# Theoretical Guarantees of Data Augmented Last Layer Retraining Methods

Monica Welfert, Nathan Stromberg and Lalitha Sankar
Arizona State University
Email: {mwelfert, nstrombe, lsankar}@asu.edu

2024 IEEE International Symposium on Information Theory (ISIT) | 979-8-3503-8284-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISIT57864.2024.10619333

***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, downsampling and mixup, 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 the optimal worst-group accuracy when modeling the distribution of the latent representations (input to the last layer) as Gaussian for each subpopulation. We evaluate and verify our results for both synthetic and large publicly available datasets.**

## I. 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 [1]. 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* [2], [3], *upweighting* [4], [5], and *mixing* [6], [7] (Section II presents precise definitions of these methods). These augmentation techniques help to account for varying proportions of individual subpopulations and enable the final model to predict well on every subpopulation.

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 worst group error (WGE $= 1 -$ WGA) and finite sample performance.

This work is supported in part by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, DMS-2134256, and SCH-2205080.

To this end, analogous to [6], we model individual subpopulations as distinct Gaussian distributions. Our primary contribution is a straightforward comparison of the three most common data augmentation techniques for WGE: downsampling, upweighting, and mixing. We evaluate the performance of the abovementioned augmentation methods in three ways using: (i) the learned linear models, (ii) the resulting WGE, and (iii) the sample complexity in the setting of a finite number of training examples. 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 WGE for each data augmentation method under Gaussian subpopulations (Theorem 2).
- Sample complexity of each method (Theorem 3).
- Empirical results that match theory for Gaussian mixtures and the CMNIST, CelebA and Waterbirds datasets.

Our work is distinct from that in [6] as follows: (i) explicit incorporation of the minority group priors; (ii) providing precise WGE guarantees (in contrast to bounds in [6]); and (iii) including downsampling and upweighting as data augmentation methods (the focus in [6] is primarily on mixing and its variants) for which we also provide comparative model and error guarantees beyond the Gaussian setting.

## II. 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 $X := \phi(X_a)$ for an embedding function $\phi : \mathcal{X}_a \to \mathcal{X} \subseteq \mathbb{R}^p$ 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 $\pi^{(y,d)} := P(Y = y, D = d)$ 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 $\theta = (w, b) \in \mathbb{R}^{p+1}$ given by

$$f_\theta(x) = w^T x + b. \tag{1}$$

The statistically optimal linear model is obtained by minimizing the risk defined as

$$R(f_\theta) := \mathbb{E}_{P_{X,Y,D}}[\ell(f_\theta(X), Y)], \tag{2}$$

where $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}_+$ is a loss function. We consider four different methods to learn a classifier: (a) standard risk minimization (SRM), (b) downsampling (DS), (c) upweighting (UW), and (d) intra-class domain mixup (MU). In particular, SRM involves minimizing (2) as is whereas downsampling involves reducing the size of each group to that of the smallest one while upweighting involves scaling the loss for each group in proportion to the inverse of the prior. Finally, intra-class domain mixup takes an arbitrary convex combination of two randomly sampled representations from the same class but from different domains.

A general formulation for obtaining the optimal $f_{\theta^*}$ is:

$$\theta^* = \arg\min_\theta \mathbb{E}_{P_{X,Y,D}}[\ell(f_\theta(X), Y)c(Y, D)], \tag{3}$$

where $c(y, d) = 1$, $(y, d) \in \mathcal{Y} \times \mathcal{D}$, for SRM, DS, and MU, 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 $\pi^{(y,d)} = 1/4$. Finally, for MU, the representation $X$ is now $X = \Lambda X_1 + (1 - \Lambda)X_2$ where $X_1 \sim P_{X|Y=y,D=S}$, $X_2 \sim P_{X|Y=y,D=T}$, $y \in \mathcal{Y}$, and the mixup parameter $\Lambda \sim \text{Beta}(\alpha, \alpha)$.

