Proceedings of Machine Learning Research 279:156–172, 2024 Algorithmic Fairness Through the Lens of Metrics and Evaluation

Fairness-Enhancing Data Augmentation Methods for Worst-Group Accuracy

Monica Welfert MWELFERT@ASU.EDU
Nathan Stromberg NSTROMBE@ASU.EDU
Lalitha Sankar LSANKAR@ASU.EDU
Arizona State University

Editors: Miriam Rateike, Awa Dieng, Jamelle Watson-Daniels, Ferdinando Fioretto, Golnoosh Farnadi

Abstract

Ensuring fair predictions across many distinct subpopulations in the training data can be prohibitive for large models. Recently, simple linear last layer retraining strategies, in combination with data augmentation methods such as upweighting and downsampling have been shown to achieve state-of-the-art performance for worst-group accuracy, which quantifies accuracy for the least prevalent subpopulation. For linear last layer retraining and the abovementioned augmentations, we present a comparison of the optimal worst-group accuracy when modeling the distribution of the latent representations (input to the last layer) as Gaussian for each subpopulation. Observing that these augmentation techniques rely heavily on well-labeled subpopulations, we present a comparison of the optimal worst-group accuracy in the setting of label noise. We verify our results for both synthetic and large publicly available datasets.

Keywords: last layer retraining, worst-group accuracy, upweighting, downsampling, label noise

1. Introduction

Last layer retraining (LLR) has emerged as a popular method for leveraging representations from large pretrained neural networks and fine-tuning them to locally available data. These methods are significantly inexpensive computationally relative to training the full model, and thus, allow transferring a model to new domains, predicting on *retraining data* with distributional shifts relative to the original, and optimizing for a different metric than that used by the original model.

In general, training data includes samples from different subpopulations (oft referred to as groups[1] which we take as a tuple of class and domain labels) (Yang et al., 2023). Assuring fair inferences across all subpopulations remains an important problem in modern machine learning. A metric which has been recently evaluated with good success for assuring fair decisions is worst-group accuracy (WGA), a worst-case metric for any prior across subpopulations. Existing methods which optimize for WGA utilize strongly regularized models along with *data augmentation* methods such as *downsampling* (DS) (Kirichenko et al., 2023; LaBonte et al., 2023) and *upweighting* (UW) (Liu et al., 2021; Qiu et al., 2023) (Section 2 presents precise definitions of these methods). These augmentation techniques

[1] we will use these terms interchangeably

help to account for varying proportions of individual groups and enable the final model to predict well on every group.

It is difficult to obtain theoretical performance guarantees for large models. However, for a fixed representation-extracting model, one can focus on evaluating LLR techniques that tune a linear last layer using (possibly augmented) representations from the pretrained model. We study this setting and model the representations of the subpopulations using tractable distributions; this allows us to directly compare different data augmentation techniques in terms of WGA.

To this end, analogous to Yao et al. (2022), we model individual sub-populations as distinct Gaussian distributions. Our primary contribution is a straightforward comparison of DS and UW, two of the most common data augmentation techniques for WGA. These methods require access to correctly labeled groups, which are often noisy in practice (Wei et al., 2022) due to domain label noise, class label noise, or both. Because class label noise generally affects classifier training in addition to data augmentation methods, we consider only domain label noise so that we can strictly compare the effect of these data augmentation techniques on WGA.

We consider two settings: (i) no domain label noise and (ii) domain label noise. Our key contributions are in providing:

- A distribution-free equivalence of the risk minimization problem, and thus the optimal models and performance, for upweighting and downsampling (Theorem 1). To the best of our knowledge, this is a new result.
- Statistical analysis of the WGA for each data augmentation method under Gaussian subpopulations in the settings of (i) no domain label noise (Theorem 3) and (ii) domain label noise (Theorem 4).
- Empirical results that match theory for Gaussian mixtures and the CMNIST, CelebA and Waterbirds datasets.

Our work is distinct from that in Yao et al. (2022) as follows: (i) explicit incorporation of the minority group priors; (ii) providing precise WGA guarantees (in contrast to bounds in Yao et al. (2022)); (iii) including downsampling and upweighting as data augmentation methods (the focus in Yao et al. (2022) is primarily on mixing and its variants) for which we also provide comparative model and error guarantees beyond the Gaussian setting; and (iv) analyzing the effect of domain label noise on WGA.

