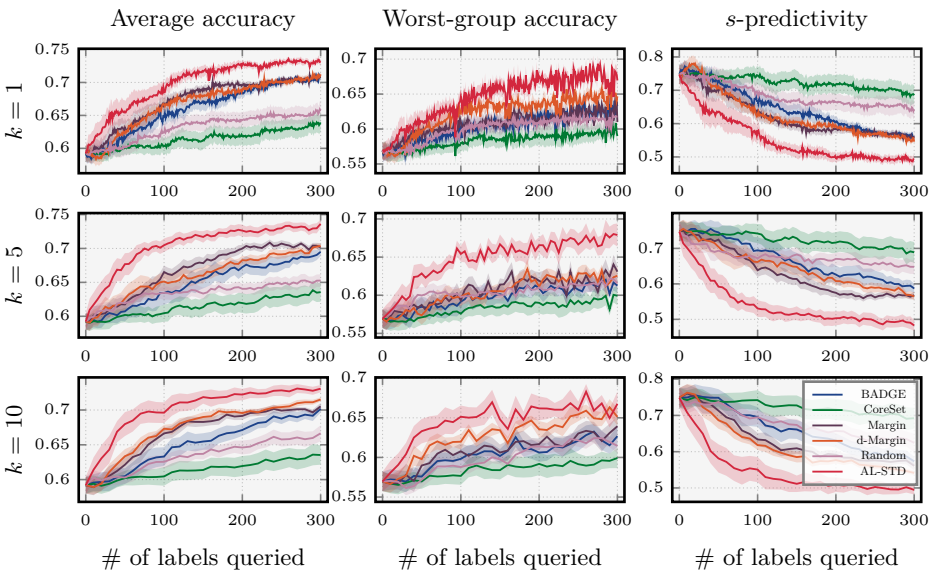

Fig. C.5: Performance on CelebA (Setup A, $\alpha = 0.02$) for varying acquisition size k. All methods are consistent across all acquisition sizes.

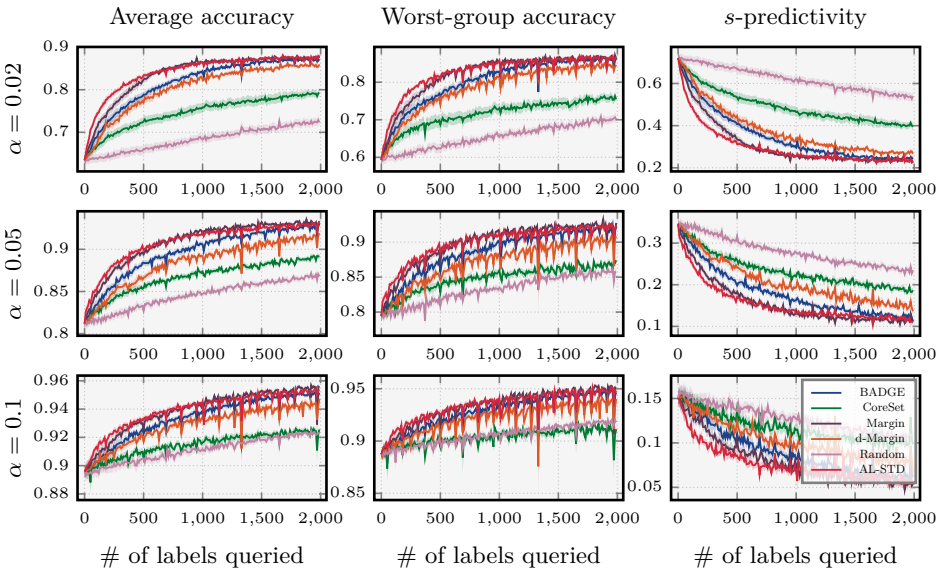

Fig. C.6: Performance on CMNIST (Setup A) for varying confounding factor α. In this configuration, we observe similar results to those in Figure C.7 except that for $\alpha = 0.05$, the performance of AL-STD is comparable to that of Margin in the early run.

