Improved Algorithm for Deep Active Learning under Imbalance via Optimal Separation

Shyam Nuggehalli [*1] **Jifan Zhang** [*1] **Lalit Jain** [2] **Robert Nowak** [1]

Abstract

Class imbalance severely impacts machine learning performance on minority classes in real-world applications. While various solutions exist, active learning offers a fundamental fix by strategically collecting balanced, informative labeled examples from abundant unlabeled data. We introduce DIRECT, an algorithm that identifies class separation boundaries and selects the most uncertain nearby examples for annotation. By reducing the problem to one-dimensional active learning, DIRECT leverages established theory to handle batch labeling and label noise – another common challenge in data annotation that particularly affects active learning methods. Our work presents the first comprehensive study of active learning under both class imbalance and label noise. Extensive experiments on imbalanced datasets show DIRECT reduces annotation costs by over 60% compared to state-of-the-art active learning methods and over 80% versus random sampling, while maintaining robustness to label noise.

1. Introduction

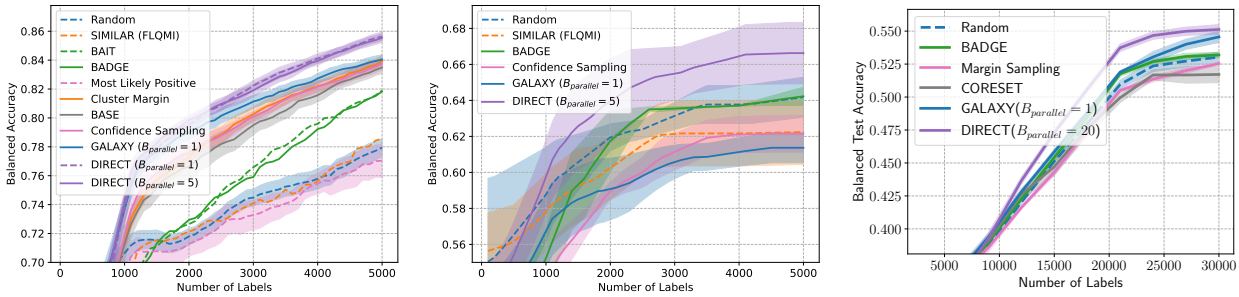
Large-scale deep learning models are playing increasingly important roles across many industries. Human feedback and annotations have played a significant role in developing such systems. Progressively over time, we believe the role of humans in a machine learning pipeline will shift to annotating rare yet important cases. However, under data imbalance, the typical strategy of randomly choosing examples for annotation becomes especially inefficient. This is because the majority of the labeling budget would be spent on common and well-learned classes, resulting in insufficient rare class examples for training an effective model. To mitigate this issue, many recent active learning algorithms have focused on labeling more class-balanced and informative examples (Aggarwal et al., 2020; Kothawade et al., 2021; Zhang et al., 2022; 2024b; Soltani et al., 2024). For many large-scale annotation jobs, this challenge of data imbalance is further compounded by label noise – a critical and common issue that results from annotator decision fatigue and perception differences. A rich body of literature on agnostic active learning (Balcan et al., 2006; Dasgupta et al., 2007; Hanneke et al., 2014; Katz-Samuels et al., 2021) addresses this challenge on low-complexity model classes (e.g. linear models). However, for deep learning models, these algorithms often becomes ineffective due to the large model class complexity. In this paper, we propose a novel active learning strategy for both class imbalance and label noise. Our algorithm DIRECT sequentially and adaptively chooses informative and more class-balanced examples for annotation while being robust to noisy annotations. To the best of our knowledge, this is the first deep active learning study to address the challenging yet prevalent scenario where both imbalance and label noise coexist.

To bridge the gap between the imbalanced deep active learning and the agnostic active learning literature, we propose a novel reduction of the imbalanced classification problem into a set of one-dimensional agnostic active learning problems. For each class, our reduction sorts unlabeled examples into an list ordered by one-vs-rest margin scores. The objective of DIRECT is to find the *optimal separation threshold* which best separates the examples in the given class from the rest. By relating our problem to that of finding the best threshold classifier, we are able to employ ideas from the agnostic active learning literature to learn the separation threshold robustly under label noise. By annotating around the threshold, the annotated examples are more class-balanced and informative.

