ON THE UNREASONABLE EFFECTIVENESS OF LAST-LAYER RETRAINING

John C. Hill[1*], **Tyler LaBonte**[2], **Xinchen Zhang**[1], **& Vidya Muthukumar**[1,2]
[1]School of Electrical and Computer Engineering, Georgia Institute of Technology
[2]H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology

ABSTRACT

Last-layer retraining (LLR) methods — wherein the last layer of a neural network is reinitialized and retrained on a held-out set following ERM training — have recently garnered interest as an efficient approach to rectify dependence on spurious correlations and improve performance on minority groups. Surprisingly, LLR has recently been found to improve worst-group accuracy even when the held-out set is an imbalanced subset of the training set. We initially hypothesize that this "unreasonable effectiveness" of LLR is explained by its ability to mitigate neural collapse through the held-out set, resulting in the implicit bias of gradient descent benefiting robustness. Our empirical investigation *does not* support this hypothesis. Instead, we present strong evidence for an alternative hypothesis: that the success of LLR is primarily due to better group balance in the held-out set. We conclude by showing how the recent algorithms CB-LLR and AFR perform implicit group-balancing to elicit a robustness improvement.

1 INTRODUCTION

The standard neural network training procedure of empirical risk minimization (ERM) (Vapnik, 1998), which minimizes the average classification loss, is well-known to overfit to spurious correlations in the training set (Geirhos et al., 2020). The conjunction of target labels and spurious features form *minority groups*, upon which ERM performance may be no better than random guessing (Shah et al., 2020). Due to the relevance of spurious correlations in high-consequence applications, *e.g.,* medicine (Zech et al., 2018) and criminal justice (Chouldechova, 2016), significant work has focused on algorithms which maximize worst-group accuracy (WGA) — also called group robustness (Sagawa et al., 2020).

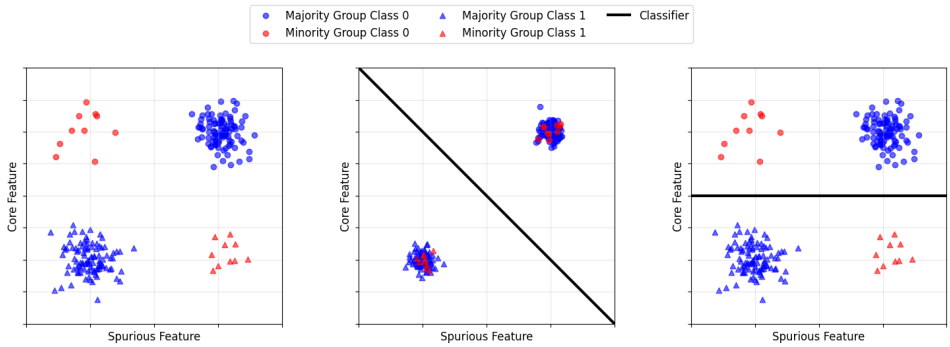
A promising family of group robustness methods are based on last-layer retraining (LLR), wherein the last layer is reinitialized and retrained on a held-out set following ERM training. The original LLR method, called deep feature reweighting (DFR), requisites the held-out set to comprise equal data from each group (Kirichenko et al., 2023). This limits its practical application, as the groups are often unknown ahead of time or difficult to annotate.[1] While DFR still performs the best, group information (even on the validation set) was recently found to be unnecessary to observe WGA improvement from LLR (Qiu et al., 2023; LaBonte et al., 2023). This surprising observation has led LLR to be termed a "free lunch" for group robustness (LaBonte et al., 2023).

Contributions. In this paper, we take a "scientific method" approach to investigate why LLR on an imbalanced held-out subset of the training set can perform so well. Our contributions include:

- We propose an initial hypothesis (visualized in Figure 1) that neural collapse on the training set results in a biased ERM classifier since the class means are dominated by majority group data. We initially hypothesize that during LLR, the features are not collapsed on the held-out set, and so the implicit bias of gradient descent elicits a maximum-margin classifier that could enjoy better robustness guarantees.

*Corresponding author. Email: jhill326@gatech.edu.

[1]However, DFR compares favorably in this respect to methods which require group annotations for the entire training set, such as group distributionally robust optimization (DRO) (Sagawa et al., 2020).



(a) Dataset with Majority and Minority Groups (b) Collapsed Features with Biased Classifier (c) Held-out Features with Optimal Classifier

Figure 1: **Visualization of our initial hypothesis.** In Figure 1(a), we visualize the feature space of a training set with two classes and two features, one core and one spurious. In Figure 1(b), we visualize our initial hypothesis that neural collapse on the training set causes the features to collapse to their class means — dominated by majority group data — resulting in a biased ERM classifier. In Figure 1(c), we visualize our initial hypothesis that during LLR, the features are not collapsed on the held-out set, and so the implicit bias of gradient descent elicits a maximum-margin classifier which is invariant to the spurious feature. Importantly, our empirical investigation *does not* support our initial hypothesis. Instead, we find that the success of LLR is primarily explained by better group balance in the held-out set.

