Boosting Test Performance with Importance Sampling–a Subpopulation Perspective

Hongyu Shen, Zhizhen Zhao

Department of Electrical and Computer Engineering,
University of Illinois at Urbana-Champaign, Champaign, IL, 61820, U.S.A.
{hongyu2, zhizhenz}@illinois.edu

Abstract

Despite empirical risk minimization (ERM) is widely applied in the machine learning community, its performance is limited on data with spurious correlation or subpopulation that is introduced by hidden attributes. Existing literature proposed techniques to maximize group-balanced or worst-group accuracy when such correlation presents, yet, at the cost of lower average accuracy. In addition, many existing works conduct surveys on different subpopulation methods without revealing the inherent connection between these methods, which could hinder the technology advancement in this area. In this paper, we identify important sampling as a simple yet powerful tool for solving the subpopulation problem. On the theory side, we provide a new systematic formulation of the subpopulation problem and explicitly identify the assumptions that are not clearly stated in the existing works. This helps to uncover the cause of the dropped average accuracy. We provide the first theoretical discussion on the connections of existing methods, revealing the core components that make them different. On the application side, we demonstrate a single estimator is enough to solve the subpopulation problem. In particular, we introduce the estimator in both attribute-known and -unknown scenarios in the subpopulation setup, offering flexibility in practical use cases. And empirically, we achieve state-of-the-art performance on commonly used benchmark datasets.

Code — https://github.com/skyve2012/DBA
Extended version — https://arxiv.org/abs/2412.13003

1 Introduction

Empirical risk minimization (ERM) often struggles with distribution shifts that manifest when the training and test distributions differ (Bickel, Brückner, and Scheffer 2007; Quionero-Candela et al. 2009; Shimodaira 2000). One ubiquitous type of distribution shift is *subpopulation shift*, which describes a scenario where the portion of the subpopulations may vary between training and testing sets. See Figure 1 for an example. This consequently leads to degraded performance when a trained model is applied to production/testing environments (Yang et al. 2023). Ensuring that machine learning models are robust against these distribution shifts hence is crucial for their reliability and safe real-world application.

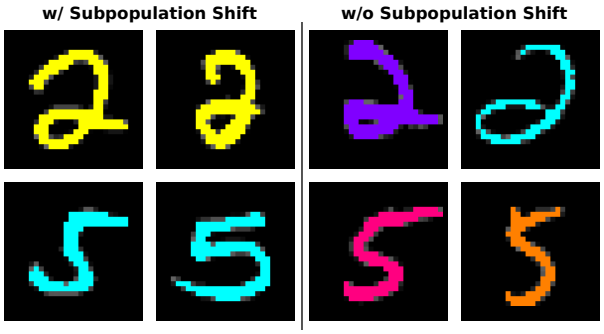

Figure 1: An image example on subpopulation shift. The left panel contains images where digits and colors are correlated, whereas the right panel does not exhibit such correlation.

Existing works proposed different methods in the forms of auxiliary losses (Li et al. 2018; Arjovsky et al. 2019; Alshammari et al. 2022), data augmentations (Zhang et al. 2018; Yao et al. 2022; Han et al. 2022), modeling objectives (Liu et al. 2021; Sagawa et al. 2020; Japkowicz 2000; Wu et al. 2023; Nam et al. 2020; Asgari et al. 2022; Rudner et al. 2024; Han and Zou 2024; Hong et al. 2023; Tsirigotis et al. 2024; Menon et al. 2021; Lin et al. 2017) and data sampling techniques (LaBonte, Muthukumar, and Kumar 2024; Izmailov et al. 2022; Japkowicz 2000). They all exhibit superior performance on worst group accuracy while maintaining high accuracy in the overall set. However, two recent works experimentally observed that most models experience a drop in average accuracy performance compared to the ERM setup despite the high worst group accuracy (Tsirigotis et al. 2024; Yang et al. 2023). Nonetheless, none of the papers is able to provide rigorous explanations on the answer to why. The lack of clarity in understanding can impede the development of appropriate models and methods, potentially stalling progress in the field.

In this work, we propose a systematic dataset bias analysis (DBA) framework that is rooted in importance sampling. With this framework, we reveal the cause of the lower-than-ERM average accuracy is the mismatch between the learning objective and the testing dataset. Moreover, we identify the flexibility of this framework in interpreting the formulation of some of the existing works that focus primarily on statistical heuristics and do not specify the underlying assumptions of

the models or data. The DBA framework, on the other hand, can close the gap, allowing us to explicitly discuss assumptions systematically and compare different existing works with the same language. We believe this analysis offers a comprehensive and theoretically grounded view to people who wish to proceed with the study of subpopulation methods.

Practically, we propose to estimate a single distribution given the conducted analysis using the DBA framework and prove that this is enough for solving the subpopulation problem under certain assumptions. Subsequently, we propose 3 different methods for estimating the distribution given different access levels to data and attributes. Empirically, we demonstrate the framework improves the test performance under subpopulation setups and achieves state-of-the-art (SOTA) results for both average and worst group accuracy while avoiding the lower-to-ERM performance.

2 Related Work

Here we cover related works about importance sampling, the survey papers of the subpopulation shift, and the associated SOTA methods. An extension is included in Appendix.

2.1 Importance Sampling

DBA interprets distributional shift as a mismatch of the weight function from an importance sampling perspective. Although primarily focused on subpopulation setups, the method's formulation applies broadly to distributional shift problems. Early works on importance sampling (Shimodaira 2000; Huang et al. 2006) address dataset shifts but lack real-world experiments and clarity for subpopulation cases. In contrast, DBA systematically formulates the application to subpopulation problems, explicitly stating assumptions and identifying key components like distributions leading to such issues. Other studies (Kanamori, Hido, and Sugiyama 2009; Fang et al. 2020) propose weight estimation methods requiring partial test set access, unlike DBA. Additionally, DBA considers the weight function as the ratio of joint distributions of x and y, addressing subpopulation and covariate shifts more realistically.

2.2 Subpopulation Survey

Yang et al. (2023) provides the first comprehensive experimental study on subpopulation methods. It uses Bayes' theorem to decompose $y|x$, accounting for attributes (spurious features), and categorizes datasets into four classes with varying label-attribute correlations. The paper benchmarks 20 subpopulation methods across these datasets but lacks statistical quantification of performance differences. Other surveys (Yu et al. 2024; Zhang et al. 2023) cover broader out-of-distribution (OOD) and domain generalization (DG) methods. While Yu et al. (2024) focuses on applications, Zhang et al. (2023) quantifies error inflation due to distribution shifts but doesn't address correction via model design. Our work extends prior studies by providing formal statistical analysis to quantify errors from both data and modeling perspectives. DBA also explains why some methods trade worst-case accuracy for lower average test accuracy.