Comparing to existing active learning algorithms such as BADGE (Ash et al., 2019), Cluster-Margin (Citovsky et al., 2021), SIMILAR (Kothawade et al., 2021), GALAXY (Zhang et al., 2022) and many others, DIRECT improves significantly in label efficiencies – less annotations needed to reach the same accuracy. Notably, most existing methods mentioned above are proposed to handle batch la-

[*]Equal contribution [1]University of Wisconsin-Madison [2]University of Washington, Seattle. Correspondence to: Jifan Zhang <jifan@cs.wisc.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.



(a) Imbalanced CIFAR-10, two classes, no label noise.

(b) Imbalanced CIFAR-100, two classes, 20% label noise.

(c) LabelBench FMoW (62 classes with imbalance), no label noise

Figure 1: Performance of DIRECT over existing baselines for both noiseless and noisy settings. The x-axis represents the total number of labeled examples so far and the y-axis shows the neural network's balanced accuracy. Both (a) and (b) are using supervised training of ResNet-18. In (c), we finetune CLIP ViT-B32 model in combination of semi-supervised training under the LabelBench framework (Zhang et al., 2024a). B_{parallel} is the batch size indicating the number of parallel annotators. $B_{\text{parallel}} = 1$ indicates the synchronous annotation requirement by GALAXY. Our algorithm DIRECT takes pre-specified B_{parallel} as input, which is determined by real world scenarios.

beling, while previous work by Zhang et al. (2022) proposes a superior performance algorithm at the cost of only allowing one annotation at a time. Our algorithm DIRECT is able to obtain the best of both worlds – practical scalability to large annotation jobs by batch labeling while also getting superior performance than all algorithms including GALAXY. On imbalanced datasets, DIRECT achieves state-of-art label efficiency on both supervised fine-tuning of ResNet-18 and semi-supervised fine-tuning of large pretrained model under the LabelBench (Zhang et al., 2024a) framework.

To summarize our main contributions:

- We propose a novel reduction that bridges the advancement in the theoretical agnostic active learning literature to imbalanced active classification for deep neural networks.
- Our algorithm DIRECT addresses the prevalent imbalance and label noise issues and annotates a more class-balanced and informative set of examples.
- Compared to state-of-art algorithm GALAXY (Zhang et al., 2022), DIRECT allows parallel annotation by multiple annotators while still maintaining significant label-efficiency improvement.
- We conduct experiments across 12 dataset settings, four levels of label noise and for both ResNet-18 and large pretrained model (CLIP ViT-B32). DIRECT consistently outperforms existing baseline algorithms by saving more than 60% annotation cost compared to the best existing algorithm, and more than 80% annotation cost compared to random sampling.

2. Related Work

Class-Balanced Deep Active Learning Active learning strategies sequentially and adaptively choose examples for annotation. Many uncertainty-based deep active learning methods extend the traditional active learning literature such as margin, least confidence and entropy sampling (Tong & Koller, 2001; Settles, 2009; Balcan et al., 2006; Kremer et al., 2014). These methods have been shown to perform among the top when fine-tuning large pretrained models and combined with semi-supervised learning algorithms (Zhang et al., 2024a). More sophisticated methods have been proposed to optimize chosen examples' uncertainty (Gal et al., 2017; Ducoffe & Precioso, 2018; Beluch et al., 2018), diversity (Sener & Savarese, 2017; Geifman & El-Yaniv, 2017; Citovsky et al., 2021), or a mix of both (Ash et al., 2019; 2021; Wang et al., 2021; Elenter et al., 2022; Mohamadi et al., 2022). However, these methods often perform poorly under prevalent and realistic scenarios such as label noises (Khosla et al., 2022) or class imbalance (Kothawade et al., 2021; Zhang et al., 2022; 2024a).