- We put forth evidence which *does not* support our initial hypothesis. First, neural collapse either does not occur or occurs after the standard number of epochs for our benchmark datasets. Moreover, convergence of the LLR classifier to the maximum-margin solution is extremely slow, and the *average* margin over all training data is much more correlated with group accuracy than the (standard) *minimum* margin over training data.
- We present strong evidence for an alternative hypothesis: that the success of LLR is primarily due to *better group balance* in the held-out set, often achieved implicitly through class balancing. Our experiments indicate that LLR does not improve over ERM when the held-out set has the same group balance as the training set, with Pearson correlation coefficient $r > 0.99$ on CelebA and CivilComments. On the other hand, LLR with a better group balance enjoys drastically improved WGA, and vice versa.
- Ultimately, we reevaluate the "free lunch" interpretation of LLR by showing that class balanced LLR (CB-LLR) (LaBonte et al., 2023) and automatic feature reweighting (AFR) (Qiu et al., 2023) owe their improved WGA primarily to implicit group-balancing on the held-out set. On a positive note, we show that LLR can recover the WGA of an optimally class-balanced model even when ERM was not optimally class-balanced. More broadly, LLR remains an effective method to achieve group robustness using group annotations only on the held-out set (or proxies thereof, such as in SELF (LaBonte et al., 2023) and AFR (Qiu et al., 2023)).

Related work. Reliance on spurious correlations is a widely-studied phenomenon in machine learning known to exacerbate bias and hinder generalization, *e.g.,* Beery et al. (2018); Geirhos et al. (2020); Sagawa et al. (2020). Last-layer retraining, proposed by Kirichenko et al. (2023), has recently garnered interest as an efficient approach to improve group robustness in a variety of settings (Qiu et al., 2023; LaBonte et al., 2023; Stromberg et al., 2024; Park et al., 2025). Our work aims to further a fundamental understanding of LLR methods, following broadly in the spirit of Izmailov et al. (2022); Ye et al. (2023); Chen et al. (2023); Welfert et al. (2024); Qiu et al. (2024). Our initial hypothesis primarily concerns neural collapse, a phenomenon introduced by Papyan et al. (2020); Han et al. (2022) which has been applied to the study of fairness and spurious correlations in the context of ERM by Lu et al. (2024); Wang et al. (2024); Chen et al. (2024); Xu et al. (2025), but never before to LLR. We also consider the implicit bias of gradient descent in logistic regression towards the maximum ℓ_2-margin classifier, studied in the separable case by Soudry et al. (2018)

and the non-separable case by Ji & Telgarsky (2019) (though a rich literature exists on this topic; see Vardi (2023) for a recent survey). The maximum-margin classifier has recently been shown to converge, under certain conditions, to the classifier with optimal worst-group accuracy (Chaudhuri et al., 2023). Finally, in Section 3 we evaluate the impact of group balance on LLR performance via a case study on class-balancing, a popular lightweight method to improve group robustness studied by Idrissi et al. (2022); Chaudhuri et al. (2023); Shwartz-Ziv et al. (2023); LaBonte et al. (2024).

Setting. We consider the setting of classification tasks with input domain $\mathcal{X}$ and target classes $\mathcal{Y}$. We assume the existence of a set of *spurious features* $\mathcal{S}$ such that each data point $x \in \mathcal{X}$ is associated with a single spurious feature $s(x) \in \mathcal{S}$. The dataset is partitioned into *groups* $\mathcal{G}$ by the Cartesian product of classes and spurious features: $\mathcal{G} := \mathcal{Y} \times \mathcal{S}$. Let $\Omega_g \subseteq \{1, ..., m\}$ denote the indices of training points belonging to group $g \in \mathcal{G}$. Similarly, let $\Omega_y \subseteq \{1, ..., m\}$ denote the indices of training points belonging to class $y \in \mathcal{Y}$. We define the *majority group(s)* to be the groups which maximize $|\Omega_g|$. All other groups are considered to be *minority groups*. Importantly, we also define the *worst group(s)* to be the group(s) with lowest test accuracy. Our goal is to find a model which, despite group imbalance in the training set, performs uniformly well across groups. To evaluate this, we use *worst-group accuracy* (WGA), or the minimum accuracy achieved among all groups (Sagawa et al., 2020).

Datasets and models. We employ four benchmark classification datasets in our experiments: Waterbirds (Welinder et al., 2010; Wah et al., 2011; Sagawa et al., 2020), CelebA (Liu et al., 2015; Sagawa et al., 2020), CivilComments (Borkan et al., 2019; Koh et al., 2021), and MultiNLI (Williams et al., 2018; Sagawa et al., 2020). We discuss each dataset at length in Appendix B.1. We use ResNet-18 and ResNet-50 (He et al., 2016) models pretrained on ImageNet-1K (Russakovsky et al., 2015) for the Waterbirds and CelebA datasets, and a BERT-Base (Devlin et al., 2019) model pretrained on Book Corpus (Zhu et al., 2015) and English Wikipedia for CivilComments and MultiNLI. Our implementation of LLR uses SGD with no learning rate schedule or regularization. Following standard practice, we use 20% of the training set for the held-out set on Waterbirds and half the validation set as the held-out set in all other datasets. Further training details are located in Appendix B.2.

2 Neural collapse and implicit bias do not explain LLR

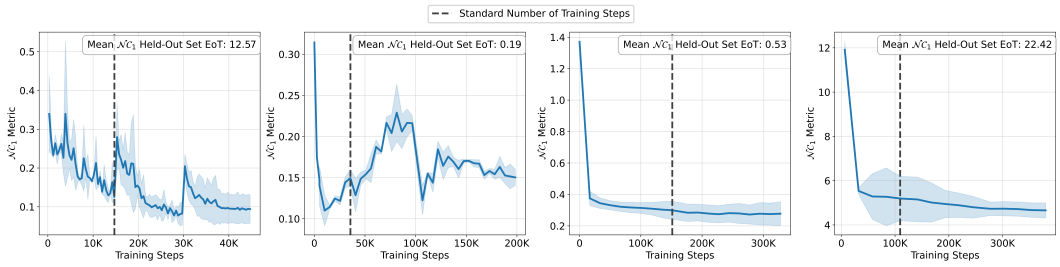
The interpretation of LLR methods as logistic regression on top of convergent ERM features lends itself to a possible understanding combining a feature learning phenomenon (*neural collapse*) and a logistic regression phenomenon (*implicit bias*). Neural collapse was first identified by Papyan et al. (2020); Han et al. (2022) and implies that the penultimate layer features collapse to their class means over the course of ERM training. On the other hand, implicit bias results state that linear classifiers trained via gradient descent with the unregularized logistic loss converge in direction to the maximum-margin SVM solution (Soudry et al., 2018; Ji & Telgarsky, 2019).