2.3 Subpopulation Method

We categorize subpopulation methods into four classes: auxiliary losses, data augmentations, modeling objectives, and data sampling techniques. Auxiliary loss methods (Li et al. 2018; Arjovsky et al. 2019; Alshammari et al. 2022) aim to mitigate the impact of spurious backgrounds via adversarial training, gradient regularization, or class-balanced adjustments. Data augmentation methods (Zhang et al. 2018; Han et al. 2022; Yao et al. 2022) use convex combinations of samples to reduce background effects. Data sampling methods (LaBonte, Muthukumar, and Kumar 2024; Izmailov et al. 2022) identify class-balanced subsets with independent spurious features for finetuning. Modeling objective methods (Sagawa et al. 2020; Wu et al. 2023; Lin et al. 2017; Rudner et al. 2024) focus on robust feature learning, subpopulation correction, or tailored loss terms like KL divergence or mutual information.

DBA stands out by explicitly stating data assumptions and connecting existing methods under a unified statistical framework (see Sec. 4). For instance, it highlights that augmentation methods (Zhang et al. 2018; Yao et al. 2022; Han et al. 2022) assume conditional similarity across subpopulations. DBA also identifies a universal assumption of identical conditional generative models across methods, which previous works did not explicitly address. Empirically, DBA outperforms SOTA methods on three datasets, confirming its effectiveness and simplicity, and leveraging importance sampling for practical implementation.

3 Method

In this section, we first describe the framework, followed by the introduction of the proposed methods.

3.1 Dataset Bias Analysis Framework

Throughout the paper, we consider the following notations: $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ indicate the random variables for the data and labels, respectively. $\mathcal{X}$ and $\mathcal{Y}$ refer to their corresponding spaces. We denote y as a discrete random variable. We use $p(\cdot)$ to denote the probability distribution and $q(\cdot)$ or $\hat{p}(\cdot)$ to represent the estimates. Subscripts "tr", "va", and "te" indicate concepts associated with train, validation, and test datasets, respectively. We use $\mathcal{D}$ to refer to the datasets. We let $\mathcal{M}_{\text{tr}} := \{q(\cdot)|q(\cdot)$ estimated with data in $\mathcal{D}_{\text{tr}}\}$ denote the model spaces for the general learning problem. s denotes the attributes/spurious variables that are present in the datasets. This is also the root of the subpopulation. And I refers to the dataset indicator, which is the abstract variable that has no real values (i.e. I_{tr}, I_{va}, and I_{te}). We use $\text{Supp}(\cdot)$ to indicate the support set. We also use the notation "$\sim$" on two datasets (e.g. $\mathcal{D}_{\text{tr}} \sim \mathcal{D}_{\text{te}}$) to represent the same data distributions for the given datasets.

The DBA framework is formulated by initially asking the question: *Which model do we pick after training?* Conventional approaches consider ERM over $\mathcal{D}_{\text{tr}}$, stop the training, and choose the model with the lowest loss value on $\mathcal{D}_{\text{va}}$. Usually, the losses are implicitly assumed to be identical across $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{va}}$, and $\mathcal{D}_{\text{te}}$. There are two drawbacks to this inattentive

assumption. First, it does not properly characterize the difference across different datasets. Second, it does not naturally take into account how people make choices on the model. As a remedy, we propose the following objective (Eq. (1)) as the foundation for the DBA framework:

$$\mathbb{E}_{(x,y)\sim p(x,y|I_{\text{va}})}[\log q(y|x, I_{\text{tr}})]. \tag{1}$$

The maximization of the objective (Eq. (2)) hence provides an intuitive view of how people choose the final model after the optimization:

$$\max_{q\in\mathcal{M}_{\text{tr}}} \mathbb{E}_{(x,y)\sim p(x,y|I_{\text{va}})}[\log q(y|x, I_{\text{tr}})]. \tag{2}$$

In this paper, we consistently focus on the predictive modeling setup (i.e. $y|x$), which is aligned with existing works. Intuitively, Eq. (2) describes the scenario where we find the best conditional predictive model q according to the highest log likelihood measured over $\mathcal{D}_{\text{va}}$. Eq. (2) differs from ERM by explicitly considering the inherent difference between different datasets. In most cases, we seek for models to perform well on the unseen $\mathcal{D}_{\text{te}}$. To characterize this, we apply a similar logic as in Eq. (1) and focus on measuring the difference between validation and test sets. We make the following universal assumption 1.

Assumption 1. *The supports of x, y on $\mathcal{D}_{tr}$, $\mathcal{D}_{va}$, and $\mathcal{D}_{te}$ follow the relationship:*
$\text{Supp}_{tr}(x, y) \supset \text{Supp}_{va}(x, y)$, $\text{Supp}_{tr}(x, y) \supset \text{Supp}_{te}(x, y)$, *and* $\text{Supp}_{va}(x, y) \supset \text{Supp}_{te}(x, y)$.

The inclusion relationship described in the Assumption 1 essentially ensures a well-defined weight function (i.e., the denominator of the weight function is not zero) in the importance sampling setup in the proposed DBA framework. With this assumption, we make the following claim on the performance of the picked model (from Eq. (2)) with $\mathcal{D}_{\text{te}}$: *How does the picked model perform on the test set?*

Claim 1. *Given Assumption 1 holds and let q^* denote the best model obtained from Eq.* (2)*. The likelihood evaluated with the test set $\mathcal{D}_{te}$ for the model q^* can be viewed as the importance sampling version of $\mathbb{E}_{(x,y)\sim p(x,y|I_{va})}[z(x, y, I_{va}, I_{te}) \log q^*(y|x, I_{tr})]$ over the validation set with the function $z(\cdot)$ defined below:*

$$z(x, y, I_{va}, I_{te}) := \frac{p(x, y|I_{te})}{p(x, y|I_{va})}. \tag{3}$$

We defer this and all the following proof details in Appendix. Claim 1 informs that the only way to guarantee the best testing performance for the picked model q^* is to have access to the distribution $p(x, y|I_{\text{te}})$. This points out a hidden pitfall that commonly exists, yet overlooked, in the current machine learning optimizations with ERM—people choose a model with the best validation performance and report the corresponding testing performance. By Claim 1, we know that this general setup is true only in the case where $p(x, y|I_{\text{va}}) = p(x, y|I_{\text{te}})$. Otherwise, one needs to provide an accurate estimation on $z(x, y, I_{\text{va}}, I_{\text{te}})$ and pick the training model via a weighted likelihood, $\mathbb{E}_{(x,y)\sim p(x,y|I_{\text{va}})}[z(x, y, I_{\text{va}}, I_{\text{te}}) \log q^*(y|x, I_{\text{tr}})]$, on the validation set, to achieve optimal performance on the test set.