Deep Active Learning under Imbalance Data imbalance and rare instances are prevalent in almost all modern machine learning applications. Active learning techniques are effective in addressing the problem in its root by collecting a more class-balanced dataset (Aggarwal et al., 2020; Kothawade et al., 2021; Emam et al., 2021; Zhang et al., 2022; Coleman et al., 2022; Jin et al., 2022; Cai, 2022; Zhang et al., 2024b; Xie et al., 2024). To this end, Kothawade et al. (2021) propose a submodular-based method that actively annotates examples similar to known examples of rare instances. GALAXY(Zhang et al., 2022) constructs one-dimensional linear graphs and applies graph-based active learning techniques in annotating a set of examples that are both class-balanced and uncertain. While GALAXY outperforms existing algorithms, due to a bisection procedure involved, it does not allow parallel annotation. In addition, bisection procedures are generally not robust

against label noises, a prevalent challenge in real world annotation tasks. Our algorithm DIRECT mitigates all of the above shortcomings of GALAXY while outperforming it even with synchronous labeling and no label noise, beating GALAXY in its own game. Lastly, we distinguish our work from Zhang et al. (2024b), where the paper studies the algorithm selection problem. Unlike our goal of proposing a new deep active learning algorithm, the paper proposes meta algorithms to choose the right active learning algorithm among a large number of candidate algorithms.

Agnostic Active Learning for Label Noise Label noise for active learning has been primarily studied under the extensive literature on agnostic learning. We refer the interested reader to the survey (Hanneke et al., 2014) for a thorough discussion. All of these works, beginning with the seminal works by Balcan et al. (2006); Dasgupta et al. (2007), follow a familiar paradigm of disagreement based learning. This involves maintaining a version space of promising hypotheses at each time and constructing a disagreement region of unlabeled examples. For any unlabeled example in the disagreement region, there exists two hypotheses in the version disagreeing on their predictions. An example then chosen for annotation by sampling from a informative sampling distribution computed over the disagreement region. Several approaches have been proposed for computing such sampling distributions, e.g. Jain & Jamieson (2019); Katz-Samuels et al. (2020; 2021); Huang et al. (2015). As described in Section 4.3, our main subroutine VReduce is equivalent to fixed-budget one dimensional threshold disagreement learning based on the ACED algorithm of Katz-Samuels et al. (2021). We remark that these algorithms tend to be overly pessimistic in training deep neural nets, and this paper hopes to close this gap.

Deep Active Learning under Label Noise Label noisy settings has rarely been studied in the deep active learning literature. Related but tangential to our work, several papers have studied to use active learning for cleaning existing noisy labels (Lin et al., 2016; Younesian et al., 2021). In this line of work, they assume access to an oracle annotator that will provide clean labels when queried upon. This is fundamentally different from our work, where our annotator may provide noisy labels. Another line of more theoretical active learning research studies active learning with multiple annotators with different qualities (Zhang & Chaudhuri, 2015; Chen et al., 2022). The primary goal in these work is to identify examples a weak annotator and a strong annotator may disagree, in order to only use the strong annotator on such instances. In our work, we assume access to a single source of annotator that is noisy, which is prevalent in annotation jobs today. Recently, Khosla et al. (2022) proposed a novel deep active learning algorithm specialized for Heteroskedastic noise, where different "regions" of examples are subject to different levels of noise. Unlike their work, our work is agnostic to the noise distributions and conduct experiments on uniformly random corrupted labels. To our knowledge, no deep active learning literature has studied the scenario where both imbalance and label noise present. Yet, this setting is the most prevalent in real-world annotation applications.

3. Preliminary

3.1. Notations

We study the pool-based active learning problem, where an initial unlabeled set of N examples $X = \{x_1, ..., x_N\}$ are available for annotation. Their corresponding labels $Y = \{y_1, ..., y_N\}$ are initially unknown. Furthermore, we study the multi-class classification problem, where the space of labels $\mathcal{Y} := [K]$ is consisted of K classes. Moreover, let $N_1, ..., N_K$ denote the number of examples in X of each class. We define the imbalance ratio as $\gamma = \frac{\min_{k\in[K]} N_k}{\max_{k'\in[K]} N_{k'}}$.