Combining these two ideas, we developed the following initial hypothesis (visualized in Figure 1). We hypothesized that the model undergoes neural collapse during ERM training, causing the features to collapse to their class means. These class means, however, are dominated by the largest groups in each class — resulting in ERM learning a biased classifier. We hypothesized that during LLR, the features are not collapsed as the held-out set was not seen during ERM; hence, the implicit bias of gradient descent elicits a robust maximum-margin classifier.

In this section, we present evidence which *does not* support our initial hypothesis. Specifically, we find that neural collapse either does not occur or occurs after the standard number of epochs of ERM training. Moreover, convergence of the LLR classifier to the maximum-margin solution is extremely slow, and the *average* margin is much more correlated with group accuracy than the (standard) *minimum* margin.

2.1 Neural collapse may not occur during standard ERM training

We will measure the collapse of variability among feature representations via a metric called $\mathcal{NC}_1$. For each class $y \in \mathcal{Y}$ and each training example $i \in \Omega_y$ in that class, we denote the penultimate layer feature vector of i as $f_{y,i}$. Neural collapse posits that features collapse to their respective



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 2: **Collapse of class feature variability occurs after standard ERM training, if at all.** We plot the empirical metric of neural collapse $\mathcal{NC}_1 := \frac{1}{|\mathcal{Y}|}\text{tr}(\Sigma_A \Sigma_R^\dagger)$ throughout training a ResNet-18 on Waterbirds and CelebA and a BERT-Tiny on CivilComments and MultiNLI. Each plot displays the mean and standard deviation for $\mathcal{NC}_1$ computed across 3 experimental seeds. We also display the mean $\mathcal{NC}_1$ metric computed on the features of the held-out set at the end of training (EoT). We believe the cyclic spikes in the Waterbirds plot are due to the cosine learning rate scheduler used when training the model.

class mean $\mu_y := \frac{1}{|\Omega_y|}\sum_{i \in \Omega_y} f_{y,i}$. We compute an empirical metric of this collapse using the intra-class covariance matrix $\Sigma_A := \frac{1}{m}\sum_{y \in \mathcal{Y}}\sum_{i \in \Omega_y}(f_{y,i} - \mu_y)(f_{y,i} - \mu_y)^\top$, the global feature mean $\mu_G := \frac{1}{|\mathcal{Y}|}\sum_{y \in \mathcal{Y}} \mu_y$, and the inter-class covariance matrix $\Sigma_R := \frac{1}{|\mathcal{Y}|}\sum_{y \in \mathcal{Y}}(\mu_y - \mu_G)(\mu_y - \mu_G)^\top$. We then define $\mathcal{NC}_1 := \frac{1}{|\mathcal{Y}|}\text{tr}(\Sigma_A \Sigma_R^\dagger)$, where $\Sigma_R^\dagger$ denotes the pseudo-inverse of Σ_R. Neural collapse is formalized by $\mathcal{NC}_1 \to 0$ (Han et al., 2022; Kothapalli, 2023).[2]

If neural collapse plays a significant role in ERM learning a biased classifier prior to LLR, then a collapse in feature variability should be observed during a standard number of ERM training steps. We display $\mathcal{NC}_1$ for a ResNet-18 on Waterbirds and CelebA in Figure 2. We train on both datasets for much longer than normal and denote the standard number of training steps in the spurious correlations literature (Sagawa et al., 2020; Kirichenko et al., 2023).

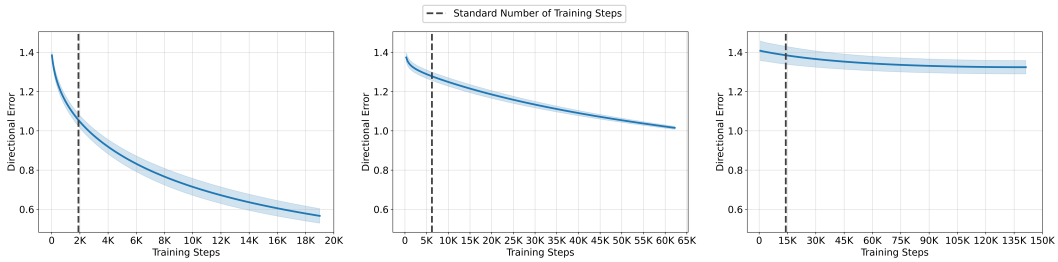
In Figure 2, we see evidence of feature collapse on Waterbirds and MultiNLI. $\mathcal{NC}_1$ is at least an order of magnitude larger for the held-out set than for the training set by the end of training. However, complete collapse is not achieved by the standard number of epochs. The CelebA and CivilComments features do not appear to collapse within any reasonable number of training steps as $\mathcal{NC}_1$ remains comparable between the training and held-out sets even after training for more than double the standard number of training steps. Also, $\mathcal{NC}_1$ measured on the CelebA and CivilComments training sets experiences no significant decrease after the first epoch. Thus, it is unlikely that neural collapse significantly affects the classifier learned by ERM prior to LLR in real-world datasets.