Simply put, Eq. (2) describes the way people pick the model during optimization, and Claim 1 points out the correct picking criterion for maximum test set performance. A natural follow-up question on these two arguments is: *Can we combine the notion of training and picking, and directly optimize q to maximize the testing performance?* The answer is affirmative under some additional assumptions. To explain, we first claim an optimization equivalence, providing the general form with which the optimization on the training set is identical to the optimization on the testing set (Claim 2). Then we derive another objective in the setup where we obtain a closed-form $g(x, y, I_{\text{tr}}, I_{\text{te}})$ (see Claim 2) after making several assumptions on the structure of the test data (Theorem 1).

Claim 2. *Given Assumption 1 holds we obtain the following equality on the objective:*

$$\begin{aligned}&\mathbb{E}_{(x,y)\sim p(x,y|I_{te})}[\log q(y|x, I_{tr})]\\ =&\mathbb{E}_{(x,y)\sim p(x,y|I_{tr})}[g(x, y, I_{tr}, I_{te}) \log q(y|x, I_{tr})],\end{aligned} \tag{4}$$

where the weight function $g(x, y, I_{tr}, I_{te}) := \frac{p(x,y|I_{te})}{p(x,y|I_{tr})}$.

The proof is similar to Claim 1 and can be found in Appendix. Note that in the language of importance sampling, the weight function $g(x, y, I_{\text{tr}}, I_{\text{te}})$ consists of the proposal distribution $p(x, y|I_{\text{tr}})$ and the data distribution $p(x, y|I_{\text{te}})$ in our setup. Compared to Eq. (1), Eq. (4) offers an objective that can be optimized with $\mathcal{D}_{\text{tr}}$ as the expectation is taken over the training set—the same space defined for models $q \in \mathcal{M}_{\text{tr}}$. Claim 2 also confirms that one must know $p(x, y|I_{\text{te}})$ to improve the testing performance of q when optimizing a model.

In this paper, we consider a uniform attribute setup that assumes the uniform distribution on the attribute/spurious variable $s \in \mathcal{S}$, which is a discrete random variable and $\text{Supp}(s) = \text{Supp}(y)$. s represents the cause of the subpopulation in our study. Formally speaking, this paper considers the following subpopulation shift:

Definition 1. The subpopulation shift is defined as the distributional difference between $p(x, y|I_{\text{tr}})$ and $p(x, y|I_{\text{te}})$ that is introduced by the spurious variable s w.r.t. the response y. Namely, $p(s, y|I_{\text{tr}}) \neq p(s, y|I_{\text{te}})$.

Specifically, we decompose the joint distribution of x and y through $\sum_s p(x, y, s|I_{\text{tr}}) = \sum_s p(x|y, s, I_{\text{tr}})p(y, s|I_{\text{tr}})$, and $\sum_s p(x, y, s|I_{\text{te}}) = \sum_s p(x|y, s, I_{\text{te}})p(y, s|I_{\text{te}})$. And the difference between datasets is on $p(s, y|I_{\text{tr}}) \neq p(s, y|I_{\text{te}})$. In the following, we describe several assumptions that lead to the major result of the paper—Theorem 1:

Assumption 2. *A universal data generator given the dataset information I, the label y, and the attribute s for the training and test sets: $p(x|y, s, I_{tr}) = p(x|y, s, I_{te})$.*

Assumption 3. *The attribute variable s follows a uniform distribution, conditional on y and I_{te}: $p(s|y, I_{te}) = 1/L$, where L is the number of outcomes for the discrete random variable s.*

Assumption 2 requires identical generative processes for x across training and testing. This can be seen as a specific type of covariate shift, attributing shifts in $p(x, y)$ to variations in $p(y, s)$ given the attribute s, rather than $p(x)$. Such

an assumption is common in conformal analysis and causal inference (Yang, Kuchibhotla, and Tchetgen Tchetgen 2024; Suter et al. 2019; Lei and Candès 2021). Assumption 3 imposes a weaker assumption compared to the literature, where uniformity and independence are generally assumed for both y and s (Tsirigotis et al. 2024). Compared to the existing work, we only assume s to follow a uniform distribution and there is no constraint on the distribution of y. The latter makes this approach applicable to class-imbalanced test data.

We further make two additional assumptions (Assumption 4 and 5) that reflect the nature of the considered subpopulation problems. This starts with studying the composition of the shifted datasets. Specifically, we introduce a random variable m that explicitly describes the substructure of the given data (i.e. $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{va}}$, and $\mathcal{D}_{\text{te}}$). Most existing works only consider the attribute variable s and its relation to labels y and data x. However, we realize that simply introducing this attribute is not enough to quantify the subpopulation as different subpopulations may have distinct relationships between s, y, and x. Therefore, the presence of m enables the quantification of such differences, making the proposed framework more flexible.

In particular, we consider m to be a binary random variable that takes values m_0 or m_1. And m_0 refers to the conceptual minority group in $\mathcal{D}_{\text{tr}}$ that shares the same statistics for s, y, and x in $\mathcal{D}_{\text{te}}$, whereas m_1 denotes the majority group that has distinct statistics of s, y, and possibly x—this explicitly characterizes the prevalent subpopulation in $\mathcal{D}_{\text{tr}}$ that causes the underperformance in $\mathcal{D}_{\text{te}}$. One may question the soundness of why we claim it is possible to find such a minority group in $\mathcal{D}_{\text{tr}}$. An intuitive, yet not strict, proof is to consider the established Assumption 1 that constrains inclusive supports across datasets. With Assumption 1, we can always find a subset of $\mathcal{D}_{\text{tr}}$ whose data statistics are close to that of $\mathcal{D}_{\text{te}}$ for any possibly large enough datasets. This leads to the following assumption:

Assumption 4. $p(y|I_{te}) = p(y|m_0, I_{tr}) = p(y|I_{tr})$.

Assumption 4 describes the scenario where there is no subpupolation on y between $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$. This assumption indicates that the subpopulation is introduced by the association between s and y, or x and y, but not solely by y itself. Since the minority group m_0 shares same data statistics as y, it is natural to have the equality $p(y|I_{\text{te}}) = p(y|m_0, I_{\text{tr}})$. It is noteworthy that there is no constraint on the number of groups specified by m. The size of 2 is considered in this paper due to its simplicity and high performance in practice (see Sec. 5).

