Improving class and group imbalanced classification with uncertainty-based active learning

Alexandru Tifrea* TIFREAA@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

John Hill* JHILL326@GATECH.EDU
Department of Computer Science, Georgia Institute of Technology

Fanny Yang FAN.YANG@INF.ETHZ.CH
Department of Computer Science, ETH Zurich

Abstract

Recent experimental and theoretical analyses have revealed that uncertainty-based active learning algorithms (U-AL) are often not able to improve the average accuracy compared to even the simple baseline of passive learning (PL). However, we show in this work that U-AL is a competitive method in problems with severe data imbalance, when instead of the *average* accuracy, the focus is the *worst-subpopulation* accuracy. We show in extensive experiments that U-AL outperforms algorithms that explicitly aim to improve worst-subpopulation performance such as reweighting. We provide insights that explain the good performance of U-AL and show a theoretical result that is supported by our experimental observations.

Keywords: Active learning, Imbalanced data, Worst-group performance

1. Introduction

In applications where data labeling is prohibitively expensive, it is important to deploy specialized algorithms that can inform which are the most useful samples for which to acquire labels. The active learning (AL) literature proposes numerous algorithms tailored for this task, such as *uncertainty sampling*, one of the most extensively analyzed AL algorithms (Settles, 2009; Balcan et al., 2007; Chaudhuri et al., 2015). Despite its theoretical guarantees, uncertainty-based active learning (U-AL) has been shown to not significantly outperform the naive baseline of passive learning (PL), especially in the context of deep learning on image or text data (Sener and Savarese, 2018; Hacohen et al., 2022). In addition, recent theoretical analyses have revealed that U-AL performs on par or even worse than PL in high-noise or high-dimensional settings (Mussmann and Liang, 2018; Tifrea et al., 2023).

Despite the known shortcomings of U-AL, we show in this work that this strategy performs significantly better than PL and other AL baselines in imbalanced classification settings, where either a class or a group within a class are underrepresented in the training data. Experiments on several datasets with class or group imbalance show that this phenomenon occurs for a broad set of real-world data distributions. While a similar observation has been made for uncertainty sampling applied to linear models in class imbalance problems (Ertekin et al., 2007), to the best of our knowledge, this is the first work to document the advantages of U-AL for *group imbalance* and for deep learning models. Furthermore, in

Section 3.2.1 we provide a theoretical result that provides insights into this phenomenon, which we then verify experimentally in Section 3.2.2. Finally, perhaps surprisingly, our experiments show that methods that try to combine representativeness and informativeness (e.g. epsilon-greedy style approaches (Brinker, 2003; Huang et al., 2014; Yang et al., 2015)) hurt worst-group performance, despite often showing a significant improvement in average accuracy (Shui et al., 2020; Ash et al., 2020).

Our results show that U-AL, while not specifically designed for imbalanced data, constitutes a strong active learning baseline for improving worst-subpopulation performance. Thus, U-AL becomes one of the few methods proposed in the context of the underexplored problem of fair active learning (Anahideh et al., 2021).

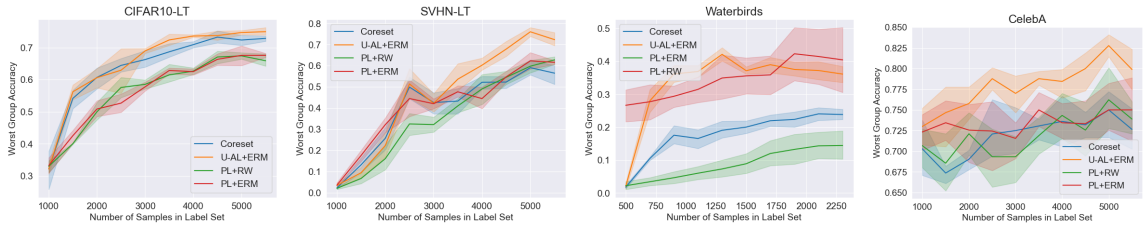
2. Problem setting: Active learning for imbalanced classification

The overarching goal in this work is to achieve good classification performance with a limited labeling budget, in situations where the data contains severely underrepresented subpopulations. In this section we elaborate on the details of the problem setting.

Active learning for classification. We consider settings where there exists a constraint on how much labeled data can be collected to serve as the training set for a classification model. In this work we focus on the standard pool-based active learning setting, where a large unlabeled dataset is available. In this active learning framework, a sampling algorithm can choose to query an oracle to obtain the label for one or more of the unlabeled data points. This process is repeated iteratively until the label budget is exhausted. The strategy employed for collecting the labeled set determines the predictive performance of the classification model trained on this dataset. A naïve baseline for assessing the quality of an active learning algorithm is *passive learning (PL)*, which consists of simply collecting the labeled data by sampling uniformly at random from the unlabeled dataset.

Uncertainty-based active learning. In this work we study uncertainty-based active learning (U-AL). As shown in Algorithm 1, this strategy trains a classification model on the currently available labeled data and queries the label of points on which the model has highest predictive uncertainty. There are numerous methods that implement a variant of this strategy, for instance, by acquiring labels where the softmax layer of a deep neural network has low confidence or querying the label of unlabeled points that lie closest to the decision boundary of the classifier trained on the currently available data (Ducoffe and Precioso, 2018). The success of this simple strategy has been questioned numerous times in the active learning literature (Mussmann and Liang, 2018; Tifrea et al., 2023). In particular, several works argue that in the context of deep learning, U-AL is not better than PL for a broad set of image and text datasets (Sener and Savarese, 2018; Hacohen et al., 2022).

Informativeness vs. representativeness-based sampling. A common explanation for the failure of U-AL for deep neural networks is that the resulting labeled dataset lacks diversity, since U-AL greedily selects the most informative samples at each sampling iteration. To mitigate this shortcoming, several prior works propose to combine the *informativeness-based* sampling of U-AL with another sampling objective aimed at choosing points that are more representative of the overall population. In fact, for deep neural networks, the best average accuracy with a limited query budget is often achieved by methods that combine an informativeness and a representativeness sampling objective, e.g. Brinker (2003); Huang et al. (2014); Gissin and Shalev-Shwartz (2019); Shui et al. (2020).



(a) CIFAR10-LT. (b) SVHN-LT. (c) Waterbirds. (d) CelebA.

Figure 1: U-AL achieves higher worst-group accuracy than PL+ERM and even PL+RW, which explicitly targets improving this metric. Furthermore, U-AL performs consistently better or on par compared to another strong AL baseline, coreset-AL.

There are several ways in which one can combine these two often competing objectives. Assuming that the unlabeled set is drawn i.i.d. from the data distribution of interest, then a straightforward *representativeness-based* sampling procedure is uniform sampling from the unlabeled set, namely passive learning. In this case, one can combine informativeness and representativeness, for instance, by means of an epsilon-greedy U-AL algorithm where with probability ϵ the next sample is selected via PL, and with probability $1-\epsilon$ it is chosen via U-AL. More sophisticated representativeness-based algorithms include coreset-based sampling (Sener and Savarese, 2018) or TypiClust (Hacohen et al., 2022).