2.2 The minimum margin of LLR is not predictive of robust generalization

Let $f_\theta : \mathcal{X} \to \{-1, 1\}$ be a binary linear classifier defined by $f_\theta(x) = x^\top \theta + b$. Let $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \{-1, 1\}$ be the training set and $\mathcal{C} := \{x_i : y_i f(x_i) > 0\}$ be the set of points correctly classified by f_θ. Then, we define the minimum ℓ_2-margin of f_θ to be $\min_{x_i \in \mathcal{C}} y_i f_\theta(x_i)$ and the average ℓ_2-margin of f_θ to be $\frac{1}{n}\sum_{i=1}^n y_i f_\theta(x_i)$. The hard margin SVM is a binary linear classifier formulated through the following optimization problem: $\min_{\theta, b} \frac{1}{2}||\theta||_2^2$ such that $y_i(x_i^\top \theta + b) \geq 1$ for all $(x_i, y_i) \in \mathcal{D}$. The classifier θ_{SVM} learned by a hard margin SVM is the separating hyperplane which maximizes the minimum ℓ_2-margin.

Soudry et al. (2018) show that linear classifiers trained via GD with the unregularized logistic loss converge in direction to θ_{SVM}. In particular, $||\frac{\theta_{\text{GD}}}{||\theta_{\text{GD}}||_2} - \frac{\theta_{\text{SVM}}}{||\theta_{\text{SVM}}||_2}||_2 \to 0$ at a worst-case rate of $O(\frac{\log \log t}{\log t})$ where t is the number of gradient steps. This result can also be extended to the nonseparable case, though, as one might expect, the implicit bias is more complex (Ji & Telgarsky,

[2] A slightly less precise formulation $\Sigma_A \to 0$ was studied by Papyan et al. (2020).



(a) Waterbirds (b) CelebA (c) CivilComments

Figure 3: **Convergence of LLR to the maximum-margin SVM solution is extremely slow.** We plot the mean and standard deviation over 3 experimental seeds of the directional error $\widehat{Err}$ between the last layer weights of a neural network model and an SVM (both trained on the features of the held-out set). We use a ResNet-18 for Waterbirds and CelebA and a BERT-Tiny for CivilComments. Here, $\widehat{Err} := ||\frac{\theta_{\text{NN}}}{||\theta_{\text{NN}}||_2} - \frac{\theta_{\text{SVM}}}{||\theta_{\text{SVM}}||_2}||_2$, where θ_{NN} denotes the last layer weights and θ_{SVM} denotes the weights of an SVM trained on the held-out set features.

2019). The SVM solution was recently studied in the context of group robustness by Chaudhuri et al. (2023), who showed that under certain conditions on the data distribution, θ_{SVM} on subsampled data converges to the classifier with optimal worst-group accuracy.

If LLR maximizes the minimum training margin and this results in a debiased classifier, then we should see a high correlation between the size of group minimum margin and group test accuracy (*i.e.,* minority group points lie near the decision boundary). However, we find that the minimum training margin for a group is not correlated with the test accuracy for that group. Interestingly, we find that *average* group training margin actually is correlated with group test accuracy. We display the Pearson correlation coefficients in Table 1 (deferred to Appendix A).

Moreover, convergence towards θ_{SVM} is slow in practice. We train a ResNet-18 and BERT-Tiny last layers on the held-out sets of Waterbirds, CelebA and CivilComments via SGD with learning rate 0.001 and unregularized logistic loss, and we plot the directional error with θ_{SVM} in Figure 3. In line with the slow $O(\frac{\log \log t}{\log t})$ rate, we find that even after $10\times$ the number of standard training steps, the LLR classifier has yet to converge to θ_{SVM}. As convergence to θ_{SVM} is slow and minimum ℓ_2-margin is uncorrelated with group test accuracy, we conclude that SVM convergence and minimum margin bounds are unlikely to explain LLR.

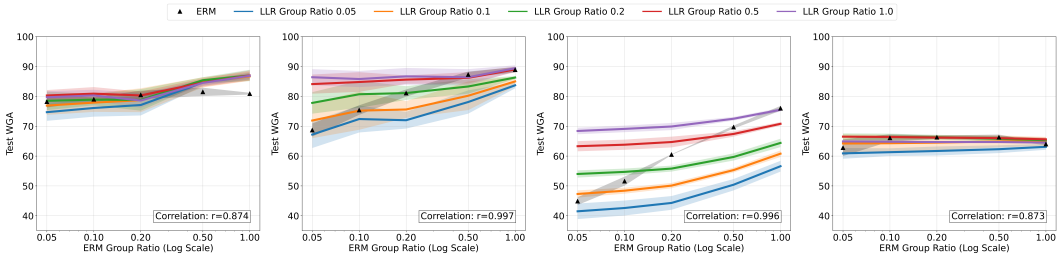
3 Group balance correlates strongly with LLR performance

The negative results of Section 2 prompted us to reconsider our initial hypothesis. We now present strong evidence for an alternative hypothesis: that the success of LLR is primarily due to *better group balance* in the held-out set than the training set. We also show how recent LLR algorithms owe their success to implicit group-balancing on the held-out set.

3.1 LLR performance is determined by held-out set group balance

To isolate the effect of group balance on LLR, we perform an ablation study, shown in Figure 4. We directly control the group ratios — defined as the ratio between the number of minority group points and the number of majority group points — by removing data until the desired group ratio is achieved for each class. We vary the group ratios for both the ERM training set and the LLR held-out set between 0.05 and 1.00, keeping the total data in each stage constant. For a fair comparison, the ERMs in Figure 4 are trained with the held-out set added in. For example, on Waterbirds we compare LLR with 20% of data following an ERM with 80% of data (colored line) to an ERM with 100% of data (gray line). Note that LLR with a group ratio of 1.00 corresponds to DFR.