2. Problem Setup

We consider the supervised classification setting and assume that the LLR methods have access to a representation of the input/*ambient* (original high-dimensional data such as images etc.) data, the ground-truth label, as well as the domain annotation. Taken together, the label and domain combine to define the group annotation for any sample. For ease of analysis, we assume binary labels (belonging to $\{0,1\}$) and binary domains (belonging to $\{S,T\}$). More formally, the training dataset is a collection of i.i.d. tuples of the random variables $(X_a, Y, D) \sim P_{X_a,Y,D}$, where $X_a \in \mathcal{X}_a$ is the ambient high-dimensional sample, $Y \in \mathcal{Y} = \{0,1\}$ is the class label, and $D \in \mathcal{D} = \{S,T\}$ is the domain label. Since the focus here is on learning the linear last layer, we denote the *latent* representation that acts as an input to this last layer by $X \coloneqq \phi(X_a)$ for an embedding function $\phi : \mathcal{X}_a \to \mathcal{X} \subseteq \mathbb{R}^m$ such that the training dataset for LLR is $(X, Y, D) \sim P_{X,Y,D}$.

The tuples (Y, D) of class and domain labels partition the examples into four different groups. Let $\pi^{(y,d)} \coloneqq P(Y = y, D = d)$ for $(y, d) \in \mathcal{Y} \times \mathcal{D}$. We denote the linear correction applied in the latent space of a pretrained model as $f_\theta : \mathcal{X} \to \mathbb{R}$, which is parameterized by a linear decision boundary $\theta = (w, b) \in \mathbb{R}^{m+1}$ given by $f_\theta(x) = w^T x + b$. The statistically optimal linear model is obtained by minimizing the risk defined as

$$R(f_\theta) \coloneqq \mathbb{E}_{P_{X,Y,D}}[\ell(f_\theta(X), Y)], \tag{1}$$

where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. We consider three different methods to learn a classifier: (a) statistical risk minimization (SRM) (also known as population risk minimization), (b) DS, and (c) UW. In particular, SRM involves minimizing (1) as is whereas DS involves reducing the size of each group to that of the smallest one and UW involves scaling the loss for each group in proportion to the inverse of the prior.

A general formulation for obtaining the optimal f_{θ^*} is:

$$\theta^* = \arg\min_\theta \mathbb{E}_{P_{X,Y,D}}[\ell(f_\theta(X), Y)c(Y, D)], \tag{2}$$

where $c(y, d) = 1$, $(y, d) \in \mathcal{Y} \times \mathcal{D}$, for SRM and DS, but $c(y, d) = 1/(4\pi^{(y,d)})$ for UW. Moreover, the priors on the groups remain the same as the true statistics, and therefore SRM, for all methods except DS where $\pi^{(y,d)} = 1/4$.

We desire a model that makes fair decisions across groups and therefore evaluate worst-group accuracy, i.e., the minimum accuracy among all groups, defined for a model f_θ as

$$\mathrm{WGA}(\theta) \coloneqq \min_{(y,d)\in\mathcal{Y}\times\mathcal{D}} A^{(y,d)}(\theta), \tag{3}$$

where $A^{(y,d)}(\theta)$ denotes the per-group classification accuracy for $(y, d) \in \mathcal{Y} \times \mathcal{D}$. Specifically, for $(y, d) \in \mathcal{Y} \times \mathcal{D}$:

$$A^{(y,d)}(\theta) \coloneqq P(\mathbb{1}\{f_\theta(X) > 1/2\} = Y | Y = y, D = d) \tag{4}$$

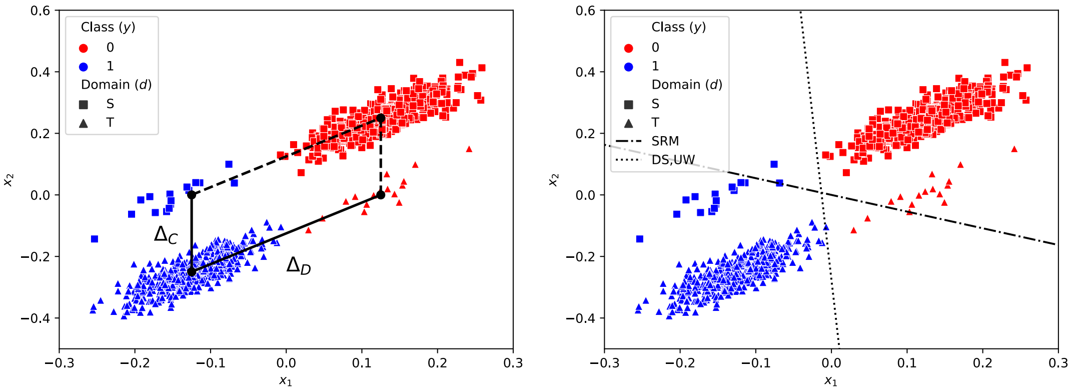
where the threshold $1/2$ is chosen to match $Y \in \{0, 1\}$.