We desire a model that makes fair decisions across groups, and therefore, we evaluate worst-group error, i.e., the maximum error among all groups, defined for a model $f_\theta$ as

$$\text{WGE}(f_\theta) := \max_{(y,d)\in\mathcal{Y}\times\mathcal{D}} E^{(y,d)}(f_\theta), \tag{4}$$

where $E^{(y,d)}(f_\theta)$ denotes the per-group misclassification error for $(y, d) \in \mathcal{Y} \times \mathcal{D}$. Specifically, for $(y, d) \in \mathcal{Y} \times \mathcal{D}$:

$$E^{(y,d)}(f_\theta) := P(\mathbb{1}\{f_\theta(X) > 1/2\} \neq Y | Y = y, D = d) \tag{5}$$

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

## III. 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 of an extended version [8] and outline a proof sketch here.

**Theorem 1.** *For any given $P_{X,Y,D}$ and loss $\ell$, the objectives in* (3) *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., $\theta^*_{DS} = \theta^*_{UW}$.*

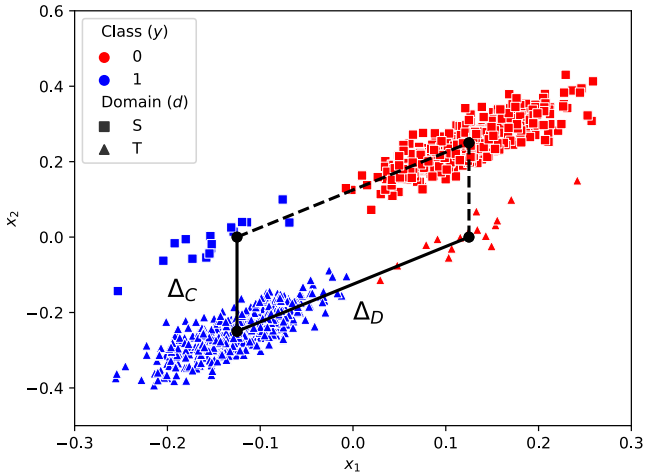


Fig. 1. $\Delta_C$ and $\Delta_D$ are shown as line segments between group means overlaid on data sampled from Gaussian mixtures satisfying Assumptions A1 to A4.

*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 [8, Appendix A].

**Remark 1.** Although we are focused on the binary class and domain label setting, Theorem 1 holds for any number of classes and domains by replacing $\pi^{(y,d)} = 1/n_g$ and $c(y, d) = 1/(n_g\pi^{(y,d)})$ for $(y, d) \in \mathcal{Y} \times \mathcal{D}$, where $n_g$ is the number of groups. To the best of our knowledge, such an analysis, albeit simple, has not been presented before.

While Theorem 1 holds for any general data distribution, to obtain more refined guarantees on WGE and model parameters for 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., [6]).

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

$$X|(Y = y, D = d) \sim \mathcal{N}(\mu^{(y,d)}, \Sigma), \tag{6}$$

for $(y, d) \in \mathcal{Y} \times \mathcal{D}$, where $\mu^{(y,d)} := \mathbb{E}[X|Y = y, D = d] \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ 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, where $\Delta_D$ and $\Delta_C$ are shown on samples drawn from a distribution satisfying Assumptions A1 to A3.

**Proposition 1.** *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$ for $\hat{y} \in \mathbb{R}$ and $y \in \mathcal{Y}$ be the mean-squared error (MSE) loss. Under Assumptions A1 to A3, the minimizers in* (3) *for DS and UW are the same, i.e.,*

$$\theta^*_{DS} = \theta^*_{UW},$$

$$w^*_{DS} = w^*_{UW} = \frac{1}{4}\left(\Sigma + \frac{1}{4}\Delta_C\Delta_C^T + \frac{1}{4}\Delta_D\Delta_D^T\right)^{-1}\Delta_D \quad (7)$$

$$b^*_{DS} = b^*_{UW} = \frac{1}{2} - \frac{1}{2}(w^*_{DS})^T\left(\mu^{(0,T)} + \mu^{(1,S)}\right), \quad (8)$$