We find that LLR performance is almost completely determined by the change in group balance between the training and held-out datasets. LLR only shows significant gains over ERM when it



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 4: **LLR performance is determined by held-out set group balance.** We compare the test WGA of ERM and LLR models while controlling the group balance of the training and held-out sets. We train a ResNet-50 for the vision datasets and a BERT-Base for the language datasets, and we plot the mean and standard deviation over 3 experimental seeds. A group ratio of 1.00 corresponds to the majority and minority groups having equal size, whereas a group ratio of 0.05 implies that the minority groups are 5% the size of the majority groups. We compute the Pearson correlation coefficient between the test WGA of ERM and the test WGA of LLR for each dataset while keeping group ratios fixed in the comparison. We find that the correlation coefficient is close to 1 for all datasets. In particular, LLR tends to improve over ERM *only if the held-out set has better group balance*.

is trained on a held-out set with a *higher group ratio* than the ERM training set, and vice versa. In Figure 4, we detail the Pearson correlation coefficients between ERM and LLR test WGA across all group ratios and observe a strong trend, with $r > 0.99$ on CelebA and CivilComments. Moreover, the test WGA of LLR trained with a fixed group ratio is nearly identical to the test WGA of ERM trained with that same group ratio (observed most impressively on CelebA).

These results indicate that the two-stage training procedure of LLR methods is not fundamentally oriented towards learning a debiased classifier as we initially hypothesized. Rather, some form of group-balancing is necessary to capture any of the robustness benefits achieved by DFR. Moreover, LLR methods are limited by how well they improve the group ratio, and thus DFR likely upper bounds the WGA of any LLR method which does not utilize group annotations.

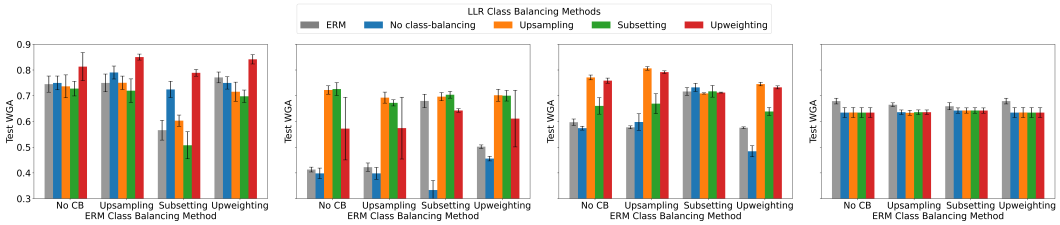
3.2 Recent LLR methods succeed via implicit group-balancing

In this section, we show how the recent LLR methods of class-balanced LLR (CB-LLR) (LaBonte et al., 2023) and automatic feature reweighting (AFR) (Qiu et al., 2023) perform *implicit group-balancing* to elicit a robustness improvement without group annotations.[3] Oftentimes, this is achieved by *class-balancing*.

We utilize three standard techniques for class-balancing: *subsetting*, wherein the size of each class is reduced to match the size of the smallest class; *upsampling*, wherein sampling probabilities for each class are adjusted so that the mini-batches are balanced in expectation; and *upweighting*, wherein minority class samples are assigned larger weights in the loss function (Idrissi et al., 2022; LaBonte et al., 2024). Since groups are defined with respect to classes, each of these class-balancing methods partially group-balance the dataset without requiring group annotations.

In Figure 5, we compare the test WGA of class-balanced ERM and class-balanced LLR, and we find that the improvement of LLR over ERM is dependent on the class-balancing methods used. We find optimally class-balanced ERM achieves nearly the same test WGA as optimally class-balanced LLR on all datasets! The CB-LLR results of LaBonte et al. (2023) use upsampling for ERM (as well as LLR), which on CelebA and CivilComments is suboptimal and explains why CB-LLR seems to work so well. On a positive note, this means that LLR can recover the WGA of an optimally class-balanced model even when ERM experiences catastrophic collapse.

[3] LaBonte et al. (2023) also propose SELF: unlike AFR and CB-LLR, it *explicitly* constructs a more group-balanced held-out set.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 5: **LLR does not significantly improve over optimally class-balanced ERM.** We plot the test WGA of ERM (trained on 100% of available data) and LLR (trained using a training/held-out set split) using no class-balancing, upsampling, upweighting, and subsetting. We train a ResNet-50 for the vision datasets and a BERT-Base for the language datasets and plot the mean and standard deviation over 3 experimental seeds. We find ERM matches the performance of LLR when ERM is trained with optimal class-balancing. Note that MultiNLI is class balanced *a priori*.

Compared to standard LLR, which utilizes the cross-entropy loss only, AFR incorporates a weighted loss function which prioritizes points upon which the ERM model performs poorly. Specifically, AFR introduces a weight for each held-out example i proportional to $\exp(-\gamma \hat{p}_i)$, where $\hat{p}_i$ is the probability for the correct class y_i and $\gamma \geq 0$ is a tunable temperature parameter.[4] Therefore, the AFR held-out set effectively has better group balance than the training set! Notably, the ablations of Qiu et al. (2023) show that setting $\gamma = 0$ reduces to CB-LLR with upweighting, thus explaining its improvement over an ERM with no class-balancing (see Figure 5).

4 Discussion

In this paper, we proposed and eventually rejected a hypothesis explaining the success of LLR through neural collapse and worst-case margin. However, we also presented strong evidence supporting an alternate hypothesis: that the improved WGA of LLR is primarily due to improved group balance in the held-out set. Overall, LLR remains an effective method to improve worst-group accuracy, especially if a limited number group annotations are available (*e.g.,* DFR), or in conjunction with implicit group-balancing techniques (*e.g.,* CB-LLR, SELF, and AFR).