Domain label noise We model noise in the domain label as symmetric label noise (SLN) with probability (w.p.) p. That is, for a sample $(X, Y, D) \sim P_{X,Y,D}$, we do not observe D directly but D w.p. $1 - p$ and $\bar{D}$ w.p. p where $\bar{D}$ is drawn uniformly at random from $\mathcal{D} \setminus \{D\}$. In the binary domain setting, this is equivalent to flipping D w.p. p.

3. Main Results

Our first result observes that, for any chosen loss, UW and DS yield the same statistically expected predictor. We collate the proofs in the Appendix and outline a proof sketch here.

Theorem 1 *For any given $P_{X,Y,D}$ and loss ℓ, the objectives in (2) when modified appropriately for DS and UW are the same. Therefore, if a minimizer exists for one of them, then the minimizer of the other is the same, i.e., $\theta^*_{DS} = \theta^*_{UW}$.*



(a) (b)

Figure 1: (a) Δ_C and Δ_D are shown as line segments between group means overlaid on data sampled from data satisfying Assumptions A1 to A4. (b) The optimal prediction planes for DS, UW, and SRM are shown overlaid on data sampled from Gaussian mixtures satisfying Assumptions A1 to A4. The SRM model largely ignores the minority group for each class.

Proof sketch. The key intuition here is that the upweighting factor is proportional to the inverse of the priors on each group. Thus, when the expected loss is decomposed into an expectation over groups, the priors from the expected loss cancel and we recover the downsampled problem. A detailed proof can be found in Appendix A.

While Theorem 1 holds for any general data distribution, to obtain more refined guarantees on WGA and model parameters for the different augmentation methods considered here, we make the following tractable assumptions on the dataset. Such assumptions have recently been introduced for tractability in the analysis of out-of-distribution robustness (e.g., Yao et al. (2022)).

Assumption A1 *$X \in \mathcal{X}$ is distributed according to the following mixture of Gaussians:*

$$X|(Y = y, D = d) \sim \mathcal{N}(\mu^{(y,d)}, \Sigma), \tag{5}$$

for $(y, d) \in \mathcal{Y} \times \mathcal{D}$, where $\mu^{(y,d)} := \mathbb{E}[X|Y = y, D = d] \in \mathbb{R}^m$ and $\Sigma \in \mathbb{R}^{m\times m}$ is symmetric positive definite. Additionally, we place priors $\pi^{(y,d)}$, $(y, d) \in \mathcal{Y} \times \mathcal{D}$, on each group and priors $\pi^{(y)} := P(Y = y)$, $y \in \mathcal{Y}$, on each class.

Assumption A2 *The minority groups have equal priors, i.e., for $\pi_0 \leq \frac{1}{4}$,*

$$\pi^{(0,T)} = \pi^{(1,S)} = \pi_0 \quad \text{and} \quad \pi^{(1,T)} = \pi^{(0,S)} = 1/2 - \pi_0.$$

Also, the class priors are equal, i.e., $\pi^{(0)} = \pi^{(1)} = 1/2$.

Assumption A3 *The difference in means between classes within a domain $\Delta_D := \mu^{(1,d)} - \mu^{(0,d)}$ is constant for $d \in \mathcal{D}$.*

Remark 2 *Assumption A3 also implies that the difference in means between domains within the same class $\Delta_C := \mu^{(y,S)} - \mu^{(y,T)}$ is also constant for each $y \in \mathcal{Y}$. We see this by noting that each group mean makes up the vertex of a parallelogram, as shown in Figure 1(a), where Δ_D and Δ_C are shown on samples drawn from a distribution satisfying Assumptions A1 to A3.*

While Theorem 1 clarifies the statistical behavior of DS and UW, comparing the resulting analytical expressions for the WGA of these two methods with those of SRM requires finer assumptions. To this end, we make the following orthogonality assumption.