*and thus, $WGE(f_{\theta^*_{DS}}) = WGE(f_{\theta^*_{UW}})$.*

*Proof sketch.* The proof of parameter equality follows directly from Theorem 1. To obtain the specific forms of the parameters, we derive the optimal parameters in (3) for the given $\ell$ and appropriate values of $c(y,d)$, $(y,d) \in \{0,1\} \times \{S,T\}$, for DS and UW. We then use Assumption A1 to obtain the WGE in terms of Gaussian CDFs (as detailed in [8, Appendix B]).

Note that if $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2 + \lambda\|w\|_1$ for $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$, and a regularizer $\lambda > 0$, then (3) simplifies to the deep feature reweighting (DFR) optimization, an $\ell_1$-regularized DS method that achieves state of the art WGA for many datasets [2].

**Corollary 1.** *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2 + \lambda\|w\|_1$ for $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$, and $\lambda > 0$. Under Assumptions A1 to A3,the minimizer in* (3) *for DFR is the same as that for UW, i.e.,*

$$\theta^*_{DFR} = \theta^*_{UW}.$$

*Thus,*

$$WGE(f_{\theta^*_{DFR}}) = WGE(f_{\theta^*_{UW}}).$$

The proof of Corollary 1 follows from Theorem 1 and Proposition 1.

As derived in the proof of Proposition 1, the WGEs result from computing Gaussian CDFs at the optimal model for each method. However, while Proposition 1 clarifies the statistical behavior of DS and UW, comparing the resulting analytical expressions for WGEs for each of the four methods requires finer assumptions. To this end, we make the following orthogonality assumption.

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

**Theorem 2.** *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$, $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$. For Assumptions A1 to A3, the optimal SRM and MU models are:*

$$w^*_{SRM} = \frac{1}{4}\left(\Sigma + 2\pi_0(1-2\pi_0)\Delta_C\Delta_C^T + \frac{1}{4}\bar{\Delta}\bar{\Delta}^T\right)^{-1}\bar{\Delta}, \quad (9)$$

*where $\bar{\Delta} := \mu^{(1)} - \mu^{(0)} = \Delta_D - (1-4\pi_0)\Delta_C$, and*

$$b^*_{SRM} = \frac{1}{2} - \frac{1}{2}(w^*_{SRM})^T(\mu^{(0,T)} + \mu^{(1,S)}), \quad (10)$$

$$w^*_{MU} = \frac{1}{4}\left(2\mathbb{E}[\Lambda^2]\Sigma + Var(\Lambda)\Delta_C\Delta_C^T + \frac{1}{4}\Delta_D\Delta_D^T\right)^{-1}\Delta_D, \quad (11)$$

$$b^*_{MU} = \frac{1}{2} - \frac{1}{2}(w^*_{MU})^T(\mu^{(0,T)} + \mu^{(1,S)}). \quad (12)$$

*Additionally, under Assumption A4 and for $\pi_0 < 1/4$,*

$$WGE(f_{\theta^*_{SRM}}) > WGE(f_{\theta^*_{DS}}) = WGE(f_{\theta^*_{UW}}) = WGE(f_{\theta^*_{MU}}).$$

*Proof sketch.* The proof follows similarly to that of Proposition 1 using the appropriate values of $c(y,d)$, $(y,d) \in \{0,1\} \times \{S,T\}$, for SRM and MU. We employ a derivative analysis to show $\mathrm{WGE}(f_{\theta^*_{\mathrm{SRM}}}) > \mathrm{WGE}(f_{\theta^*_{\mathrm{DS}}})$ and then show the remaining equalities. See Figure 2 for a plot showing the optimal planes for each method for data satisfying Assumptions A1 to A4. See [8, Appendix C] for a detailed proof.