Our work highlights the need for further research towards two open questions. First, how do neural networks learn both core and spurious features during ERM training? Second, what is the mechanism by which LLR with better group ratio than ERM reweights the core and spurious features?

Acknowledgments

We thank Jacob Abernethy for compute assistance. V.M. acknowledges support from the NSF (awards CCF-223915 and IIS-2212182), Adobe Research and Amazon Research.

[4]Loss-based adjustments are common among group robustness methods not using group annotations, *e.g.,* Liu et al. (2021).

REFERENCES

Sara Beery, Grant van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 2018.

Daniel Borkan, Lucas Dixon, Jeffrey Sorenson, Nithium Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *World Wide Web (WWW)*, 2019.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error? In *International Conference on Machine Learning (ICML)*, 2023.

Yongqiang Chen, Wei Huang, Kaiwen Zhou, Yatao Bian, Bo Han, and James Cheng. Understanding and improving feature learning for out-of-distribution generalization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Zhikang Chen, Min Zhang, Sen Cui, Haoxuan Li, Gang Niu, Mingming Gong, Changshui Zhang, and Kun Zhang. Neural collapse inspired feature alignment for out-of-distribution generalization. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *Conference on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*, 2019.

William Falcon and the PyTorch Lightning maintainers and contributors. PyTorch Lightning. *GitHub*, 2019.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020.

X.Y. Han, Vardan Papyan, and David Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations (ICLR)*, 2022.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(1):357–362, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

John D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007.

Badr Youbi Idrissi, Martín Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning (CLeaR)*, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Ziwei Ji and Matus Telgarsky. The implicit bias of gradient descent on nonseparable data. In *Conference on Learning Theory (COLT)*, 2019.

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning (ICML)*, 2021.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *Transactions on Machine Learning Research (TMLR)*, 2023.

Tyler LaBonte. Milkshake: Quick and extendable experimentation with classification models. http://github.com/tmlabonte/milkshake, 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.

Tyler LaBonte, John C. Hill, Xinchen Zhang, Vidya Muthukumar, and Abhishek Kumar. The group robustness is in the details: Revisiting finetuning under spurious correlations. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Evan Zheran 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)*, 2021.

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

Shenyu Lu, Junyi Chai, and Xiaoqian Wang. Neural collapse inspired debiased representation learning for min-max fairness. In *Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

Vardan Papyan, X.Y. Han, and David Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (40):24652–24663, 2020.

Juhyeon Park, Seokhyeon Jeong, and Taesup Moon. TLDR: text based last-layer retraining for debiasing image classifiers. In *Winter Conference on Applications of Computer Vision (WACV)*, 2025.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *Conference on Neural Information Processing Systems (NeurIPS) Workshop on Automatic Differentiation*, 2017.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence of spurious correlations. In *International Conference on Machine Learning (ICML)*, 2024.

Shikai Qiu, Andres Potapczynski, Pavel Izmailov, and Andrew Gordon Wilson. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning (ICML)*, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(1): 211–252, 2015.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Ravid Shwartz-Ziv, Micah Goldblum, Yucen Lily Li, C Bayan Bruss, and Andrew Gordon Wilson. Simplifying neural network training under class imbalance. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research (JMLR)*, 19: 1–57, 2018.

Nathan Stromberg, Rohan Ayyagari, Monica Welfert, Sanmi Koyejo, and Lalitha Sankar. Robustness to subpopulation shift with domain label noise via regularized annotation of domains. *Transactions on Machine Learning Research (TMLR)*, 2024.

Saeid Asgari Taghanaki, Aliasghar Khani, Fereshte Khani, Ali Gholami, Linh Trana, Ali Mahdavi-Amiri, and Ghassan Hamarneh. MaskTune: mitigating spurious correlations by forcing to explore. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

TorchVision maintainers and contributors. TorchVision: PyTorch's computer vision library. *GitHub*, 2016.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6): 86–93, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Yining Wang, Junjie Sun, Chenyue Wang, Mi Zhang, and Min Yang. Navigate beyond shortcuts: Debiased learning through the lens of neural collapse. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Monica Welfert, Nathan Stromberg, and Lalitha Sankar. Theoretical guarantees of data augmented last layer retraining methods. In *International Symposium on Information Theory (ISIT)*, 2024.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Technical report, California Institute of Technology, 2010.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Association for Computational Linguistics (NAACL)*, 2018.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP) System Demonstrations*, 2020.

Jingxuan Xu, Wuyang Chen, Linyi Li, Yao Zhao, and Yunchao Wei. Collapsed language models promote fairness. In *International Conference on Learning Representations (ICLR)*, 2025.

Haotian Ye, James Zou, and Linjun Zhang. Freeze then train: Towards provable representation learning under spurious correlations and feature noise. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15, 2018.

Michael Zhang, Nimit S. Sohoni, Hongyang R. Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning (ICML)*, 2022.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *International Conference on Computer Vision (ICCV)*, 2015.

A Additional Experiments

In this section, we present additional experiments deferred from the main body of the work. First, in Table 1, we include the Pearson correlation coefficient between the test accuracy on each group and its minimum/average margin, calculated as defined in Section 2.2. We find that the minimum training margin on a group is surprisingly uncorrelated with the test accuracy on that group; instead, the average training margin on a group is more correlated with its test accuracy.