Assumption A4 *Δ_D and Δ_C are orthogonal w.r.t. the Σ^{-1}–inner product, i.e., $\Delta_C^T \Sigma^{-1} \Delta_D = 0$.*

Theorem 3 *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$, $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$ (MSE loss). Under Assumptions A1 to A4, for any $\pi_0 \le 1/4$,*

$$WGA(\theta^*_{SRM}) \le WGA(\theta^*_{DS}) = WGA(\theta^*_{UW}),$$

with equality when $\pi_0 = 1/4$.

Proof sketch. The proof of equality for DS and UW follows directly from Theorem 1. To compare with SRM, we derive the optimal parameters in (2) for the given ℓ and appropriate values of $c(y, d)$, $(y, d) \in \{0, 1\} \times \{S, T\}$ for each method. We then use Assumption A1 to obtain the WGA in terms of Gaussian CDFs. We employ a derivative analysis to show $\text{WGA}(\theta^*_{\text{SRM}}) < \text{WGA}(\theta^*_{\text{DS}})$. A detailed proof is in Appendix B. See Figure 1(b) for a plot showing the optimal planes for each method for data satisfying Assumptions A1 to A4.

We now show that in the setting of symmetric domain label noise, UW and DS achieve identical WGA, which degrades with increasing noise, and outperform SRM, which is unaffected by domain label noise.

Theorem 4 *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$, $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$. Under Assumptions A1 to A4 and symmetric domain label noise with parameter p, let $\theta^{(p)}_{UW}$ and $\theta^{(p)}_{DS}$ denote the solution to (2) for UW and DS, respectively. Then, for any $\pi_0 \le 1/4$ and $p \le 1/2$, the WGA of DS and UW is strictly decreasing in p, and*

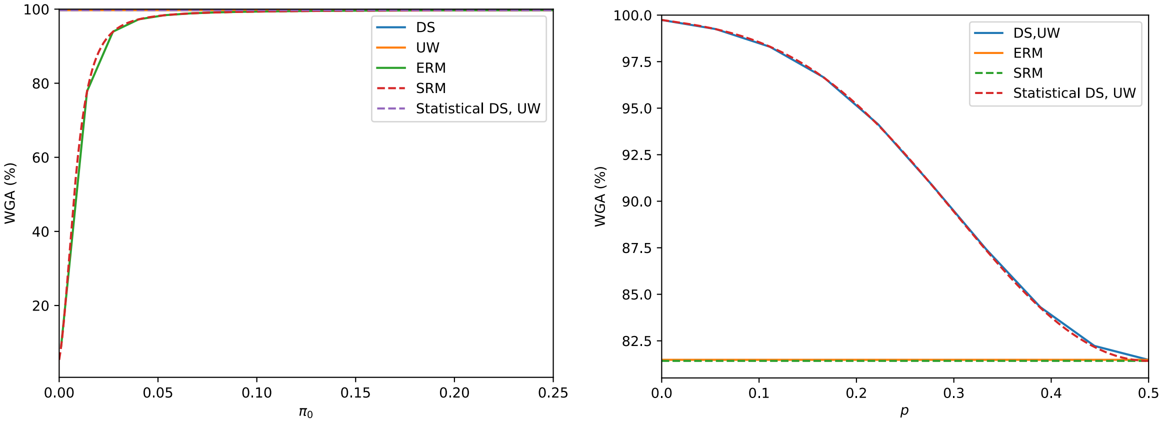
$$WGA(\theta^*_{SRM}) \le WGA(\theta^{(p)}_{DS}) = WGA(\theta^{(p)}_{UW}),$$

with equality when $\pi_0 = 1/4$ or $p = 1/2$.

Proof sketch. The proof is presented in Appendix C and involves showing that the WGA for DS under domain label noise is the same that for SRM (which is noise agnostic) but with a different prior dependent on p.

Fundamentally, this result can be seen as an effect of the domain label noise on the *perceived* (noisy) priors $\pi_0^{(p)}$ of the minority groups, which can be derived as

$$\pi_0^{(p)} := (1 - p)\pi_0 + p\,(1/2 - \pi_0)\,. \tag{6}$$



(a)

(b)

Figure 2: (a) A plot of the WGA of each data augmentation technique vs π_0, the prevalence of the minority groups, for data satisfying Assumptions A1 to A4. We see that as π_0 approaches 1/4 (a balanced dataset), the WGA of vanilla ERM increases to match that of the data augmented methods. (b) A plot of the WGA of each data augmentation technique vs p, the prevalence of domain label noise, for data satisfying Assumptions A1 to A4. The WGA of DS and UW decreases as p increases to 1/2. At the extreme point, the WGA of ERM is recovered.