Assumption 5, on the other hand, quantifies explicitly that there is a portion (i.e. m_1) of samples in $\mathcal{D}_{\text{tr}}$ whose attributes s are identical to the labels y. Rather than treating it as an assumption, it is more of a characterization on the subpopulation that widely presents in the real-world data (e.g., Waterbirds and ColorMNIST, or others described in (Yang et al. 2023)), where attributes strongly mislead the model prediction by such correlation.

Assumption 5. $p(s|y, m_1, I_{tr}) = \mathbf{1}_{\{y=s\}}$, *where* $\mathbf{1}_{\{y=s\}}$ *is the indicator function.*

With all ingredients, we propose the following theorem on the modeling objective:

Theorem 1. *Given Assumption 1, 2, 3, 4, and 5 hold, the optimization of Eq.* (4) *with the following weight function* $g(x, y, I_{tr}, I_{te})$ *directly maximizes the testing performance:*

$$g(x, y, I_{tr}, I_{te})^{-1} := p(m_0|I_{tr}) + \frac{p(m_1|I_{tr}) \cdot \frac{L}{p(y|I_{tr})} \cdot p(y|m_1, I_{tr})}{1 + \left[\frac{p(m_0|I_{tr}) \cdot p(y|I_{tr})/L + p(y|m_1, I_{tr})}{p(m_0|I_{tr}) \cdot p(y|I_{tr})/L}\right] \cdot \frac{1 - p(s=y|y,x,I_{tr})}{p(s=y|y,x,I_{tr})}}, \tag{5}$$

where $p(y|m_1, I_{tr}) = \frac{p(y|I_{tr}) - p(m_0|I_{tr}) \cdot p(y|I_{tr})}{p(m_1|I_{tr})}$. $p(m_0|I_{tr})$ *and* $p(m_1|I_{tr})$ *represent the probability of a binary random variable* m *taking the value* m_0 *or* m_1*, respectively. Namely, the random variable* m *denotes the split of* $\mathcal{D}_{tr}$ *into the majority and minority groups.*

The corresponding proof can be found in Appendix. With this formulation, $p(s = y|y, x, I_{\text{tr}})$ is the only unknown term to be estimated. Theorem 1 provides a closed form objective with which models trained with $\mathcal{D}_{\text{tr}}$ perform optimally on $\mathcal{D}_{\text{te}}$. In the following, we consider 3 different setups on the accessibility of s and the relationship between $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{va}}$. In each setup, we provide a method to estimate Eq. (5). We further showcase the performance of the proposed methods in the experiment section (Sec. 5). In Appendix, we include a discussion on the limitations of this approach concerning the restriction and possible relaxation of the assumptions.

3.2 Dataset Bias Correction Method

In this section, we provide 3 different approaches to estimate $Eq.$ (5). We summarize these approaches with a general name: dataset bias correction method (DBCM). The 3 approaches essentially provide different ways of estimating the only missing term $p(s|y, x, I_{\text{tr}})$ in Eq. (5). Once the term is estimated, we employ a universal algorithm (see Algorithm 1) to train the model with $\mathcal{D}_{\text{tr}}$. To facilitate the use of this approach in more real-world applications, we describe the scenarios where the three following approaches can be applied in Appendix.

Attribute s is Known When we have access to the attribute s, we can make a direct estimation on the only unknown term $p(s = y|y, x, I_{\text{tr}})$ using the data $(x, y, s) \in \mathcal{D}_{\text{tr}}$ and apply Algorithm 1 therein. As $p(s = y|y, x, I_{\text{tr}})$ increases, the weight function g decreases (see Eq. (5)), because stronger spurious correlations make $p(s = y|y, x, I_{\text{tr}})$ larger. Down-weighting these samples during training helps performance by reducing reliance on spurious correlations.

Attribute s is Unknown and $\mathcal{D}_{\text{tr}} \sim \mathcal{D}_{\text{va}}$ When we do not have access to the attribute s and $\mathcal{D}_{\text{tr}} \sim \mathcal{D}_{\text{va}}$, we propose to use the following term to estimate $p(s|y, x, I_{\text{tr}})$:

$$\hat{p}(s = y|y, x, I_{\text{tr}}) \propto \exp\left(\frac{|\log \hat{p}(y|x, I_{\text{tr}}) - \log \hat{p}(y|x, I_{\text{va}})|}{\tau}\right)^{-1}, \tag{6}$$

where $\hat{p}(y|x, I_{\text{tr}})$ and $\hat{p}(y|x, I_{\text{va}})$ are the predictive models learned with $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{va}}$, respectively. And τ is the temperature hyperparameter. In practice, we find $\tau = 1$ consistently produces good results. We explicitly introduce τ to allow flexibility in the control of the estimation in Eq. (6). Specifically, we first overfit two independent predictive models on both $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{va}}$ and then measure the difference on the two approximate laws with the training data. Note that $p(s = y|y, x, I_{\text{tr}})$ captures how likely s shares the same label as y, which is the only unknown term evaluated in Eq. (5). Therefore, we do not need to recover the full distribution $p(s|y, x, I_{\text{tr}})$. Instead, we only need to quantify $\hat{p}(s = y|y, x, I_{\text{tr}})$—"how likely the bias is biased towards the true label y." This is captured by Eq. (6), as if two models (trained separately on training and validation data) produce similar likelihoods (i.e. the difference in Eq. (6) is smaller) on a given input, then the input must associate with the attribute s that is same as y. To summarize this approach in one line: *two overfitted models act as a bias corrector!*

Attribute s is Unknown and $\mathcal{D}_{\text{tr}} \nsim \mathcal{D}_{\text{va}}$ On the other hand, when $\mathcal{D}_{\text{tr}} \nsim \mathcal{D}_{\text{va}}$, we cannot utilize the predictive model estimated with $\mathcal{D}_{\text{va}}$. Instead, we propose to use the following term as an alternative,

$$\hat{p}(s = y|y, x, I_{\text{tr}}) \propto \exp\left(-\frac{\log \hat{p}(y|x, I_{\text{tr}})}{\tau}\right)^{-1}. \quad (7)$$

This is according to the observation that machine learning models tend to learn the correlated attributes s with y easily (Asgari et al. 2022). In our case, we simply use $\hat{p}(y|x, I_{\text{tr}})$ as the proxy to characterize such correlation. In this case, samples with high accuracy should be down-weighted, as the model easily learns spurious correlations.