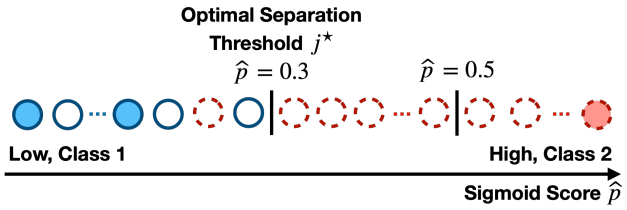
A deep active learning algorithm iteratively chooses batches of examples for annotation. During the t-th iteration, the algorithm is given labeled and unlabeled sets of examples, L_t and U_t respectively, where $L_t \cup U_t = X$ and $L_t \cap U_t = \emptyset$. The algorithm then chooses B examples from the unlabeled set $X^{(t)} \subseteq U_t$ and then obtains their corresponding labels $Y^{(t)}$. The labeled and unlabeled sets are then updated, i.e., $L_{t+1} \leftarrow L_t \cup X^{(t)}$ and $U_{t+1} \leftarrow U_t \backslash X^{(t)}$. Based on new labeled set L_{t+1} and its corresponding labels, a neural network $f_t : X \to [K]$ is trained to inform the choice for the next iteration. The ultimate goal of deep active learning is to obtain high predictive accuracy for the trained neural network while annotating as few examples as possible.

3.2. Limitations of Existing Imbalanced Active Learning Algorithms

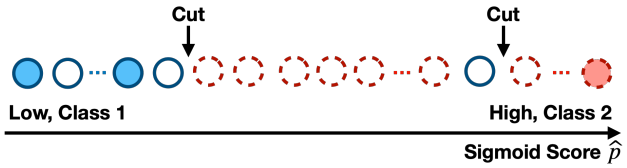
Below we document the several active learning algorithms and how their progressive improvement. At the end, we highlight the shortcomings of the state-of-art algorithm GALAXY (Zhang et al., 2022) and motivate DIRECT's objective of adaptively finding the *optimal separation threshold*. We first consider an imbalanced binary classification case, where $N_1 < N_2$ without loss of generality.

Random Sampling. After annotating a significant number of examples, random sampling would annotate a subset of X with an imbalance ratio close to $\frac{N_1}{N_2}$. This approach suffers from annotating examples that are neither class-balanced nor informative.

Uncertainty Sampling. In the binary classification case, uncertainty sampling methods, such as confidence (Settles, 2009), margin (Tong & Koller, 2001; Balcan et al., 2006) and entropy (Kremer et al., 2014) sampling, simply sort examples based on their predictive sigmoid scores $\widehat{p}$ and



(a) Uncertainty based methods that query around $\widehat{p} = .5$ could annotate examples only in the majority class.



(b) GALAXY spends approximately equal annotation budget around both cuts, while the cut on the right would yield examples mostly in the majority class.

Figure 2: Demonstration of existing imbalance active learning algorithms. Ordered lists of examples are ranked by the predictive sigmoid score $\widehat{p}$. The ground truth label of each example is represented by its border – solid blue for class 1 and dotted red for class 2. Annotated examples are shaded.

annotate examples closest to .5 as demonstrated in Figure 2a. As shown in our results in Figure 1 and Section 5, uncertainty sampling, despite improving over random sampling, significantly underperforms DIRECT and GALAXY and consistently collects less balanced annotations. This shortcoming suggests there are significantly more majority examples than minorities around the decision boundary of $\widehat{p} = .5$.

Objective of DIRECT. To mitigate the above issue with the decision boundary, we propose to identify the *optimal separation threshold*. The threshold best separates the minority and majority classes and approximately equalizes the number of examples from both classes around its vicinity (see Section 4.1 for formal definition). We note the optimal separation threshold could be relatively distant from $\widehat{p} = .5$, as shown in Figure 2a. Our overall objective is to label examples that are *both uncertain and class-balanced*, and can be decomposed into the following two-phased procedure:

1. Identify the *optimal separation threshold* $j^\star$ that best separates the minority class from the majority class, as shown in Figure 2a.
2. Annotate equal number of examples next to $j^\star$ from both sides.