Data with imbalanced subpopulations. Most of the prior experimental and theoretical analyses of U-AL study the average-case accuracy in settings where there is no severe underrepresentation in the data distribution. In contrast, in this work we focus on problems with a pronounced imbalance in the data. We consider both *class imbalance* (i.e. one or more classes have many fewer samples than other classes) and *group imbalance* (e.g. a class consists of data from two groups, with one of them significantly overrepresented compared to the other). Moreover, in addition to the average-case accuracy, we are also interested in the *worst-subpopulation* accuracy, namely the worst-case accuracy among all the classes or groups, for class and group imbalance, respectively.

Both the class and group imbalanced problem have been studied extensively in the supervised learning literature (Sagawa et al., 2020a; Kini et al., 2021), but it remains not well understood what active learning strategy is more suitable for deep neural networks that operate in these settings.

3. The benefits of U-AL for imbalanced classification

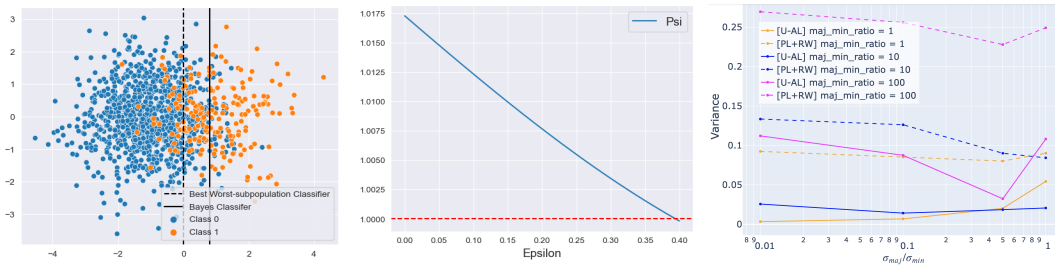
In this section we show that U-AL can significantly improve the worst-subpopulation accuracy of deep neural networks on several image classification datasets. We provide an explanation for this phenomenon and explore it further in several ablation studies.

3.1 Experimental study

We begin by providing experimental evidence that U-AL improves the worst-subpopulation accuracy compared to several commonly considered active and passive learning baselines.

3.1.1 Experiment details

Datasets. We consider both class and group imbalanced image dataset. The CIFAR10-LT and SVHN-LT datasets are a variant of CIFAR10 and SVHN, respectively, that are used in the class imbalanced literature (Cui et al., 2019). To construct these datasets, 5 of the classes have been subsampled to be 10% of the size of the remaining 5 classes. In



(a) Imbalanced classification. (b) $\Psi(\epsilon)$. (c) Estimator variance.

Figure 2: a) In the imbalanced case, the optimal classifier (wrt average accuracy) is shifted towards the mean of the minority class. Thus, the proportion of minority points near the decision boundary is higher than the proportion in the entire distribution. b) Function $\Psi(\epsilon)$ that appears in Proposition 1. Our lower bound on $\mathbb{E}[\eta_s]$ exceeds the original η for all $\epsilon < 0.38$. c) The variance of the PL+RW estimator is larger than for ERM trained on the dataset collected with U-AL.

addition to these datasets, we also consider CelebA and Waterbirds, two datasets which contain extremely underrepresented groups determined by a binary attribute (i.e. gender or background, respectively) (Sagawa et al., 2020a,b).

Baselines. We consider the following baselines to compare with U-AL (see Appendix 7 for more details). Collecting the data with PL and then running ERM for prediction (PL+ERM) as well as the coreset-AL method (Sener and Savarese, 2018) do not explicitly target underrepresented subpopulations. In contrast, reweighting (PL+RW) (Sagawa et al., 2020b) aims to improve the worst-subpopulation performance by assigning a larger weight in the training loss to samples from minority groups.

3.1.2 Main experimental results

Figure 1 reveals that U-AL improves significantly the worst-group and worst-class accuracy compared to PL+ERM for all datasets. Moreover, even when using an explicit mitigation to improve worst-group performance, such as reweighting (i.e. PL+RW), passive learning is still surpassed by U-AL. We discuss the potential cause of this gap in worst-subpopulation performance between U-AL in PL+RW in Section 3.3. Additionally, we see U-AL outperforms coreset-AL in worst-group performance as well. Finally, we note that, as shown in Appendix 7, the improvement in average accuracy induced by U-AL is almost always less significant than the improvement in worst-group accuracy.

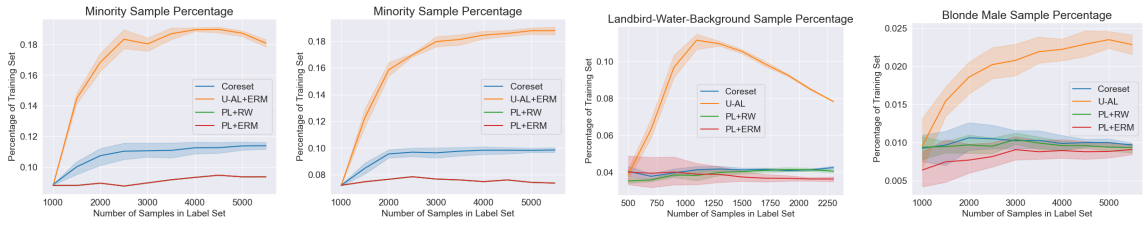
3.2 Intuitive explanation: U-AL collects a more balanced labeled set

In this section we investigate the cause that leads to the good performance of U-AL in imbalanced data settings. In what follows, we argue theoretically and experimentally that U-AL collects a more balanced labeled dataset for both class and group imbalance, and hence, mitigates the problems that stem from data imbalance.

3.2.1 Theoretical analysis for class imbalance

To understand why U-AL tends to perform well on settings with imbalanced data, we consider a simple binary classification problem with class imbalance. Later, we explain how the insights transfer from this setting to more general class and group imbalance problems.

Consider a distribution D with binary labels $y \in \{0, 1\}$, where the features $x \in \mathbb{R}^d$ are drawn from Gaussian class-conditional distributions with shared covariance matrix, Σ.



(a) CIFAR10-LT. (b) SVHN-LT. (c) Waterbirds. (d) CelebA.

Figure 3: In the case of both class and group imbalance, we see that U-AL samples minority points at rates far higher than PL and coreset-AL.

More specifically, $x \sim \mathcal{N}(\mu_0, \Sigma)$ with probability $p = \frac{1}{\eta+1}$ and $x \sim \mathcal{N}(\mu_1, \Sigma)$ with probability $1 - p = \frac{\eta}{\eta+1}$, where η denotes the population imbalance ratio between the two classes and μ_0 and μ_1 are the means of the two class-conditional distributions.

We focus on this setting, since it accepts a linear Bayes-optimal classifier, making it more amenable for a theoretical analysis. The following proposition shows a bound on the expected imbalance ratio $\mathbb{E}[\eta_s]$ of the labeled set collected after running one iteration of uncertainty sampling (Algorithm 1) to label a batch of n_q samples.