3.3 Choose Models

Similarly, we discuss different approaches for choosing a model. Unlike conventional methods that consistently use $\mathcal{D}_{\text{va}}$ to decide which model to choose, we propose to consider different ways for choosing a model when relationships between $\mathcal{D}_{\text{va}}$ and $\mathcal{D}_{\text{te}}$ are different. When $\mathcal{D}_{\text{va}} \sim \mathcal{D}_{\text{te}}$, according to Eq. (3), $z(x, y, I_{\text{va}}, I_{\text{te}}) = 1$. This indicates that evaluating models on validation set is equivalent to evaluating on the test set, which corresponds to the conventional approach. However, things change when $\mathcal{D}_{\text{va}} \nsim \mathcal{D}_{\text{te}}$. This suggests that $\mathcal{D}_{\text{va}}$ is not sufficient in measuring the model performance for the test set as $z(x, y, I_{\text{va}}, I_{\text{te}}) \neq 1$. In this case, we can adopt the similar approach outlined in Sec. 3.2 to estimate $z(x, y, I_{\text{va}}, I_{\text{te}})$, which focuses on $\mathcal{D}_{\text{va}}$ and $\mathcal{D}_{\text{te}}$, rather than $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$.

4 DBA Interpretation on Existing Work

In this section, we showcase how some representative existing works can be related to the DBA framework. Such discussion should complement the existing survey papers on subpopulation/distributional shifts and provide insights on the methodological development in the future. We follow the previously introduced categorization.

Algorithm 1 The universal algorithm for optimizing $q(y|x, I_{\text{tr}})$.

Input The initialized model $q(y|x, I_{\text{tr}})$; dataset $\mathcal{D}_{\text{tr}}$; The estimation $\hat{p}(s|y, x, I_{\text{tr}})$.
Output: the optimized $q(y|x, I_{\text{tr}})$.

1: Obtain $\hat{g}(x, y, I_{\text{tr}}, I_{\text{te}})$ given $\hat{p}(s = y|y, x, I_{\text{tr}})$ (see Eq. (5)).
2: Perform the following optimization using $\mathcal{D}_{\text{tr}}$:

$$\max_{q \in \mathcal{M}_{\text{tr}}} \mathbb{E}_{(x,y)\sim p(x,y|I_{\text{tr}})}[\hat{g}(x, y, I_{\text{tr}}, I_{\text{te}}) \log q(y|x, I_{\text{tr}})]. \quad (8)$$

`The model objective class`: Liu et al. (2021) and Nam et al. (2020) can be viewed as proposing different forms of the $p(s|y, x, I_{\text{tr}})$ estimation, where the former utilizes the classification accuracy and the latter considers generalized cross-entropy. Sec. 3.2 and 3.2 provide rationale on the validity of these terms—essentially they characterize the probability $p(s = y|y, x, I_{\text{tr}})$. Besides the variants of $\hat{p}(s = y|y, x, I_{\text{tr}})$, they propose different training schemes to correct. Liu et al. (2021) subsamples the training set with their $\hat{p}(s = y|y, x, I_{\text{tr}})$ and Nam et al. (2020) proposes a parallel model to reweight samples according to $\hat{p}(s = y|y, x, I_{\text{tr}})$ from the generalized cross entropy. Nonetheless, none of them is alike DBCM, which is statistically consistent in directly improving the testing performance.

`The data sampling class`: The methods in the data sampling class share great similarity to ours, as the proposed DBCM is essentially an importance sampling (reweighing) mechanism. ReWeight and ReSample (Japkowicz 2000) can be treated as variants of the sampling technique. Precisely, ReWeight adjusts each sample weight according to the class ratio, in order to recover the class-balanced setup. Similarly, ReSample bootstraps the dataset with class-balanced weights. Essentially, they can be treated as the direct estimation of $g(x, y, I_{\text{tr}}, I_{\text{te}}) := \frac{p(x,y|I_{\text{te}})}{p(x,y|I_{\text{tr}})}$, assuming $p(y|I_{\text{tr}})$ is uniform. When considering the presence of attribute s, $g(x, y, I_{\text{tr}}, I_{\text{te}})$ becomes,

$$g(x, y, I_{\text{tr}}, I_{\text{te}}) := \frac{p(x, y|I_{\text{te}})}{p(x, y|I_{\text{tr}})} = \frac{\sum_s p(x|y, s, I_{\text{te}})p(y, s|I_{\text{te}})}{\sum_s p(x|y, s, I_{\text{tr}})p(y, s|I_{\text{tr}})}. \quad (9)$$

Their setups, in this case, further assume $p(y, s|I_{\text{te}})$ is uniform and $p(y, s|I_{\text{tr}}) = p(y, s|I_{\text{te}})$, which is a stronger assumption compared to the proposed.

`The auxiliary loss class`: Tsirigotis et al. (2024) and Menon et al. (2021) are commonly used logit adjustment methods. With the DBA framework, they can be viewed as a two-step method. First, both methods propose an estimation of $p(y, s = y|, x, I_{\text{tr}})$. Then the estimates are used as a penalty term to regularize the ERM of the predictive model $q(y|x, I_{\text{tr}})$. In the first step, Tsirigotis et al. (2024) applies a similar approach to one described in (Liu et al. 2021). Both share conceptual similarity to the DBCM variant in Sec. 3.2. Menon et al. (2021), on the other hand, simply enforces the uniform class balance assumption. Once

$\hat{p}(y, s = y|x, I_{\text{tr}})$ is obtained, they optimize w.r.t.

$$\mathbb{E}_{(x,y)\sim p(x,y|I_{\text{tr}})}[\log q(y|x, I_{\text{tr}}) + \log \hat{p}(y, s = y|x, I_{\text{tr}})]. \tag{10}$$

To compare the difference between Eq. (10) and the optimal objective (Eq. (4)), we prove the following theorem with two additional assumptions on the label y.

Assumption 6. *The label y given I_{tr} follows a uniform distribution.*

Assumption 7. *The training set contains only the dominant group m_1: $p(m_1|I_{tr}) = 1$.*

Theorem 2. *Given Assumption 1, 2, 3 6, and 7 hold, the optimization of Eq.* (4) *with the following weight function $g(x, y, I_{tr}, I_{te})$ directly maximizes the testing performance:*

$$g(x, y, I_{tr}, I_{te})^{-1} := L \cdot p(y, s = y|x, I_{tr}). \tag{11}$$

And the objective Eq. (4) *is of form:*

$$\begin{aligned}\mathbb{E}_{(x,y)\sim p(x,y|I_{tr})}\big[&g(x, y, I_{tr}, I_{te})\big(\log q(y|x, I_{tr}) \\ &+ \log p(y, s = y|x, I_{tr})\big) \\ &+ g(x, y, I_{tr}, I_{te}) \log L \cdot g(x, y, I_{tr}, I_{te})\big]. \end{aligned} \tag{12}$$

The proof is in Appendix. Eq. (10) differs Eq. (12) by 2 aspects. First, Eq. (10) ignores the weight function $g(x, y, I_{\text{tr}}, I_{\text{te}})$ before the summation. Second, the regularization $g(x, y, I_{\text{tr}}, I_{\text{te}}) \log L \cdot g(x, y, I_{\text{tr}}, I_{\text{te}})$ in Eq. (12) is missing. Without these terms, Eq. (12) is not guaranteed to optimize for a class-balance dataset, as indicated in (Tsirigotis et al. 2024; Menon et al. 2021). Consequently, these methods may underperform.