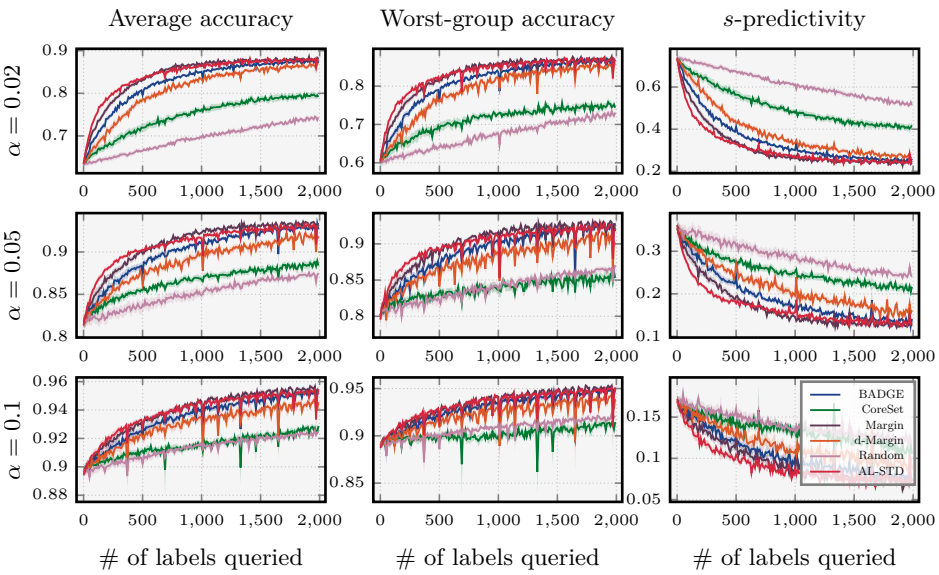

Fig. C.7: Performance on CMNIST (Setup B) for varying confounding factor α. As α increase, the advantage of AL-STD is less evident. Particularly, AL-STD exhibits a small margin of improvement in the early run.

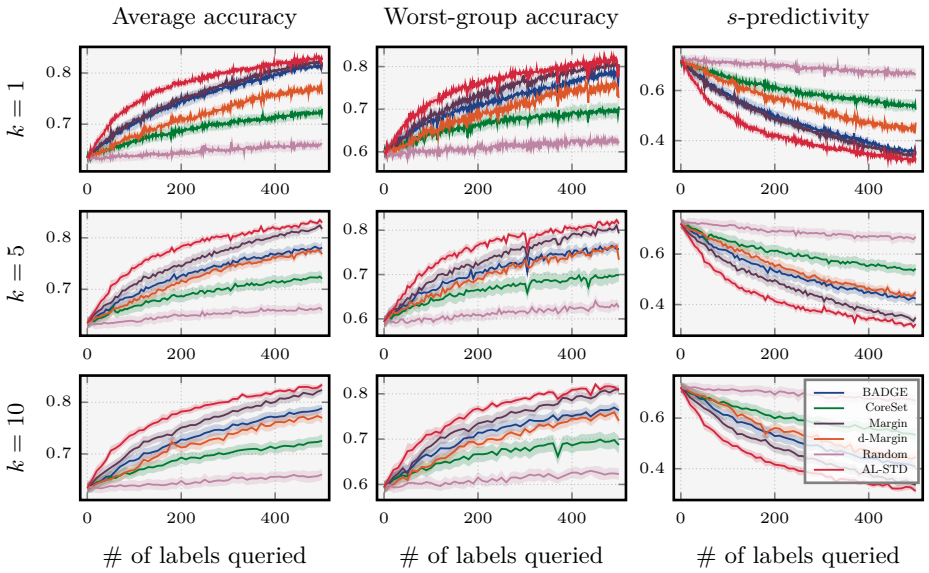

Fig. C.8: Performance on CMNIST (Setup A and $\alpha = 0.02$) for varying acquisition size k. Similar to Figure C.5, all methods perform consistently across various acquisition sizes expect BADGE, where it converged to around average accuracy of 81% when using a single batch size ($k = 1$), but around 78% when using $k > 1$.