Proposition 1. *Let L_n be the empirical logistic loss function for a labeled dataset $D_n = \{(x_i, y_i)\}_{i=1}^n$ made up of n i.i.d. samples from the same distribution D as defined above with imbalance ratio $\eta \in (0, 1]$. Also, assume that n is sufficiently large. Let $f_\theta : \mathbb{R}^d \to \{0, 1\}$ and let $\hat{\theta} = \arg\min_\theta L_n(f_\theta(X), Y)$. Assume that we label a batch of n_q points from D which are at most β away from the decision boundary parameterized by $\hat{\theta}$ in Euclidean distance. There exists a function $\Psi(\epsilon)$, such that for large enough n, n_q and small enough η, ϵ and β we have that $\Psi(\epsilon) \geq 1$ and with high probability,*

$$\mathbb{E}[\eta_s] \geq \eta \cdot \Psi(\epsilon). \tag{1}$$

The formal statement and the proof of this result can be found in Appendix 3.4. The proposition implies that, under some assumptions, the expected imbalance ratio of the dataset collected with U-AL is closer to 1 (and hence the dataset is more balanced) compared to the original $\eta \ll 1$. Ertekin et al. (2007) allude to a similar phenomenon to the one captured in Proposition 1, without a clear and rigorous justification. We plot function Ψ in Figure 2b, which reveals that for sufficiently small ϵ the expected imbalance ratio of the collected labeled set will be closer to 1 (and hence, more balanced) than the expected imbalance ratio of PL, η.

At the core of this result lies the observation that an overrepresented class is likely to contain more outliers than an underrepresented class. This, in turn, leads to the decision boundary being shifted towards the minority class, like in Figure 2. This insight holds more generally than for linear classifiers and GMM data. In fact, a concurrent work (Chaudhuri et al., 2023) shows that for a broad family of class-conditional distributions, decision boundaries are shifted towards the minority subpopulation for both class and group imbalance.

3.2.2 Empirical verification of theoretical insights

We now check whether the insight from Theorem 1 that U-AL selects a more balanced labeled set than passive learning holds in more general settings. Indeed, Figure 3 shows that for both data with class and group imbalance, U-AL improves the imbalance ratio of the collected training set. Furthermore, in Appendix 7 we study the impact of using a less

biased classifier for sampling the labeled set for active learning. As implied by Theorem 1, a less biased classifier will *not* improve the imbalance ratio of the collected data as well as a more biased classifier.

Finally, we note that more complicated uncertainty metrics can also be used to improve the imbalance ratio substantially. We demonstrate this result in Appendix Section 3.4 through the inclusion of BADGE (Ash et al., 2020), an active learning algorithm which measures uncertainty as the gradient magnitude with respect to parameters in the final network layer.

3.3 Balanced data is better than a balancing algorithm

Figure 1 shows that U-AL improves the worst-subpopulation accuracy compared to not only PL where the classifier is trained with ERM, but also when a reweighted objective is used for training (i.e. PL+RW). The reason for this gap lies with the different consequences of either changing the data or changing the learning algorithm in order to improve worst-case performance. Francazi et al. (2023) show that for (passive) supervised learning, reweighting increases the statistical variance of the resulting estimator compared to resampling (i.e. subsampling the majority subpopulations). Intuitively, reweighting the few samples of the minority subpopulation amplifies not only the signal, but also the noise in that small subset of the data. In contrast, U-AL selects more data from the underrepresented populations, which is a mitigation that does not increase the variance significantly.

Indeed, we can see this trend when comparing the variance of ERM trained on data collected via U-AL (i.e. U-AL+ERM) and the variance of a reweighted classifier trained on a uniformly sampled training set (i.e. PL+RW). For simplicity, we consider a 1D problem, with data drawn from a 1D mixture of 2 Gaussians, where the imbalance ratio and the ratio between the standard deviations of the class-conditional normal distributions are varied. Figure 2c shows that the variance of U-AL+ERM is significantly lower than for PL+RW, and that the gap grows the larger the imbalance ratio. Finally, note that reweighting can also be used alongside U-AL (instead of ERM) to further improve worst-group accuracy.

3.4 Representativeness-based AL hurts worst-subpopulation accuracy

Several recent works suggest that striking a balance between informativeness and representativeness is necessary for the success of AL for deep learning (Gissin and Shalev-Shwartz, 2019; Ash et al., 2020; Shui et al., 2020). However, if data is imbalanced and the goal is worst-subpopulation accuracy, we see that employing a representativeness-based objective leads to collecting a more imbalanced labeled set, which in turn induces poorer performance.

To show this, we consider an ϵ-greedy strategy, in which at every round, an ϵ fraction of the samples are selected using uniform sampling, while the remaining points are chosen with uncertainty sampling. Appendix 7 indicates that a larger fraction of points sampled with the representativeness-based objective, leads to greater imbalance in the collected labeled set and consequently, hurts the worst-group performance. We expect that a similar trend occurs as well for methods like BADGE (Ash et al., 2020) which implicitly interpolate between a diversity-inducing objective and uncertainty sampling.

References

Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning, 2021.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds, 2020.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In Nader H. Bshouty and Claudio Gentile, editors, *Learning Theory*, pages 35–50, 2007.

Klaus Brinker. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning*, 2003.

Kamalika Chaudhuri, Sham M Kakade, Praneeth Netrapalli, and Sujay Sanghavi. Convergence rates of active learning for maximum likelihood estimation. *Proceedings in Advances in Neural Information Processing Systems 28*, 2015.

Kamalika Chaudhuri, Kartik Ahuja, Martin Arjovsky, and David Lopez-Paz. Why does throwing away data improve worst-group error?, 2023.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples, 2019.

Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

Seyda Ertekin, Jian Huang, Leon Bottou, and Lee Giles. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, 2007.

Emanuele Francazi, Marco Baity-Jesi, and Aurelien Lucchi. A theoretical analysis of the learning dynamics under class imbalance, 2023.

Daniel Gissin and Shai Shalev-Shwartz. Discriminative active learning. *arXiv preprint arXiv:1907.06347*, 2019.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

Sheng-Jun Huang, Rong Jin, and Zhi-Hua Zhou. Active learning by querying informative and representative examples. *Proceedings in Advances in Neural Information Processing Systems 27*, 2014.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In *Advances in Neural Information Processing Systems*, 2021.

Stephen Mussmann and Percy Liang. On the relationship between data efficiency and error for uncertainty sampling. In *Proceedings of the 34th International Conference on Machine Learning*, 2018.

Dmitrii Ostrovskii and Francis Bach. Finite-sample analysis of m-estimators using self-concordance, 2020.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *Proceedings of the 8th International Conference on Learning Representations*, 2020a.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations, 2020b.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *Proceedings of the 6th International Conference on Learning Representations*, 2018.

Burr Settles. Active learning literature survey. 2009.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, 2020.

Alexandru Tifrea, Jacob Clarysse, and Fanny Yang. Margin-based sampling in high dimensions: When being active is less efficient than staying passive, 2023.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 2015.