`The augmentation class`: Despite existing works provide augmentation techniques in the form of linear combination (Zhang et al. 2018; Yao et al. 2022; Han et al. 2022), none of the papers provides statistical interpretation on why such techniques work better than ERM. We see our DBA framework as the first to provide support for the soundness of the augmentation technique. In short, the augmentation to combine data samples can be viewed as variations of the direct recovery of $g(x, y, I_{\text{tr}}, I_{\text{te}}) := \frac{p(x,y|I_{\text{te}})}{p(x,y|I_{\text{tr}})}$ under a different set of assumptions. Specifically, we provide the following Theorem 3 to support this statement. The proof can be found in Appendix. In the following theorem, m_0 and m_1 are identical to the terms introduced in Theorem 1. We first describe the assumptions.

Assumption 8. *The data generator of $\mathcal{D}_{tr}$ are conditionally identical given different group information m: $p(x|m_0, I_{tr}) = p(x|m_1, I_{tr}) = p(x|I_{tr})$.*

Assumption 9. *The predictive model on $\mathcal{D}_{te}$ shares the same law with the model that is conditioned on the group m_0 for $\mathcal{D}_{tr}$: $p(y|x, I_{te}) = p(y|x, m_0, I_{tr})$.*

Note that Assumption 9 is conceptually similar to the setup for Theorem 1.

Theorem 3. *Given Assumption 1, 8, and 9 hold, the weight function $g(x, y, I_{tr}, I_{te})$ has the following form:*

$$\begin{aligned} g(x, y, I_{tr}, I_{te})^{-1} :=& \lambda_0(x, I_{tr}, I_{te}) \cdot p(x|I_{tr}) \\ &+ \lambda_1(x, y, I_{tr}, I_{te}) \cdot p(x|I_{tr}). \end{aligned} \tag{13}$$

$\lambda_0(x, I_{\text{tr}}, I_{\text{te}}) := \frac{p(m_0|I_{\text{tr}})}{p(x|I_{\text{te}})}$ and $\lambda_1(x, y, I_{\text{tr}}, I_{\text{te}}) := \frac{p(m_1|I_{\text{tr}})p(y|x,m_1,I_{\text{tr}})}{p(y|x,I_{\text{te}})p(x|I_{\text{te}})}$. This means the weight function $g(x, y, I_{\text{tr}}, I_{\text{te}})$ is a reweighing of the original $p(x|I_{\text{tr}})$. The commonly used augmentation can be viewed as a sample-level adjustment to the weight function. From Theorem 3 we know that the sum of λ_0 and λ_1 need not be 1, which is different from some existing augmentation approaches (Zhang et al. 2018; Yao et al. 2022) [1]. The theorem also offers statistical rationale on why the weighted linear combination works (Han et al. 2022). Since both λ_0 and λ_1 depend on the data statistics from the testing set, methods that utilize sample-independent coefficient (Zhang et al. 2018; Yao et al. 2022) should experience degraded performance. We believe this provides insights into the advancement of augmentation-based techniques in the future.

5 Experiment

We compare different DBCM variants (see Sec. 3.2) benchmarking models with three benchmarking datasets. We showcase the SOTA performance of our models, to demonstrate the consistency of the theory developed in Sec. 3. In addition, we provide experimental evidence that complements the theory on explaining why existing works would sacrifice average accuracy for higher worst group accuracy.

5.1 Experimental Setup

To ensure a fair comparison, we consider models and datasets prepared by Yang et al. (2023). Specifically, we consider two vision datasets: `Waterbirds` (Sagawa et al. 2020) and `ColorMNIST` (Nam et al. 2020; Tsirigotis et al. 2024), and one language dataset: `CivilComments` (Borkan et al. 2019), in order to cover the two popular data types. We modify the `ColorMNIST` dataset such that it aligns with the setup in Tsirigotis et al. (2024), which is a harder setup. This is because the vanilla version in Yang et al. (2023) consists of only two types of attributes, whereas the version in Nam et al. (2020); Tsirigotis et al. (2024) contains 10 attributes. The modified `ColorMNIST` contains a "ratio" indicator that specifies the portion of samples that do not correlate with labels and attributes. In our experiment, we consider ratios 2% and 0.5%, as they are the intermediate and the hardest setups. In practice, we also treat $p(m_1|I_{\text{tr}})$ and $p(m_0|I_{\text{tr}})$ serve as prior knowledge/hyperparameters of training composition. Specifically for `ColorMNIST`, where spurious sample ratio is known, we directly assign 0.5% or 2% for $p(m_0|I_{\text{tr}})$ (i.e., $1 - p(m_1|I_{\text{tr}})$). When the composition ratio is unknown, $p(m_0|I_{\text{tr}})$ is treated as a hyperparameter and empirically we identify $p(m_0|I_{\text{tr}}) = 0.85$ performed well across datasets.

The evaluation consists of 8 benchmarking models from Yang et al. (2023) that fall into the 4 different classes (see Sec. 1 and 4): Mixup (Zhang et al. 2018); LISA (Yao et al. 2022); JTT (Liu et al. 2021); Focal Loss (Lin et al.

[1] We consider the reformulation of the objectives in Zhang et al. (2018); Yao et al. (2022) according to Han et al. (2022), where the objective with the mixup random variable $\tilde{y}$ can be transformed to the mixup of two weighted terms with the random variable y. See Sec. 3.2 of (Han et al. 2022) for details.