In practice, we only have access to a finite number of samples. In this setting, we can approximate the risk in (2) by the empirical risk, defined for a given dataset of $n$ samples $(x_i, y_i, d_i) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{D}$, $i = 1, \ldots, n$, drawn i.i.d. from $P_{X,Y,D}$ and a loss $\ell$ as

$$\hat{R}(f_\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell(f_\theta(x_i), y_i). \quad (13)$$

We consider the same four methods as before where SRM is now just the empirical risk minimization (ERM). The empirically optimal $f_{\hat{\theta}} = \hat{w}^T x + \hat{b}$ is obtained from

$$\hat{\theta} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n}\ell(f_\theta(x_i), y_i)c(y_i, d_i), \quad (14)$$

where again $c(y,d) = 1$, $(y,d) \in \mathcal{Y} \times \mathcal{D}$ for ERM, DS, and MU, but $c(y,d) = n/(4n^{(y,d)})$ for UW with $n^{(y,d)}$ being the number of samples in the group $(y,d)$. In the case of DS, rather than using $n$ samples, we use $4n_{\min} := \min_{(y,d)} n^{(y,d)}$ samples. Finally, for MU, we use $x_i = \lambda_i x_{i_1} + (1-\lambda_i)x_{i_2}$ where $i_1$ and $i_2$ are uniformly chosen from the indices of samples in the groups $(y_i, G)$ and $(y_i, R)$, respectively, and the mixup parameter $\lambda_i \sim \mathrm{Beta}(\alpha, \alpha)$. The following result compares the sample complexity of each of the four methods. We use the notation $O_p(\cdot)$ for the stochastic boundedness of a sequence of random variables [9]. More formally, for a sequence of random variables $X_n$ and a sequence of positive scalars $a_n$, $\|X_n\|_2 = O_p(a_n)$ if for any $\varepsilon > 0$ there exist finite $M > 0$ and $N > 0$ such that

$$P\left(\|X_n/a_n\|_2 \le M\right) \ge 1 - \varepsilon, \quad \text{for all } n > N. \quad (15)$$

**Theorem 3.** *Let $\ell(\hat{y}, y) = \|y - \hat{y}\|_2^2$, $\hat{y} \in \mathbb{R}$, $y \in \mathcal{Y}$. Consider $n$ i.i.d. samples generated according to* (6)*, with $n_{min}$ being the number of samples in the minority groups. Then*

$$\|\theta^*_{ERM} - \hat{\theta}_{ERM}\|_2^2 = \|\theta^*_{UW} - \hat{\theta}_{UW}\|_2^2 = O_p(p/n),$$

$$\|\theta^*_{DS} - \hat{\theta}_{DS}\|_2^2 = O_p(p/n_{min}),$$

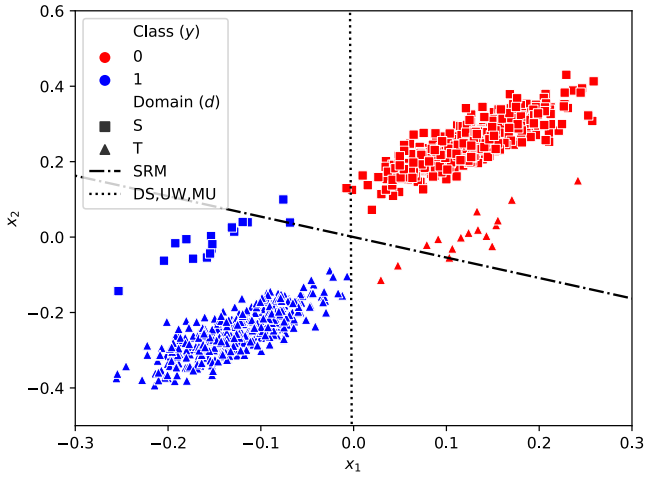


Fig. 2. The optimal prediction planes for DS, UW, MU, and SRM are shown overlaid on data sampled from Gaussian mixtures satisfying Assumptions A1 to A4. The SRM model largely ignores the minority group for each class.