Appendix A.

In this appendix we provide a rigorous proof of Proposition 1 from Section 3.2.1. We begin by presenting the pseudocode for the U-AL procedure that we analyze in this work.

Algorithm 1: Uncertainty-based AL (U-AL)

Input: Seed set $\mathcal{D}_{seed}$, unlabeled set $\mathcal{D}_u$, budget n_ℓ, uncertainty score u, loss function $\mathcal{L}$
Result: Prediction model $\hat{f}$

$\mathcal{D}_\ell \leftarrow \mathcal{D}_{seed}$
for $n \in \{|\mathcal{D}_{seed}| + 1, ..., n_\ell\}$ **do**
  $\hat{f} \leftarrow \arg\min_f \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y)\in\mathcal{D}_\ell} \mathcal{L}(f, y)$
  $x_{\text{query}} \leftarrow \arg\max_{x\in\mathcal{D}_u} u(x; \hat{f})$
  $y_{\text{query}} \leftarrow \mathit{AcquireLabel}(x_{\text{query}})$
  $\mathcal{D}_\ell \leftarrow \mathcal{D}_\ell \cup \{(x_{\text{query}}, y_{\text{query}})\}; \mathcal{D}_u \leftarrow \mathcal{D}_u \setminus \{x_{\text{query}}\}$
return $\arg\min_f \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y)\in\mathcal{D}_\ell} \mathcal{L}(f, y)$

We now define formally our chosen setting. We consider binary logistic regression classifiers in 2-dimensions, but it should be noted that we make no major use of this assumption and a similar result holds for arbitrary dimension. Assume that our data $x \in \mathbb{R}^2$ is sampled from a distribution D such that only the first component of each sample corresponds to a signal dimension. The remaining dimension is isotropic Gaussian noise.

The signal component of our data is distributed according to two Gaussians with means $\mu \in \{\mu_0, \mu_1\}$ and standard deviation $\sigma > 0$, corresponding to the two classes with labels $y \in \{0, 1\}$. $x \sim \mathcal{N}(\mu_0, \Sigma)$ with probability $p = \frac{1}{\eta+1}$ and $x \sim \mathcal{N}(\mu_1, \Sigma)$ with probability $1 - p = \frac{\eta}{\eta+1}$, where η denotes the population imbalance ratio between the two classes.

We define our model, $f_\theta : \mathbb{R}^3 \to \{0, 1\}$, as

$$f_\theta(\tilde{x}) = \begin{cases} 1 & \text{if } \mathbf{S}(\theta^\top \tilde{x}) = \mathbf{S}\left(w^\top x + b\right) \geq 0.5 \\ 0 & \text{if } \mathbf{S}(\theta^\top \tilde{x}) = \mathbf{S}\left(w^\top x + b\right) < 0.5 \end{cases} \tag{2}$$

where $\theta = [w, b]^\top$, $\tilde{x} = [x, 1]^\top$, and $\mathbf{S}$ denotes the sigmoid function.

Let $L^*(\theta) = \mathbb{E}_{x\sim D}[L(f_\theta(x), y)]$ be the expected logistic loss of classifier θ over data points sampled from our distribution. Then, let $\theta^* = \arg\min_\theta L^*(\theta)$. $\theta^* = [w^*, b^*]^\top$ and it is well known that in the case of data sampled from two gaussians sharing the same covariance matrix, the optimal logistic regression classifier has the following closed form: $w^* = \Sigma^{-1}(\vec{\mu}_1 - \vec{\mu}_0)$ and $b^* = -\frac{1}{2} w^{*\top}(\vec{\mu}_1 + \vec{\mu}_0) + \log(\eta)$.

Now, let L_n be the logistic loss function for a dataset D_n made up of n i.i.d. samples from our distribution D. Also, let $\hat{\theta} = \arg\min_\theta L_n(f_\theta(X), Y)$. Applying Theorems 1.1 and 1.2 from (Ostrovskii and Bach, 2020) we know that with probability $1 - \delta$,

$$||\hat{\theta} - \theta^*||_2 \leq \epsilon \tag{3}$$

for $\epsilon, \delta > 0$ and $n = O(\frac{\log(1/\delta)}{\epsilon^2})$.

We know that $\hat{\theta} = [\hat{w}, \hat{b}]^\top$, so we get

$$|\hat{w}_1 - w_1^*| = ||\hat{w}_1 - w_1^*||_2 \leq ||\hat{\theta} - \theta^*||_2 \leq \epsilon \tag{4}$$

$$|\hat{w}_2| = |\hat{w}_2 - w_2^*| = ||\hat{w}_2 - w_2^*||_2 \leq ||\hat{\theta} - \theta^*||_2 \leq \epsilon \tag{5}$$

$$|\hat{b} - b^*| = ||\hat{b} - b^*||_2 \leq ||\hat{\theta} - \theta^*||_2 \leq \epsilon \tag{6}$$

where the right most inequalities of (4), (5), and (6) are just a trivial application of (3).

We are now ready to state a rigorous version of Theorem 1.

Proposition 2. *Assume we have a logistic regression classifier $\hat{\theta} = [\hat{w}, \hat{b}]^\top$ trained on a dataset $D_n = \{x_i, y_i\}_{i=1}^n$ which is sampled i.i.d. from distribution D. If n is sufficiently large to apply Theorems 1.1 and 1.2 from (Ostrovskii and Bach, 2020) with $\epsilon, \delta > 0$, then after querying n_q additional points within L2 distance β of the decision boundary defined by $\hat{\theta}$, the expected imbalanced ratio of the labeled set, $\mathbb{E}[\eta_s]$, can be bounded from below with probability $1 - \delta$:*

$$\mathbb{E}[\eta_s] \geq \eta \cdot \frac{n + n_q \cdot \frac{1}{\sqrt{2\pi}} \int_v^u \exp(\frac{-1}{2} z^2) dz}{n + n_q \cdot \frac{1}{\sqrt{2\pi}} \int_y^x \exp(\frac{-1}{2} z^2) dz} \tag{7}$$

for

$$u = -\frac{\mu_1(w_1^* + \epsilon) + b^* + \epsilon - \beta \cdot (w_1^* - \epsilon)}{\sigma(w_1^* - \epsilon)}$$

$$v = -\frac{\mu_1(w_1^* - \epsilon) + b^* - \epsilon + \beta \cdot (w_1^* - \epsilon)}{\sqrt{\sigma^2(w_1^* + \epsilon)^2 + \epsilon^2}}$$

$$x = -\frac{\mu_0(w_1^* + \epsilon) + b^* - \epsilon - \beta \cdot (w_1^* + \epsilon)}{\sigma(w_1^* - \epsilon)}$$

$$y = -\frac{\mu_0(w_1^* - \epsilon) + b^* + \epsilon + \beta \cdot (w_1^* + \epsilon)}{\sqrt{\sigma^2(w_1^* + \epsilon)^2 + \epsilon^2}}$$

Proof. Note that the decision boundary of our empirical classifier, $\hat{\theta} = [\hat{w}, \hat{b}]^\top$, is a hyperplane with normal vector $\hat{w}$ and offset $\hat{b}$. Therefore, the set of points at most β away from our decision boundary in terms of Euclidean distance will be the set: $\Omega = \{x : x^\top \hat{w} + \hat{b} + \beta \cdot ||\hat{w}||_2 \geq 0 \text{ and } x^\top \hat{w} + \hat{b} - \beta \cdot ||\hat{w}||_2 \leq 0\}$.