	ColorMNIST(0.5%)		ColorMNIST(2%)		Waterbirds		CivilComments		Availability of s
	average	worst	average	worst	average	worst	average	worst	
ERM	81.69 ± 0.10	1.14 ± 0.40	95.23 ± 0.07	56.82 ± 0.23	88.25 ± 0.16	67.76 ± 0.30	87.59 ± 0.38	48.17 ± 2.61	s Known
Mixup (Zhang et al. 2018)	81.12 ± 2.20	0.00 ± 0.00	96.09 ± 0.20	80.00 ± 2.22	88.52 ± 0.22	59.97 ± 2.01	87.67 ± 0.12	53.10 ± 2.11	
LISA (Yao et al. 2022)	89.45 ± 1.57	21.50 ± 8.51	97.32 ± 0.37	87.27 ± 6.55	93.63 ± 0.66	76.95 ± 4.25	87.22 ± 0.13	40.62 ± 4.32	
JTT (Liu et al. 2021)	81.98 ± 1.17	2.00 ± 0.08	95.03 ± 0.10	56.82 ± 2.21	88.32 ± 0.20	68.80 ± 2.99	87.78 ± 0.29	47.06 ± 2.94	
Focal Loss (Lin et al. 2017)	67.37 ± 0.44	0.00 ± 0.00	94.62 ± 0.25	43.00 ± 2.33	87.75 ± 0.36	54.67 ± 2.67	87.74 ± 0.16	43.73 ± 3.66	
GroupDRO (Sagawa et al. 2020)	82.88 ± 0.09	9.00 ± 0.08	95.19 ± 1.01	40.91 ± 1.20	92.03 ± 0.16	$\mathbf{83.64} \pm 1.88$	86.78 ± 0.18	56.51 ± 1.93	
MMD (Li et al. 2018)	11.35 ± 1.30	0.00 ± 0.00	11.35 ± 2.26	0.00 ± 0.00	88.33 ± 0.51	53.58 ± 2.38	82.08 ± 0.63	0.00 ± 0.00	
ReSample (Japkowicz 2000)	94.95 ± 0.19	66.37 ± 2.33	98.34 ± 0.23	92.00 ± 1.66	93.72 ± 0.22	80.69 ± 1.86	84.59 ± 1.23	$\mathbf{62.17} \pm 1.72$	
ReWeight (Japkowicz 2000)	92.43 ± 0.21	57.84 ± 1.78	97.83 ± 0.19	91.46 ± 1.80	93.86 ± 0.30	81.15 ± 2.20	87.04 ± 0.74	58.27 ± 2.14	
DBCM(Sec. 3.2, known s)	$\mathbf{96.67} \pm 0.27$	$\mathbf{84.62} \pm 2.02$	$\mathbf{98.76} \pm 0.20$	$\mathbf{92.31} \pm 1.73$	$\mathbf{94.01} \pm 0.19$	83.18 ± 2.00	$\mathbf{87.85} \pm 0.15$	43.33 ± 2.15	
ERM	81.69 ± 0.10	1.14 ± 0.40	95.23 ± 0.07	56.82 ± 0.23	88.25 ± 0.16	67.76 ± 0.30	87.59 ± 0.38	48.17 ± 2.61	s Unknown
Mixup (Zhang et al. 2018)	81.03 ± 2.30	0.00 ± 0.00	95.26 ± 0.17	42.05 ± 3.61	90.65 ± 0.30	67.29 ± 1.93	87.48 ± 0.11	54.84 ± 2.13	
LISA (Yao et al. 2022)	68.09 ± 2.06	0.00 ± 0.00	94.46 ± 0.53	15.91 ± 13.11	89.80 ± 1.11	66.82 ± 3.87	87.18 ± 0.28	49.21 ± 2.11	
JTT (Liu et al. 2021)	81.80 ± 0.19	2.00 ± 0.08	95.42 ± 0.02	48.86 ± 1.85	88.83 ± 0.28	66.36 ± 3.10	87.78 ± 3.84	47.06 ± 8.09	
Focal Loss (Lin et al. 2017)	67.12 ± 0.50	0.00 ± 0.00	94.53 ± 0.32	37.00 ± 4.10	89.92 ± 0.43	61.68 ± 3.01	87.74 ± 0.12	50.08 ± 4.10	
ReSample (Japkowicz 2000)	81.55 ± 0.21	0.00 ± 0.00	95.70 ± 0.15	65.00 ± 1.63	87.99 ± 0.12	64.17 ± 1.98	83.24 ± 1.73	$\mathbf{68.91} \pm 4.51$	
ReWeight (Japkowicz 2000)	76.77 ± 0.37	0.00 ± 0.00	94.93 ± 0.08	54.55 ± 0.10	87.81 ± 0.18	67.60 ± 1.67	87.02 ± 1.12	58.73 ± 4.60	
DBCM(Sec. 3.2, $\mathcal{D}_{\text{tr}} \sim \mathcal{D}_{\text{va}}$)	$\mathbf{94.63} \pm 0.35$	$\mathbf{57.95} \pm 2.30$	$\mathbf{97.64} \pm 0.10$	$\mathbf{81.00} \pm 1.40$	88.25 ± 0.05	$\mathbf{70.56} \pm 0.12$	$\mathbf{87.86} \pm 0.30$	43.41 ± 2.20	
DBCM(Sec. 3.2, $\mathcal{D}_{\text{tr}} \nsim \mathcal{D}_{\text{va}}$)	86.12 ± 0.29	3.41 ± 0.70	96.08 ± 0.20	61.36 ± 1.90	$\mathbf{91.04} \pm 0.07$	62.77 ± 0.10	87.62 ± 0.20	53.89 ± 2.16	

Table 1: Results on the three benchmarking datasets with both cases where s is known and unknown, respectively. We report both average and worst group accuracy, with mean and standard deviation ("$\pm$") for each of the considered methods after 3 independent runs. The **boldfaced** values indicate the highest accuracy in comparison.

2017); GroupDRO (Sagawa et al. 2020); MMD (Li et al. 2018); ReSample (Japkowicz 2000); ReWeight (Japkowicz 2000). For each model, we consider two setups, where the first allows the presence of attributes and the second does not. We retrain all the considered models from Yang et al. (2023) and pick the best models according to the average validation accuracy, which is different from the worst-group-accuracy criterion in Yang et al. (2023) to match the objective in Eq. 4 (i.e. the framework considers on average accuracy by design). For model optimization, we consider default optimizers and learning rates in Yang et al. (2023). Details are in Appendix.

Attribute s is Known This section presents results with accessible attribute s. In addition to the 8 benchmarking models, we also include results with ERM as the baseline. We consider the DBCM variant in Sec. 3.2. Results are summarized in the upper half of Table 1. It is clear that when the attribute s presents, the proposed DBCM(Sec. 3.2, known s) achieves the highest average accuracy among all the considered datasets. And all the accuracy of our model exceeds the ERM baseline. This provides the empirical evidence for Theorem 2 and 1. Although there is no theoretical quantification on the worst group accuracy, DBCM achieves two highest and one competing (i.e., Waterbirds) worst group accuracy.