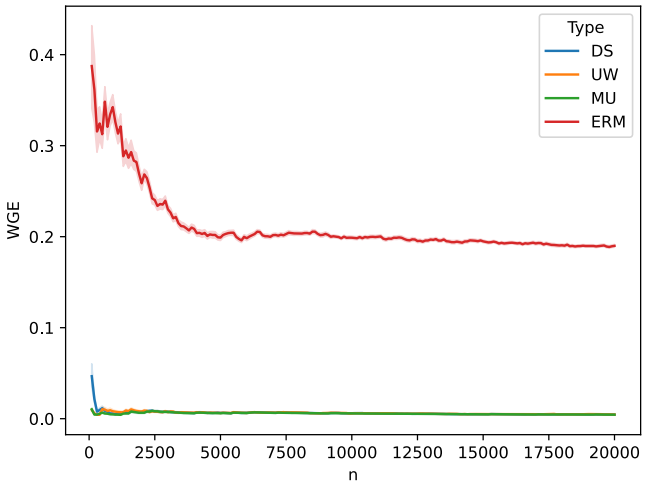


Fig. 3. WGE for UW, DS, MU and ERM for the data in Figure 2. As the number of samples $n$ increases, UW, DS, and MU perform better than ERM.

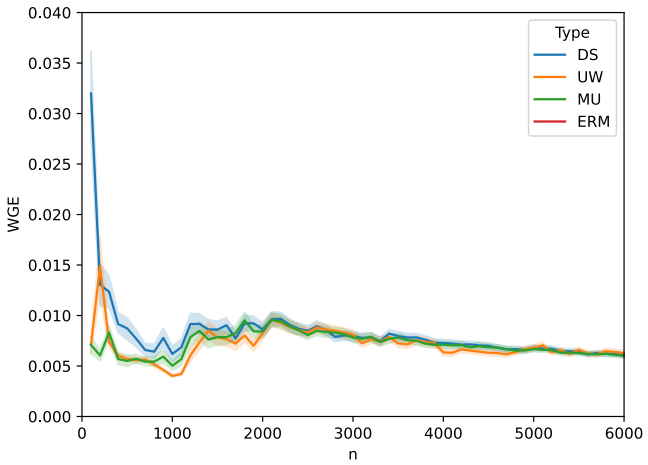


Fig. 4. Zoomed in version of Figure 3 where we see the differences between data augmentation methods, especially for small $n$.

*and*

$$\|\theta^*_{MU} - \hat{\theta}_{MU}\|_2^2 = O_p(p\log(n)/n + p/n_{min}).$$

*Proof sketch.* The sample complexity bound for ERM follows from a standard application of the weak law of large numbers [10]. Furthermore, the bound for DS is obtained by setting $n = 4n_{\min}$ while that for UW is a straightforward generalization of that for ERM to the weighted least squares setting. The bound for MU is from [6].

## IV. Experimental Results

We present numerical results for both synthetic and real-world data for all the augmentation techniques and ERM.

### A. Orthogonal Latent Gaussians

We first examine a numerical analog to the mixture Gaussian model given in Assumptions A1 to A4 to empirically study the convergence of DS, UW, and MU methods in terms of MSE from the corresponding statistical solutions. 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 (14) 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 and MU. 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 generate group-conditional Gaussian data with the following parameters satisfying Assumptions A1 to A4:

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

In Figure 3, we compare the WGE of each training method as a function of the number of samples. We see that each of the data augmentation methods outperforms (non-augmented vanilla) ERM, and they all converge to the same WGE. The equivalence of these methods for large $n$ is implied by Theorem 2. However, we see interesting behavior at small $n$ in Figure 4: DS achieves worse WGE than UW or MU at very small $n$. This may be explained by the fact that DS often throws away data while MU and UW keep all available data. Therefore, DS may not be well-suited to limited data regimes.