Given a 2-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$, the total probability mass P contained in the set $\Omega^+ = \{x : x^\top w + b \geq 0\}$ is equal to:

$$P = \frac{1}{(2\pi)\sqrt{\det \Sigma}} \int_{\Omega^+} \exp(-\frac{1}{2}(x - \vec{\mu})^\top \Sigma^{-1}(x - \vec{\mu})) dx$$

Through careful re-parameterizations we get that for our setting:

$$P = \frac{1}{\sqrt{2\pi}} \int_\alpha^\infty \exp(-\frac{1}{2} z^2) dz = 1 - \Phi(\alpha)$$

where $\alpha = -\frac{(\mu w_1 + b)}{||\sqrt{\Sigma} w||_2} = -\frac{(\mu w_1 + b)}{\sqrt{\sigma^2 w_1^2 + w_2^2}}$ and Φ is the standard normal CDF. This formula allows us to compute the probability mass of our majority and minority distributions contained in the set Ω and thus the probability of points sampled from these distributions falling in this region.

$$\mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_0, \Sigma)) = (1 - \Phi(-\frac{(\mu_0 \hat{w}_1 + \hat{b} + \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2 \hat{w}_1^2 + \hat{w}_2^2}})) - (1 - \Phi(-\frac{(\mu_0 \hat{w}_1 + \hat{b} - \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2 \hat{w}_1^2 + \hat{w}_2^2}}))$$

$$= \Phi(-\frac{(\mu_0\hat{w}_1 + \hat{b} - \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}}) - \Phi(-\frac{(\mu_0\hat{w}_1 + \hat{b} + \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}})$$

$$\mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_1, \Sigma)) = (1-\Phi(-\frac{(\mu_1\hat{w}_1 + \hat{b} + \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}}))-(1-\Phi(-\frac{(\mu_1\hat{w}_1 + \hat{b} - \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}}))$$

$$= \Phi(-\frac{(\mu_1\hat{w}_1 + \hat{b} - \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}}) - \Phi(-\frac{(\mu_1\hat{w}_1 + \hat{b} + \beta \cdot ||\hat{w}||_2)}{\sqrt{\sigma^2\hat{w}_1^2 + \hat{w}_2^2}})$$

Let us now recall some results from earlier. With probability $1-\delta$, for sufficiently large n, $w_i^* - \epsilon \leq \hat{w}_i \leq w_i^* + \epsilon$ and $b^* - \epsilon \leq \hat{b} \leq b^* + \epsilon$. Additionally, $\Big|\, ||\hat{w}||_2 - ||w^*||_2 \,\Big| \leq ||\hat{w} - w||_2$ by the reverse triangle inequality, so for the same n as before: $\Big|\, ||\hat{w}||_2 - ||w^*||_2 \,\Big| \leq ||\hat{\theta} - \theta^*||_2 \leq \epsilon$ with high probability. Thus, $w_1^* - \epsilon = ||w^*||_2 - \epsilon \leq ||\hat{w}||_2 \leq ||w^*||_2 + \epsilon = w_1^* + \epsilon$. Let's also assume that our parameters are well chosen, so that $|w_1^*|, |b^*| > \epsilon$ and $\mu_0 < 0 < \mu_1$.

Therefore, with probability $1-\delta$,

$$\mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_0, \Sigma)) \leq \Phi\Big(-\frac{(\mu_0(w_1^* + \epsilon) + b^* - \epsilon - \beta \cdot (w_1^* + \epsilon))}{\sigma(w_1^* - \epsilon)}\Big) - \Phi\Big(-\frac{(\mu_0(w_1^* - \epsilon) + b^* + \epsilon + \beta \cdot (w_1^* + \epsilon))}{\sqrt{\sigma^2(w_1^* + \epsilon)^2 + \epsilon^2}}\Big) \quad (8)$$

We can also lower bound the probability for our minority distribution,

$$\mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_1, \Sigma)) \geq \Phi\Big(-\frac{(\mu_1(w_1^* + \epsilon) + b^* + \epsilon - \beta \cdot (||w^*||_2 - \epsilon))}{\sigma(w_1^* - \epsilon)}\Big) - \Phi\Big(-\frac{(\mu_1(w_1^* - \epsilon) + b^* - \epsilon + \beta \cdot (||w^*||_2 - \epsilon))}{\sqrt{\sigma^2(w_1^* + \epsilon)^2 + (d-2)\epsilon^2}}\Big) \quad (9)$$

Finally, we conclude by noting that:

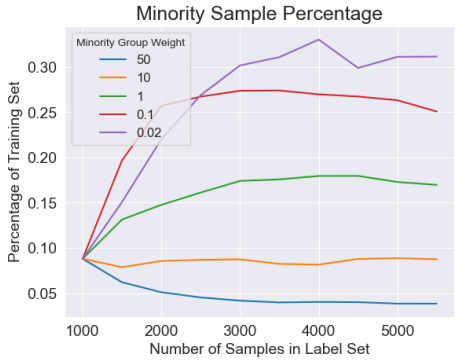
$$\mathbb{E}[\eta_s] = \eta \cdot \frac{n + n_q \cdot \mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_1, \Sigma))}{n + n_q \cdot \mathbb{P}(x \in \Omega \mid x \sim \mathcal{N}(\mu_0, \Sigma))} \quad (10)$$

which we have shown to be bounded from below by the expression given in Theorem 2. □

Appendix B.

We studied the impact of using a less biased classifier when sampling the labeled set for active learning. To do this we varied the weight used for re-weighting the minority class across a range of values, shown in Figure 4.

As explained before, a more bias classifier will be pushed towards underrepresented subpopulations, leading to more minority points being queried at higher frequencies and thus



(a) CIFAR10-LT: Proportion of Minority Data Points in the Labeled Set for Varied Minority Group Weights.

(b) SVHN-LT: Proportion of Minority Data Points in the Labeled Set for Varied Minority Group Weights.

Figure 4: We compare the imbalance ratio achieved by U-AL across a variety of minority group weights. For both datasets, we see that the higher the minority group weight and hence, the less bias the classifier, the worse the imbalance ratio becomes.

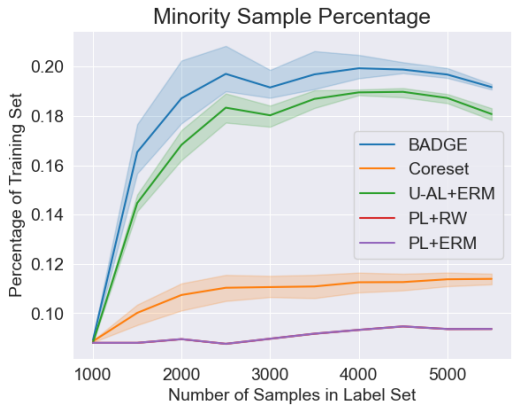
improves the imbalanced ratio of the labeled set. A less biased classifier will not be shifted towards these minority sub-populations as severely, which as we see in Figure 4 leads them to *not* improve the imbalance ratio of the labeled set as much as a more biased classifier.