Table 1: **Correlation between types of training margins and group test accuracy.** We compute the Pearson correlation coefficient between the test group accuracy of a ResNet-18 (BERT-Tiny) trained on Waterbirds and CelebA (CivilComments and MultiNLI) and the minimum/average margin of each group across 3 experimental seeds. We see that average training margin is more correlated in general with group test accuracy than minimum training margin. We leave an explanation of the predictive power of average margin for future work.

(a) **Waterbirds**

Margin Metric	Group 0	Group 1	Group 2	Group 3
Minimum Margin	0.349	0.380	-0.026	0.282
Average Margin	-0.418	0.585	0.662	-0.185

(b) **CelebA**

Margin Metric	Group 0	Group 1	Group 2	Group 3
Minimum Margin	0.245	0.326	0.385	0.370
Average Margin	0.469	0.290	0.475	0.495

(c) **CivilComments**

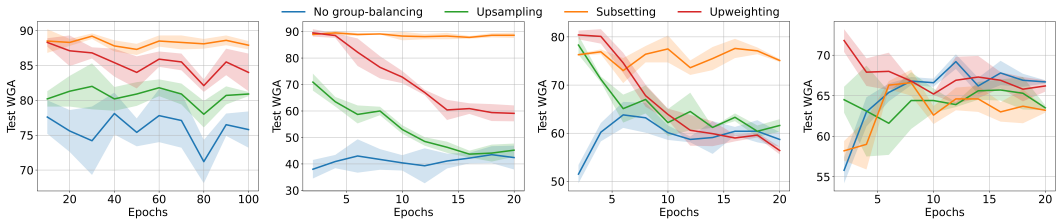
Margin Metric	Group 0	Group 1	Group 2	Group 3
Minimum Margin	0.214	0.237	0.227	0.113
Average Margin	0.209	0.271	0.298	0.247

(d) **MultiNLI**

Margin Metric	Group 0	Group 1	Group 2	Group 3	Group 4	Group 5
Minimum Margin	0.574	0.266	0.027	0.726	0.492	0.591
Average Margin	0.865	0.926	0.858	0.910	0.823	0.897

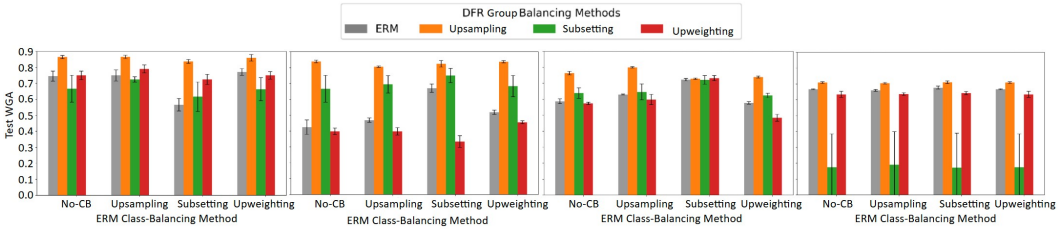
Second, we study an extension of the *implicit* group-balancing discussion from Section 3.2 to *explicit* group balancing. As discussed in Section 3.2, we investigate three explicit balancing methods: subsetting, upsampling, and upweighting. It has previously been observed that using upsampling or upweighting for class-balancing can result in drastic decreases to test WGA during training (LaBonte et al., 2024). If the performance improvement from LLR is indeed primarily due to the implicit group-balancing achieved by class-balancing, we would expect similar collapse when we explicitly group-balance using these same techniques.

Figure 6 shows the effects that different group-balancing methods have on WGA during training. We find that similarly to class-balancing, upsampling and upweighting lead to catastrophic collapse on CelebA and CivilComments. Additionally, group-balanced upweighting leads to a catastrophic collapse on MultiNLI and a mild decrease in WGA on Waterbirds. Overall, the training dynamics which result from the choice of subsetting, upweighting, and upsampling for group-balancing are quite similar to the dynamics observed by LaBonte et al. (2024) for class-balancing. The notable exceptions are MultiNLI, which is class balanced *a priori* but not group balanced, and subsetting on Waterbirds. Class-balanced subsetting on Waterbirds overly penalizes the minority group in the majority class which is problematic due to Waterbirds' small size. Group-balanced subsetting does not have this issue and hence performs very well on Waterbirds.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 6: **Group-balancing with upsampling and upweighting can lead to catastrophic collapse.** We compare three group-balancing methods: *subsetting*, *upsampling*, and *upweighting*. We plot the mean and standard deviation over 3 experimental seeds for a ResNet-50 on the vision datasets and a BERT-Base on the language datasets. We note a dramatic decrease in test WGA during training for both upsampling and upweighting on CelebA and CivilComments. We also see a collapse in WGA for upweighting on MultiNLI.



(a) Waterbirds (b) CelebA (c) CivilComments (d) MultiNLI

Figure 7: **Choice of group-balancing method is important to DFR success.** We compare class-balanced ERM to group-balanced DFR across three balancing methods: *subsetting*, *upsampling*, and *upweighting*. We plot the mean and standard deviation over 3 experimental seeds for a ResNet-50 on the vision datasets and a BERT-Base on the language datasets. We note that the choice of balancing method has a dramatic effect on the test WGA for both ERM and DFR. Two seeds of MultiNLI subsetting failed, achieving 0% accuracy on the minority group, leading to high variance.

In Figure 7, we compare the test WGA of class-balanced ERM and group-balanced DFR. We observe that the choice of balancing method has a large effect on the success of both ERM and DFR. However, optimally group-balanced DFR improves over optimally class-balanced ERM on all datasets. This is not the case for LLR, highlighting the value of explicit group-balancing when annotations are available.