Attribute s is Unknown This section presents results without accessing the attribute s. We omit results for GroupDRO (Sagawa et al. 2020), MMD (Li et al. 2018) as both methods naturally require the knowledge of s (Yang et al. 2023). DBCM(Sec. 3.2, $\mathcal{D}_{\text{tr}} \sim \mathcal{D}_{\text{va}}$) and DBCM(Sec. 3.2, $\mathcal{D}_{\text{tr}} \nsim \mathcal{D}_{\text{va}}$) are two variants of the proposed method. Results are summarized in the lower half of Table 1. We observe that the proposed DBCM variants achieve the highest average accuracy among all the compared datasets, and 3 out of 4 highest worst group accuracy, suggesting the validity of the methods when s is unknown. And in the case of the worst group accuracy for ColorMNIST(0.5%), almost all but DBCM cannot correctly classify the worst group samples (i.e. worst group accuracy = 0), suggesting that DBCM method is robust to the change of spurious association between the attributes and the labels.

5.2 Observation on the Degraded Average Accuracy

From Table 1 we empirically identify an interesting phenomenon—compared to all other methods, DBCM is the only model that consistently outperforms the results of ERM. This observation is aligned with Yang et al. (2023); Tsirigotis et al. (2024). Yet the previous work did not provide systematic reasoning on why. We argue that the cause is an incorrect model objective that is different from the data composition in $\mathcal{D}_{\text{te}}$. Specifically, the reduced average accuracy is the result of the misspecified $p(x, y|I_{\text{te}})$ in $g(x, y, I_{\text{tr}}, I_{\text{te}})$. For the full explanation, please refer to Appendix. It is noteworthy that we are the first to provide such a statistical interpretation of the degradation phenomenon.

6 Conclusion

In summary, we present the DBA framework to identify the true model objective that improves the test performance. The paper proposes different DBCM variants with weaker assumptions compared to the existing works and demonstrates the SOTA performance. Additionally, we reinterpret the existing work with the proposed framework, which explains the issue of the degraded average accuracy. With the analysis, we convey a message that to achieve decent test performance (even without the access to test during training), one must comprehensively investigate the relationship between those datasets and the model objective. For this purpose, we hope the proposed framework could act as a complementary tool to all the existing work, help people analyze such gaps, and facilitate the development of the corresponding model solutions.

Acknowledgments

This work was partially supported by Alfred P. Sloan foundation and National Science Foundation's grant #1934757. This work utilizes computational resources supported by the National Science Foundation's Major Research Instrumentation program, grant #1725729, as well as the University of Illinois at Urbana-Champaign.

References

Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6897–6907.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Asgari, S.; Khani, A.; Khani, F.; Gholami, A.; Tran, L.; Mahdavi Amiri, A.; and Hamarneh, G. 2022. Masktune: Mitigating spurious correlations by forcing to explore. *Advances in Neural Information Processing Systems*, 35: 23284–23296.

Bickel, S.; Brückner, M.; and Scheffer, T. 2007. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, 81–88.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, 491–500.

Fang, T.; Lu, N.; Niu, G.; and Sugiyama, M. 2020. Rethinking importance weighting for deep learning under distribution shift. *Advances in neural information processing systems*, 33: 11996–12007.

Han, Y.; and Zou, D. 2024. Improving Group Robustness on Spurious Correlation Requires Preciser Group Inference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.

Han, Z.; Liang, Z.; Yang, F.; Liu, L.; Li, L.; Bian, Y.; Zhao, P.; Wu, B.; Zhang, C.; and Yao, J. 2022. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35: 37704–37718.

Hong, F.; Yao, J.; Lyu, Y.; Zhou, Z.; Tsang, I.; Zhang, Y.; and Wang, Y. 2023. On Harmonizing Implicit Subpopulations. In *The Twelfth International Conference on Learning Representations*.

Huang, J.; Gretton, A.; Borgwardt, K.; Schölkopf, B.; and Smola, A. 2006. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19.

Izmailov, P.; Kirichenko, P.; Gruver, N.; and Wilson, A. G. 2022. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532.

Japkowicz, N. 2000. The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on artificial intelligence*, volume 56, 111–117.

Kanamori, T.; Hido, S.; and Sugiyama, M. 2009. A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10: 1391–1445.

LaBonte, T.; Muthukumar, V.; and Kumar, A. 2024. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36.

Lei, L.; and Candès, E. J. 2021. Conformal inference of counterfactuals and individual treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5): 911–938.

Li, H.; Pan, S. J.; Wang, S.; and Kot, A. C. 2018. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5400–5409.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 6781–6792. PMLR.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684.

Quionero-Candela, J.; Sugiyama, M.; Schwaighofer, A.; and Lawrence, N. D. 2009. *Dataset Shift in Machine Learning*. The MIT Press. ISBN 0262170051.

Rudner, T. G.; Zhang, Y. S.; Wilson, A. G.; and Kempe, J. 2024. Mind the GAP: Improving Robustness to Subpopulation Shifts with Group-Aware Priors. In *International Conference on Artificial Intelligence and Statistics*, 127–135. PMLR.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2): 227–244.

Suter, R.; Miladinovic, D.; Schölkopf, B.; and Bauer, S. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 6056–6065. PMLR.

Tsirigotis, C.; Monteiro, J.; Rodriguez, P.; Vazquez, D.; and Courville, A. C. 2024. Group Robust Classification Without Any Group Information. *Advances in Neural Information Processing Systems*, 36.

Wu, S.; Yuksekgonul, M.; Zhang, L.; and Zou, J. 2023. Discover and cure: Concept-aware mitigation of spurious correlation. In *International Conference on Machine Learning*, 37765–37786. PMLR.

Yang, Y.; Kuchibhotla, A. K.; and Tchetgen Tchetgen, E. 2024. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*.

Yang, Y.; Zhang, H.; Katabi, D.; and Ghassemi, M. 2023. Change is Hard: A Closer Look at Subpopulation Shift. In *International Conference on Machine Learning*.

Yao, H.; Wang, Y.; Li, S.; Zhang, L.; Liang, W.; Zou, J.; and Finn, C. 2022. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, 25407–25437. PMLR.

Yu, H.; Liu, J.; Zhang, X.; Wu, J.; and Cui, P. 2024. A Survey on Evaluation of Out-of-Distribution Generalization. *arXiv preprint arXiv:2403.01874*.

Zhang, H.; Cissé, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhang, S.; Luo, Y.; Wang, Q.; Chi, H.; Chen, X.; Han, B.; and Li, J. 2023. Mixture Data for Training Cannot Ensure Out-of-distribution Generalization. arXiv:arXiv:2312.16243.