Appendix C.

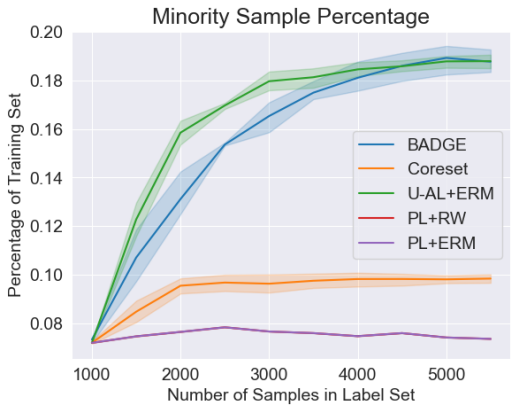
In this section, we include detailed results for our experiments outlined in Section 3.1.2. We present these results in figures: 5, 6, and 7. Alongside, the active learning techniques discussed previously, we include BADGE as a baseline. Note that BADGE incorporates an uncertainty-based objective for sampling and therefore is expected to perform well under imbalance. This follows from the same intuition we presented for U-AL. All experiments for this section were run using a ResNet-50 and any graph with confidence intervals represents the average across 5 independent experimental trials. Any experiment run on CIFAR10-LT, SVHN-LT, or CelebA used an initial labeled set of size 1000 and queried 500 additional labeled points per round. For Waterbird experiments, the initial labeled set contained 500 datapoints with 200 additional labeled points being queried per round.

Appendix D.

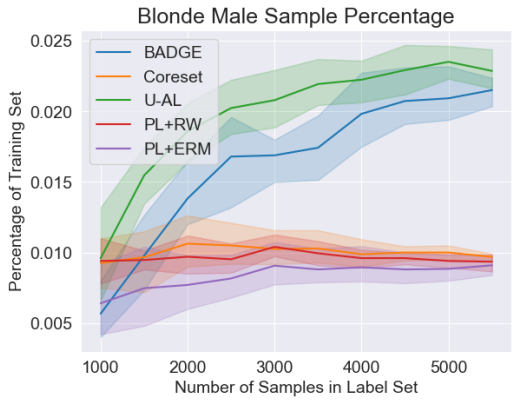
In this section we include the results for our ϵ-greedy experiments introduced in Section 3.4. In figures 8, 9, and 10, we see how group imbalance, average accuracy and worst-group accuracy are affected as ϵ is varied between 0 and 1. Similar to Section 3.4, these experiments were run using a ResNet-50 and any graph with confidence intervals represents the average across 5 independent experimental trials. Again, any experiment run on CIFAR10-LT, SVHN-LT, or CelebA used an initial labeled set of size 1000 and queried 500 additional labeled points per round. For Waterbird experiments, the initial labeled set contained 500 datapoints with 200 additional labeled points being queried per round.



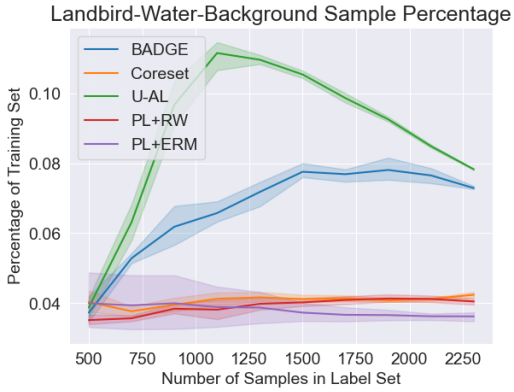
(a) CIFAR10-LT: Proportion of Minority Data Points in the Labeled Set.



(b) SVHN-LT: Proportion of Minority Data Points in the Labeled Set.

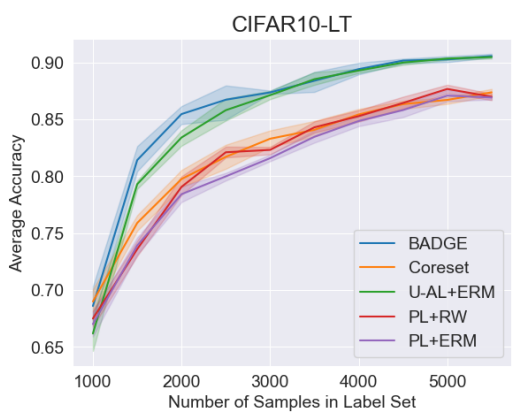


(c) CelebA: Proportion of Blonde Male Data Points in the Labeled Set.

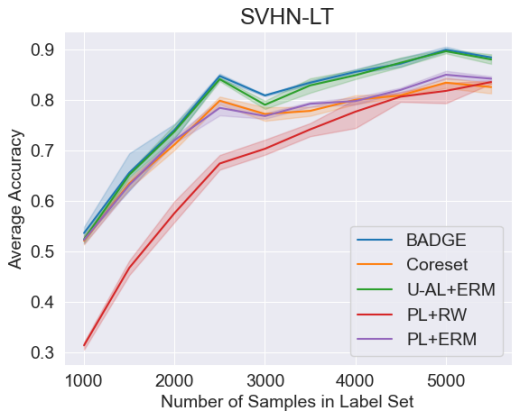


(d) Waterbirds: Proportion of Landbird-Water-Background Data Points in the Labeled Set.

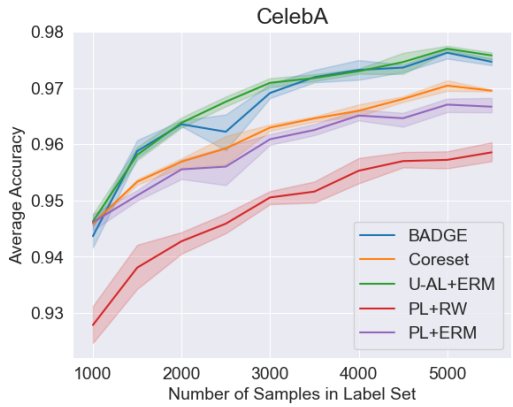
Figure 5: These figures demonstrate that across a wide range of datasets, U-AL increases the proportion of minority points in the labeled set by consistently querying minority points at higher frequencies than their overall prevalence in the dataset.



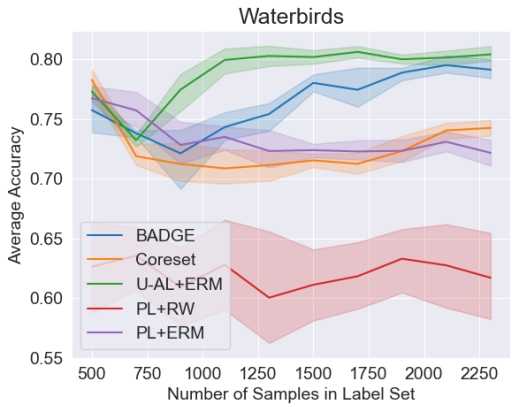
(a) CIFAR10-LT Average Accuracy.