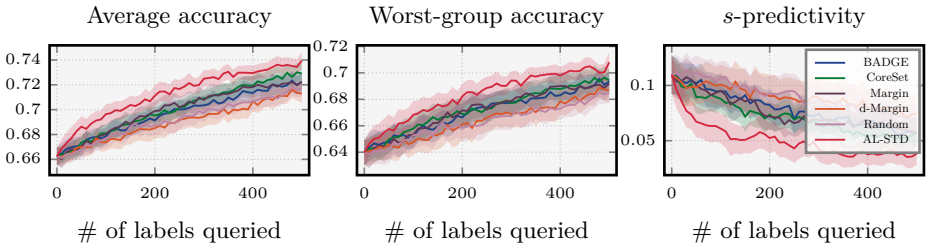

Fig. C.9: Evaluation on CivilComments ($\alpha = 0.2$). We can see that due to the moderate distributional shift, the worst-group accuracy is close to the average accuracy. In terms of average accuracy, Random and d-Margin have similar performance. AL-STD converged to a worst-group accuracy of around 67% as opposed to all other baselines' 63%.

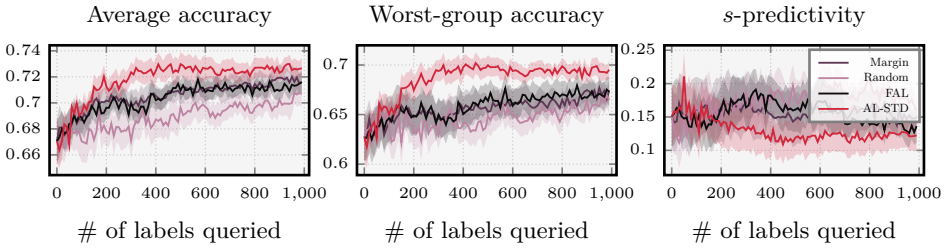

Fig. C.10: Evaluation on Adult Income (inherent shift). The large difference between the average accuracy and the worst-case accuracy indicates that the dataset is biased (or imbalanced over subgroups). We observed that AL-STD improves both metrics as more samples are acquired, while the worst-group accuracies of baselines remained close to the initial values.

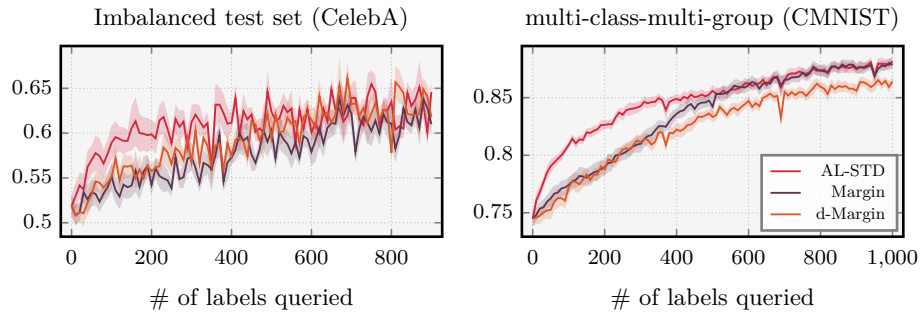

Fig. C.11: The y-axis in the figures represents the worst-group accuracy. The figure on the left shows the worst-group accuracy evaluated on an imbalanced test set which is in accordance with the average accuracy shown in main text. While the right figure shows the worst-group accuracy on the CMNIST multi-class-multi-group task, the curve appears to be less noisy compared to the binary case. The worst-group performance of AL-STD is almost 10% higher than that of baselines in the early run.