UW would thus weight the loss in inverse proportion to the corresponding noisy (and not the true) prior for each group. The *true prior* of the minority group after DS can be derived as (see Appendix C)

$$\pi_{\text{DS}}^{(p)} := \frac{(1-p)\pi_0}{4\pi_0^{(p)}} + \frac{p\pi_0}{4(1/2-\pi_0^{(p)})}. \tag{7}$$

Thus, with noisy domain labels, instead of the desired balanced group priors after downsampling, from (7), DS results in a true minority prior that decreases from 1/4 to π_0.

4. Experimental Results

We present numerical results for both synthetic and real-world data for UW, DS, and *empirical risk minimization* (ERM) (the empirical, or finite sample, version of SRM).

4.1. Orthogonal Latent Gaussians

We first examine a numerical analog to the mixture Gaussian model given in Assumptions A1 to A4 to empirically verify our theoretical results. We generate n data points and calculate the empirical weights for each method by performing the corresponding data augmentation and then computing the sample variance (of X) and covariance (of X, Y) matrices used in the closed-form solution to (2) with $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$, $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$. This training step is repeated 10 times to account for randomness introduced by DS. Furthermore, we average over 10 runs (data generation and training) for different random seeds to account for randomness in the training data. We average over these runs when reporting statistics. We compute WGA by plugging in the empirical weights into the statistical forms derived in Theorem 3. We generate group-conditional Gaussian data with the following parameters satisfying Assumptions A1 to A4:

$$\Delta_C = \begin{pmatrix}0 & 1/4\end{pmatrix}^T \qquad \Delta_D = \begin{pmatrix}-1/4 & -1/4\end{pmatrix}^T$$
$$\Sigma = \begin{pmatrix}.002 & .002\\ .002 & .003\end{pmatrix} \qquad \pi_0 = 1/64.$$

We first demonstrate that both DS and UW are robust to the prevalence of the minority group. For fixed $n = 1\,000\,000$, we train each method with varying $\pi_0 \in [0, 1/4]$ and plot the corresponding WGA along with the true statistical WGA for each method in Figure 2(a). We see that both DS and UW achieve high constant WGA across all values of π_0 and are therefore robust to even very small minority groups. We additionally note that the WGA of ERM approaches that of both UW and DS as $\pi_0 \to 1/4$, the prior for a group-balanced dataset.

Next, we consider the setting of domain label noise. For fixed $n = 1\,000\,000$, we train each method with varying $p \in [0, 1/2]$ and fixed $\pi_0 = 1/64$ and plot the corresponding WGA along with the true statistical WGA for each method in Figure 2(b). DS and UW still outperform ERM; however, the WGA of both DS and UW decreases with increasing noise, dropping to the WGA of ERM, which remains constant because ERM does not use domain information, once $p = 1/2$.

4.2. Publicly Available Large Datasets

We next consider the CMNIST (Arjovsky et al., 2019), CelebA (Liu et al., 2015), and Waterbirds (Sagawa* et al., 2020) datasets, which are oft-used in LLR (Yang et al., 2023). CMINST (Arjovsky et al., 2019) is a variant of the MNIST handwritten digit dataset in which digits 0-4 are labeled $y = 0$ and digits 5-9 are labeled $y = 1$. The domain is given by color: 90% of digits labeled $y = 0$ are colored green and 10% are colored red and vice-versa for those labeled $y = 1$.