(b) SVHN-LT Average Accuracy.

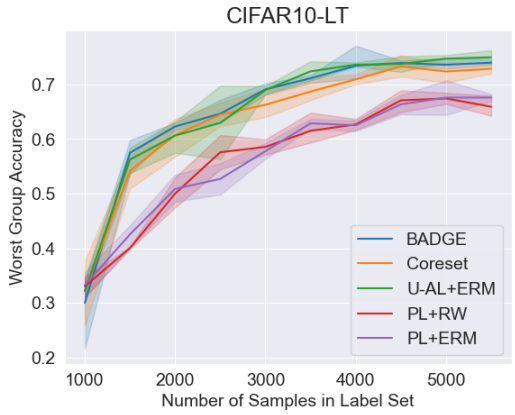


(c) CelebA Average Accuracy.

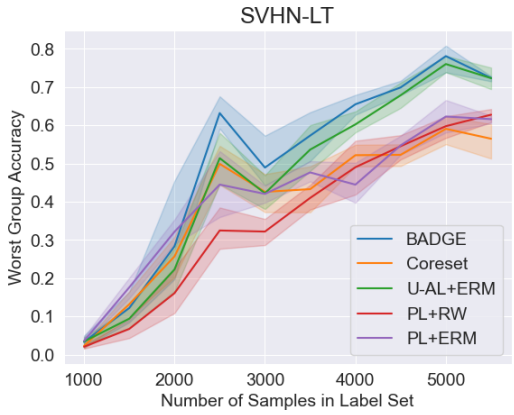


(d) Waterbirds Average Accuracy.

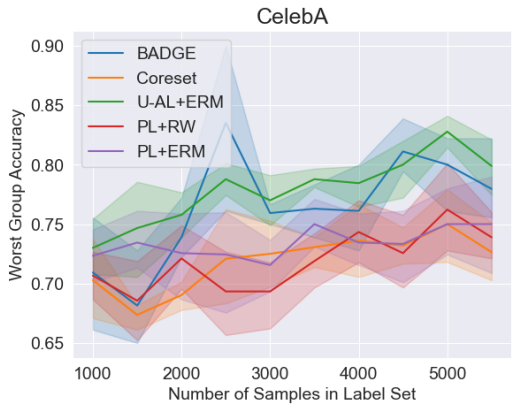
Figure 6: U-AL is still able to perform similarly to more representative sampling methods despite querying points from underrepresented sub-populations disproportionately.



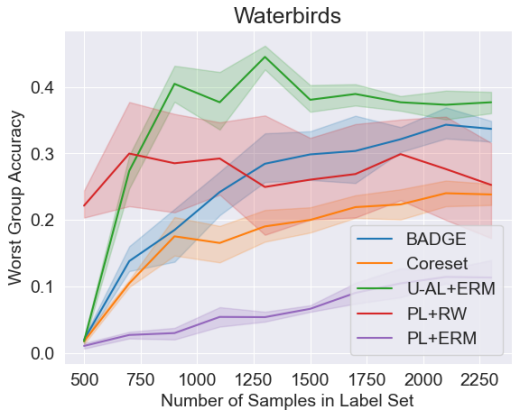
(a) CIFAR10-LT Worst Group Accuracy.



(b) SVHN-LT Worst Group Accuracy.

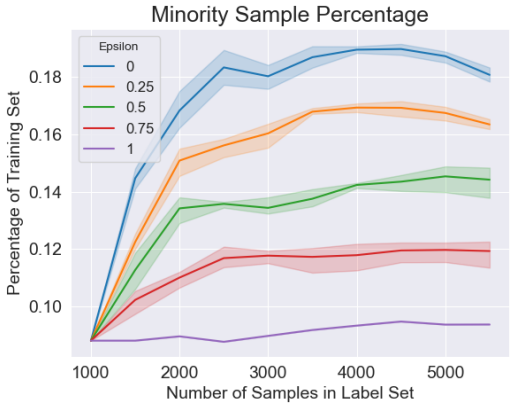


(c) CelebA Worst Group Accuracy.

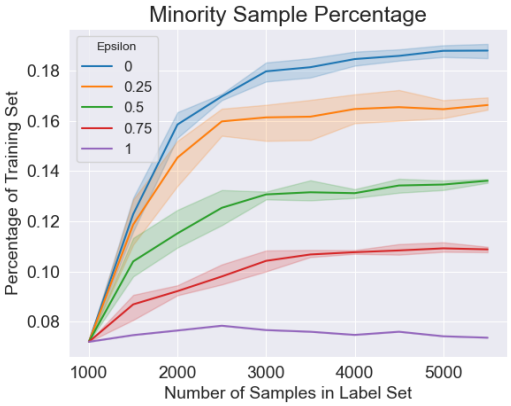


(d) Waterbirds Worst Group Accuracy.

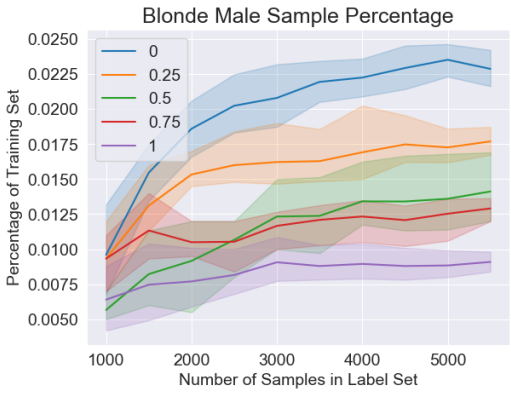
Figure 7: Uncertainty based active learning methods achieve highest worst group accuracy on all datasets after 10 sampling rounds. Uncertainty based method significantly outperform representative methods such as PL+ERM and PL+RW, in some cases by as much as 10%.



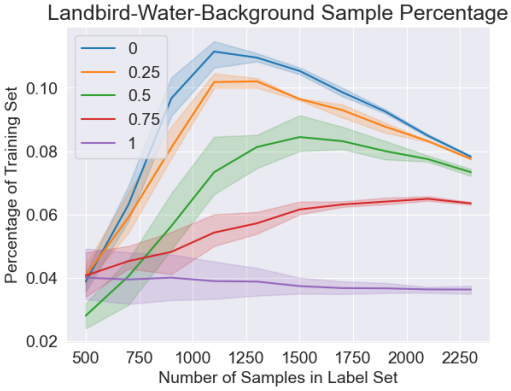
(a) CIFAR10-LT: Proportion of Minority Data Points in the Labeled Set.



(b) SVHN-LT: Proportion of Minority Data Points in the Labeled Set.