We next compare the empirical weights and bias obtained by each method to the corresponding statistically optimal weights and bias as calculated in (7), (8) for DS and UW, in (9), (10), for SRM, and in (11), (12) for MU. We report the MSE as a function of $n$ in Figure 5 and compare our results to the bounds found in Theorem 3. We see that each method converges at a similar rate, suggesting that tighter sample complexity analysis may be possible.

Finally we demonstrate that each of the data augmentation methods is robust to the prevalence of the minority group. For a fixed $n = 10,000$, we train each method with varying $\pi_0$.

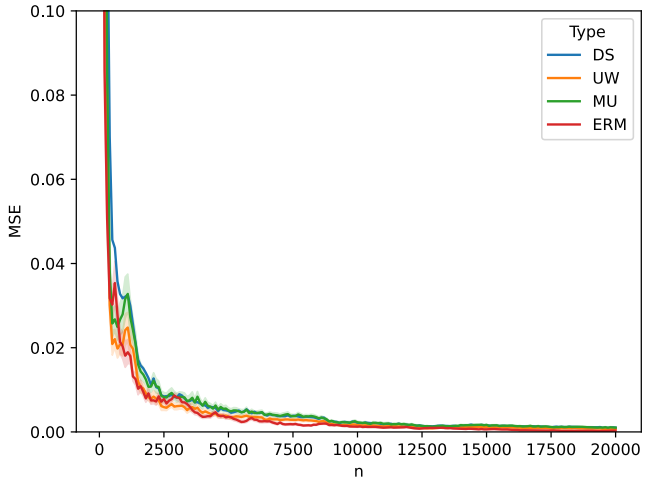


Fig. 5. Mean squared error of the estimated weights from data as compared to the expected weights. We see that each method converges quickly to the expected weights as a function of $n$.

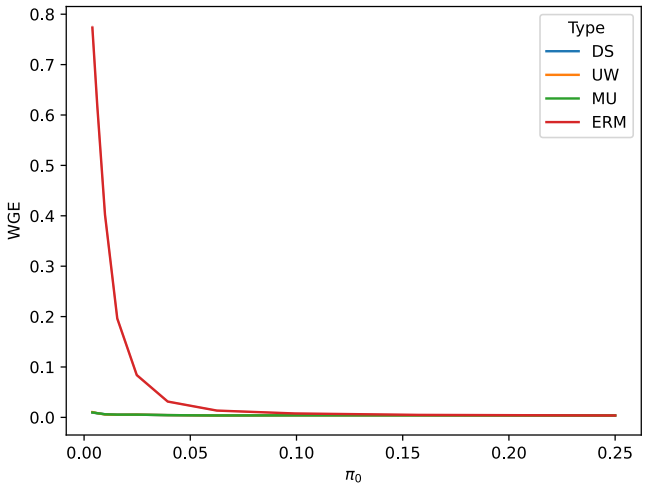


Fig. 6. Worst-group error on latent Gaussian subpopulations for each data augmentation technique as a function of $\pi_0$, the prevalence of the minority groups. We see that as $\pi_0$ approaches $1/4$ (a balanced dataset), the WGE of vanilla ERM decreases to match that of the data augmented methods.

We see in Figure 6 that the data augmentation methods are robust to even very small minority groups. We additionally note that the performance of ERM approaches that of the data augmented methods as $\pi_0 \to 1/4$, i.e., the prior for a group-balanced dataset.

### *B. Publicly Available Large Datasets*

We next consider the CMNIST [11], CelebA [12], and Waterbirds [13] datasets, which are oft-used in LLR [1]. CMINST [11] 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 [12] 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.