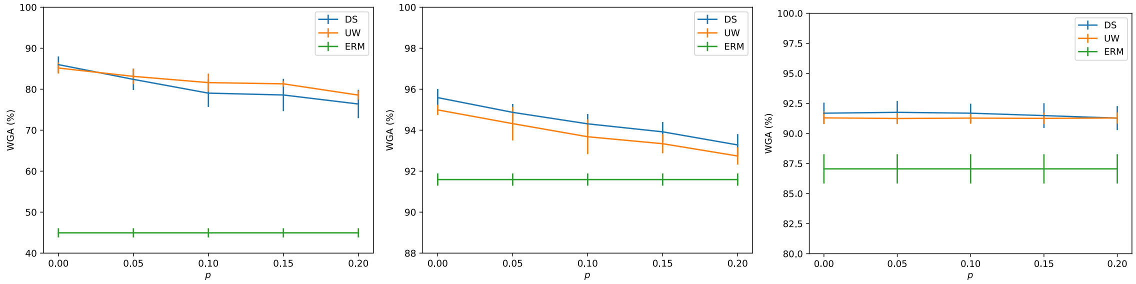
CelebA (Liu et al., 2015) is a dataset of celebrity faces. We predict hair color as either blonde ($y = 1$) or non-blonde ($y = 0$), while the domain label is either male ($d = 1$) or female ($d = 0$). There is a naturally induced correlation between hair color and gender in the dataset due to the prevalence of blonde females.

Waterbirds (Sagawa* et al., 2020) is a semi-synthetic image dataset comprised of land birds ($y = 1$) or sea birds ($y = 0$) on land ($d = 1$) or sea backgrounds ($d = 0$). There is a correlation between background and bird type in the training data (sea birds being more present with sea backgrounds) but this correlation is absent in the domain-balanced validation data.

Each dataset is split into training, validation, and test sets. The training data is used to train a large model (ResNet-50 architecture) from which we extract the embedding function $\phi(\cdot)$ used to obtain the latent representations. We view the validation data as a retraining dataset whose representations are used to retrain the last layer of the pretrained model.

Table 1: WGA (higher is better) Mean ± StDev (averaged over 10 runs)

	CMNIST	**CelebA**	**Waterbirds**
DS	93.0 ± 0.4	80.7 ± 3.1	90.1 ± 0.8
UW	94.6 ± 0.0	78.3 ± 0.0	90.0 ± 0.0
ERM	90.9 ± 0.0	43.3 ± 0.0	85.5 ± 0.0



(a) (b) (c)

Figure 3: A plot of the WGA vs. p, the prevalence of domain label noise for (a) CelebA, (b) CMNIST and (c) Waterbirds. For CelebA and CMNIST, we see that DS and UW degrade with increasing noise but still outperform ERM. Additionally, DS and UW have similar performance as expected. For Waterbirds, we see that domain noise does not strongly affect the performance of DS and UW as the dataset is domain balanced, although they still perform better than ERM.

In practice, state-of-the-art methods do not employ the MSE loss. Instead, common methods such as DFR (Kirichenko et al., 2023) use highly regularized losses such as log loss with ℓ_1 penalty. We proceed following this example and train logistic models with strong ℓ_1 regularization.

For each of these datasets, we see in Table 1 that UW and DS perform similarly and outperform ERM without augmentations. This suggests that our analysis may hold more generally than just on latent Gaussian subpopulations. We see that UW and ERM have no variance over runs which is due to the fact that both are deterministic methods, whereas DS introduces randomness.

Next, we consider the setting of domain label noise with varying $p \in [0, 1/5]$. We see in Figure 3(a) that for the CelebA dataset, DS and UW both achieve similar WGA even in the presence of domain label noise as suggested by Theorem 4. Furthermore, we see that their performance degrades as the level of noise increases, though both still outperform ERM. We similar trends in Figure 3(b) for the CMNIST dataset. We see in Figure 3(c) that DS and UW both are able to outperform ERM on the Waterbirds dataset, but their performance does not degrade with increasing domain label noise. This is likely due to Waterbirds having a domain-balanced validation set.

5. Conclusion

We have presented a new result that the well-known data augmentation techniques of DS and UW have statistically identical performance. For LLR, when the latent representations that are input to the last layer are modeled as Gaussian mixtures, DS and UW outperform SRM, even in the presence of domain label noise. Our results are validated for a synthetic Gaussian mixture dataset and appear to hold for several large publicly available datasets.

Acknowledgments

This work is supported in part by NSF grants CIF-1901243, CIF-2007688, DMS-2134256, and SCH-2205080.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on LearnIning Representations (ICLR)*, 2023.

Tyler LaBonte, Vidya Muthukumar, and Abhishek Kumar. Towards last-layer retraining for group robustness with fewer annotations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning (ICML)*, volume 139, pages 6781–6792, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2015.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *ICML*, volume 202, pages 28448–28467, 2023.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *ICLR*, 2020.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *ICLR*, 2022.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: a closer look at subpopulation shift. In *ICML*, 2023.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In *ICML*, volume 162, pages 25407–25437, 2022.