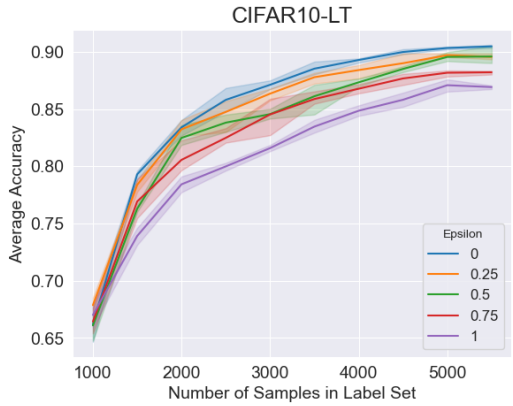


(c) CelebA: Proportion of Blonde Male Data Points in the Labeled Set.

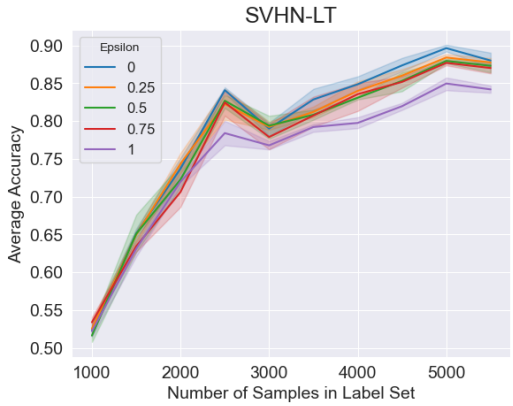


(d) Waterbirds: Proportion of Landbird-Water-Background Data Points in the Labeled Set.

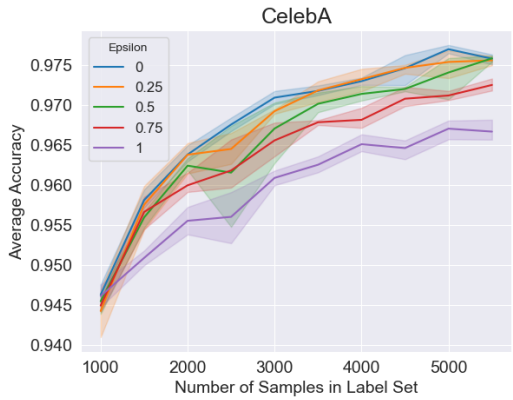
Figure 8: These figures show the minority group sample percentage for each dataset across a wide range of ϵ for ϵ-greedy models. We see that for all datasets, increasing ϵ (and by extension the representativeness of the sampling objective) leads to fewer minority points being sampled. The opposite is true as ϵ decreases.



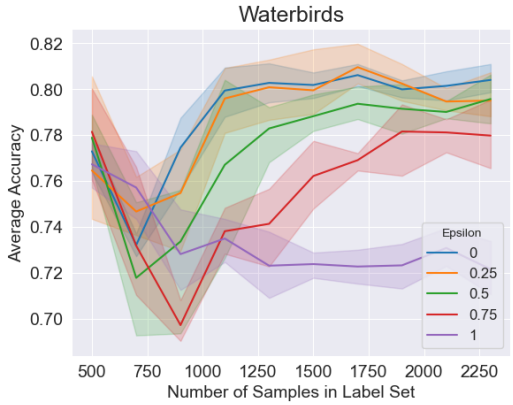
(a) CIFAR10-LT Average Accuracy.



(b) SVHN-LT Average Accuracy.

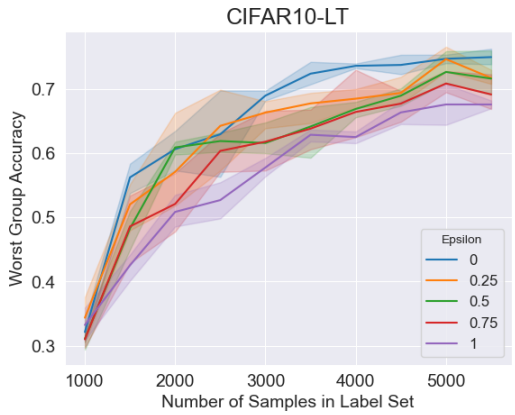


(c) CelebA Average Accuracy.

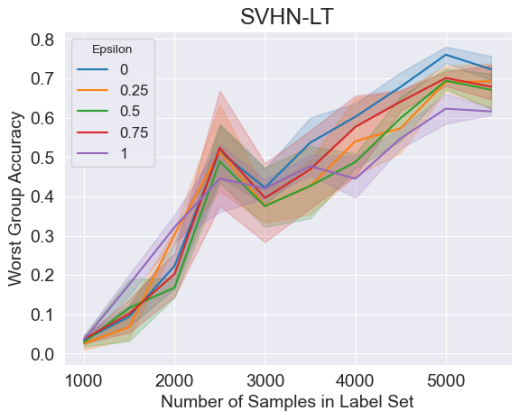


(d) Waterbirds Average Accuracy.

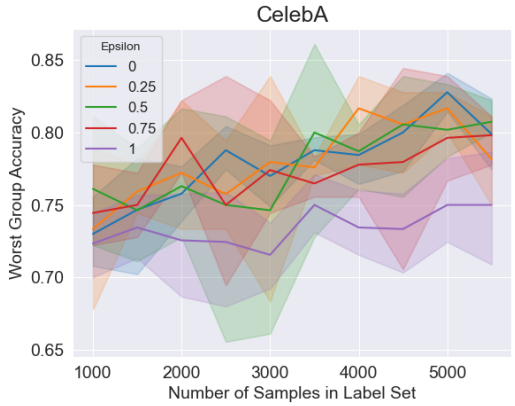
Figure 9: These figures show the effect on average accuracy caused by varying ϵ for ϵ-greedy sampling strategies. We see that interestingly, for all datasets, smaller values of ϵ tend to result in improved average accuracy. This is somewhat surprising, as decreasing ϵ leads to a less representative sampling strategy.



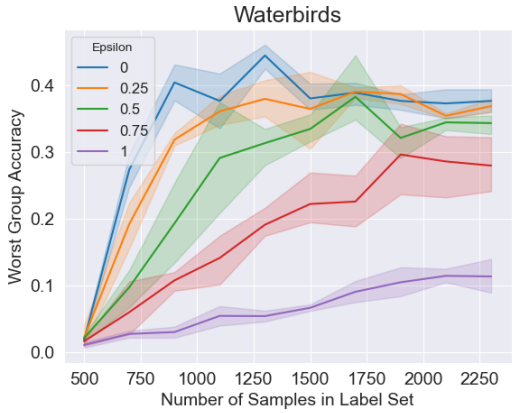
(a) CIFAR10-LT Worst Group Accuracy.



(b) SVHN-LT Worst Group Accuracy.



(c) CelebA Worst Group Accuracy.



(d) Waterbirds Worst Group Accuracy.

Figure 10: These figures show how the worst-group accuracy is impacted by varying ϵ for ϵ-greedy sampling strategies. We see that despite high variance, there is a noticeable trend that decreasing ϵ leads to better performance on underrepresented sub-populations. This is to be expected, since smaller ϵ leads to a more balanced labeled set.