TABLE I
WGE (LOWER IS BETTER) MEAN $\pm$ STDEV (AVERAGED OVER 10 RUNS)

| | CMNIST | CelebA | Waterbirds |
|---|---|---|---|
| DS | 7.0 $\pm$ 0.4 | 19.3 $\pm$ 3.1 | 9.9 $\pm$ 0.8 |
| UW | 5.4 $\pm$ 0.0 | 21.7 $\pm$ 0.0 | 10.0 $\pm$ 0.0 |
| MU | 6.2 $\pm$ 0.4 | 22.9 $\pm$ 1.5 | 10.0 $\pm$ 0.7 |
| ERM | 9.1 $\pm$ 0.0 | 56.7 $\pm$ 0.0 | 14.5 $\pm$ 0.0 |

Waterbirds [13] 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 group- and class-balanced validation data.

Each dataset is broken into training, validation, and test data. 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.

In practice, state-of-the-art methods do not employ the MSE loss. Instead, common methods such as DFR [2] use highly regularized losses such as log loss with $\ell_1$ penalty. We proceed following this example, and train logistic models with strong $\ell_1$ regularization.

For each of these datasets, we see in Table I that all of the data augmentation methods perform similarly and outperform ERM alone. This suggests that the analysis provided here 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 and MU introduce randomness. Additionally, these results suggest that DS – the most common data augmentation method for WGA – may not have strong advantages over UW or MU, which have advantages in variance, though not in computational complexity.

## V. 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, MU also achieves the same statistical worst-group accuracy as DS and UW, all of which are better than SRM. Our results are validated for a synthetic Gaussian mixture dataset and appear to hold for several large publicly available datasets. A natural extension is to obtain more refined sample complexity, or equivalently excess risk bounds, when explicitly accounting for the size of each group/subpopulation. An equally compelling question to address is characterizing the finite sample differences between UW and DS for a larger class of distributions building upon the work in [14].

## REFERENCES


[1] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi, "Change is hard: a closer look at subpopulation shift," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[2] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," in *The Eleventh International Conference on LearnIning Representations (ICLR)*, 2023.

[3] T. LaBonte, V. Muthukumar, and A. Kumar, "Towards last-layer re-training for group robustness with fewer annotations," in *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[4] E. Z. Liu, B. Haghgoo, A. S. Chen, A. Raghunathan, P. W. Koh, S. Sagawa, P. Liang, and C. Finn, "Just train twice: Improving group robustness without training group information," in *Proceedings of the 38th ICML*, vol. 139, 2021, pp. 6781–6792.

[5] S. Qiu, A. Potapczynski, P. Izmailov, and A. G. Wilson, "Simple and fast group robustness by automatic feature reweighting," in *Proceedings of the 40th ICML*, vol. 202, 2023, pp. 28 448–28 467.

[6] H. Yao, Y. Wang, S. Li, L. Zhang, W. Liang, J. Zou, and C. Finn, "Improving out-of-distribution robustness via selective augmentation," in *Proceedings of the 39th ICML*, vol. 162, 2022, pp. 25 407–25 437.

[7] G. Giannone, S. Havrylov, J. Massiah, E. Yilmaz, and Y. Jiao, "Just mix once: Mixing samples with implicit group distribution," in *NeurIPS 2021 Workshop on Distribution Shifts*, 2021.

[8] M. Welfert, N. Stromberg, and L. Sankar, "Theoretical guarantees of data augmented last layer retraining methods," *arXiv:2405.05934*, 2024.

[9] Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete multivariate analysis: Theory and practice*. Springer Science & Business Media, 2007.

[10] M. Krikheli and A. Leshem, "Finite sample performance of linear least squares estimation," *Journal of the Franklin Institute*, vol. 358, no. 15, pp. 7955–7991, 2021.

[11] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2019.

[12] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[13] S. Sagawa*, P. W. Koh*, T. B. Hashimoto, and P. Liang, "Distributionally robust neural networks," in *ICLR*, 2020.

[14] K. Chaudhuri, K. Ahuja, M. Arjovsky, and D. Lopez-Paz, "Why does throwing away data improve worst-group error?" in *Proceedings of the 40th ICML*, 2023.