B Additional Dataset and Training Details

B.1 Dataset Details

We study four benchmarks for group robustness across vision and language tasks, outlined below and detailed in Table 2. Note that Waterbirds is the only dataset that has a distribution shift and MultiNLI is the only dataset which is class-balanced *a priori*.

- *Waterbirds* (Welinder et al., 2010; Wah et al., 2011; Sagawa et al., 2020) is an image dataset where birds are classified into land species ("landbirds") or water species ("waterbirds"). The spurious feature is the image background: landbirds are more frequently associated with land backgrounds, and waterbirds are more often found with water backgrounds.[5]

[5]It is worth noting that the Waterbirds dataset contains incorrect labels (Taghanaki et al., 2022). We report results based on the original, uncorrected version, as is customary in the literature.

- *CelebA* (Liu et al., 2015; Sagawa et al., 2020) is an image dataset where celebrities are categorized as either blond or non-blond. The spurious feature is gender, with a 16× greater number of blond women than blond men in the training data.
- *CivilComments* (Borkan et al., 2019; Koh et al., 2021) is a language dataset where online comments are classified as either toxic or non-toxic. The spurious feature involves the presence of one of the following identity categories: male, female, LGBT, black, white, Christian, Muslim, or other religion.[6] Toxic comments tend to contain one of these identity categories more often than non-toxic comments, and vice versa.
- *MultiNLI* (Williams et al., 2018; Sagawa et al., 2020) is a language dataset where pairs of sentences are classified as a contradiction, entailment, or neither. The spurious feature is the presence of a negation in the second sentence — contradictions tend to have negations more often than entailments or neutral pairs.

Table 2: **Dataset composition.** The class probabilities change dramatically when conditioned on the spurious feature. Probabilities may not sum to 1 due to rounding.

Dataset	Group g		Training distribution $\hat{p}$			Data quantity		
	Class y	Spurious s	$\hat{p}(y)$	$\hat{p}(g)$	$\hat{p}(y\|s)$	Train	Val	Test
Waterbirds	landbird	land	.768	.730	.984	3498	467	2225
	landbird	water		.038	.148	184	466	2225
	waterbird	land	.232	.012	.016	56	133	642
	waterbird	water		.220	.852	1057	133	642
CelebA	non-blond	female	.851	.440	.758	71629	8535	9767
	non-blond	male		.411	.980	66874	8276	7535
	blond	female	.149	.141	.242	22880	2874	2480
	blond	male		.009	.020	1387	182	180
CivilComments	neutral	no identity	.887	.551	.921	148186	25159	74780
	neutral	identity		.336	.836	90337	14966	43778
	toxic	no identity	.113	.047	.079	12731	2111	6455
	toxic	identity		.066	.164	17784	2944	8769
MultiNLI	contradiction	no negation	.333	.279	.300	57498	22814	34597
	contradiction	negation		.054	.761	11158	4634	6655
	entailment	no negation	.334	.327	.352	67376	26949	40496
	entailment	negation		.007	.104	1521	613	886
	neither	no negation	.333	.323	.348	66630	26655	39930
	neither	negation		.010	.136	1992	797	1148

B.2 Training Details

We utilize ResNet-18 and ResNet-50 (He et al., 2016) models pretrained on ImageNet-1K (Russakovsky et al., 2015) for Waterbirds and CelebA. We also utilize a BERT (Devlin et al., 2019) model pretrained on Book Corpus (Zhu et al., 2015) and English Wikipedia for CivilComments and MultiNLI. These pretrained models are used as the initialization for ERM finetuning under the cross-entropy loss. We use standard ImageNet normalization with standard flip and crop data augmentation for the vision tasks and BERT tokenization for the language tasks (Izmailov et al., 2022). Our implementation uses the following packages: `NumPy` (Harris et al., 2020), `PyTorch` (Paszke et al., 2017; 2019), `Lightning` (Falcon & the PyTorch Lightning maintainers and contributors, 2019), `TorchVision` (TorchVision maintainers and contributors, 2016), `Matplotlib` (Hunter, 2007), `Transformers` (Wolf et al., 2020), and `Milkshake` (LaBonte, 2023).

[6]This version of CivilComments includes four identity groups, as used in this work and by Sagawa et al. (2020); Idrissi et al. (2022); Izmailov et al. (2022); Kirichenko et al. (2023); LaBonte et al. (2023). There exists another version where identity categories are not merged into one spurious feature; that version is employed by Liu et al. (2021); Zhang et al. (2022); Qiu et al. (2023). Both versions use the WILDS split (Koh et al., 2021).

To our knowledge, the licenses of Waterbirds and CelebA are unknown. CivilComments is released under the CC0 license, and information about MultiNLI's license may be found in Williams et al. (2018).

Our experiments were conducted on two local 24GB Nvidia RTX A5000 GPUs. We provide our ERM finetuning hyperparameters in Table 3. Our LLR experiments were run for the same number of epochs as ERM on a held-out dataset (20% of the training set for Waterbirds and half the validation set for the other three datasets); we used SGD with learning rate 0.01 with no weight decay or learning rate schedule.

Our code is available at `https://github.com/tmlabonte/understanding-llr`.

Table 3: **Default ERM and LLR hyperparameters.**

Dataset	Optimizer	Initial LR	LR schedule	Batch size	Weight decay	Epochs
Waterbirds	AdamW	1×10^{-5}	Cosine	32	1×10^{-4}	100
CelebA	AdamW	1×10^{-5}	Cosine	32	1×10^{-4}	20
CivilComments	AdamW	1×10^{-5}	Linear	32	1×10^{-4}	20
MultiNLI	AdamW	1×10^{-5}	Linear	32	1×10^{-4}	20