Limitation of GALAXY(Zhang et al., 2022). As discussed above, the neural network decision boundary $\widehat{p} = .5$ does not necessarily best separate minority and majority class examples. GALAXY draws inspiration from graph-based active learning. It relies on the fact that the best separation threshold must be a cut, namely thresholds with a minority class example to the left and a majority class example to the right (see Figure 2b). The algorithm aims to find *all* cuts in the sorted graph as shown in Figure 2b. However, GALAXY suffers from three weaknesses:

1. During active learning, the neural network is still under training and cannot perfectly separate the two classes of examples yet. Therefore, the sorted graph could have a significant number of cuts. As an example in Figure 2b, when annotating around all of such cuts, the algorithm could waste a significant portion of the annotation budget around misclassified outliers, leading to a large number of majority class annotations.
2. Under label noise, the incorrect annotation could lead to more cuts in the sorted graph, further exacerbating the above issue.
3. GALAXY finds all cuts through a modified bisection procedure, which only allows for sequential labeling and prevents multiple annotators labeling in parallel.

In this paper, we take a DIRECT approach by identifying only the optimal separation threshold and address all of the shortcomings above.

4. A Robust Algorithm for Active Learning under Imbalance and Label Noise

In this section, we formally define the optimal separation threshold and pose the problem of identifying it as an 1-dimensional reduction to the agnostic active learning problem. We then propose an algorithm inspired by the agnostic active learning literature (Balcan et al., 2006; Dasgupta et al., 2007; Hanneke et al., 2014; Katz-Samuels et al., 2021).

4.1. An 1-D Reduction to Agnostic Active Learning

We start by considering the imbalanced binary classification setting mentioned in Section 3.2. When given a neural network model, we let $\widehat{p} : X \to [0, 1]$ be the predictive function mapping examples to sigmoid scores. Here, a higher sigmoid score represents a higher confidence of the example being in class 2. We sort examples by their sigmoid predictive score similar to Section 3.2. Formally, we now define the optimal separation threshold as described in Section 3.2.

Definition 4.1. Let $0 = q_{(0)} \leq q_{(1)} \leq \cdots \leq q_{(N)}$, where $\{q_{(i)} \in \mathbb{R}\}_{i=1}^N$ is a sorted permutation of $\{\widehat{p}(x_i)\}_{i=1}^N$. Further we let $\{x_{(i)}\}_{i=1}^N$ and $\{y_{(i)}\}_{i=1}^N$ denote the sorted list's corresponding examples and labels. We define the *optimal separation threshold* as $j^\star \in \{0, 1, ..., N\}$ such that

$$j^\star = \arg\max_j \left(|\{y_{(i)} = 1 : i \leq j\}| - |\{y_{(i)} = 2 : i \leq j\}|\right)$$
$$= \arg\max_j \left(|\{y_{(i)} = 2 : i > j\}| - |\{y_{(i)} = 1 : i > j\}|\right). \quad (1)$$

In other words, on either side of j^*, it has the largest discrep-

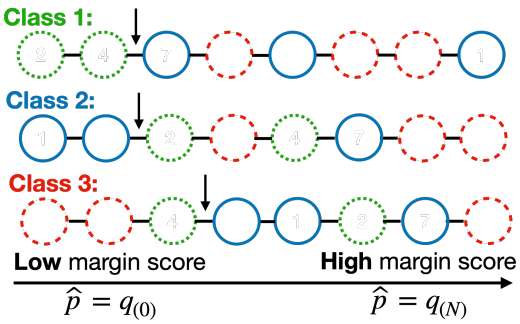


Figure 3: Visualization of multi-class classification. For each class, we formulate the problem as a one-vs-rest binary classification problem by sorting examples based on margin scores. The black arrows indicates the optimal separation thresholds for each class.

ancy in the number of examples between the two classes. This captures the intuition of Figure 2a — our goal is to find a threshold that best separates one class from the other. We quickly remark that ties are broken by choosing the largest $j^\star$ that attains the argmax if class 1 is the minority class and the lowest $j^\star$ otherwise.

1D Reduction. We now provide a reduction of finding $j^\star$ to an 1-dimensional agnostic active learning problem. We define the hypothesis class $\mathcal{H} = \{h_0, h_1, ..., h_N\}$ where each hypothesis h_j is defined as $h_j(q) = \begin{cases} 1 & \text{if } q \le q_{(j)} \\ 2 & \text{if } q > q_{(j)} \end{cases}$. Here, $q_{(0)} = 0$ defines the hypothesis h_0 that predicts class 2 at all times. The empirical zero-one loss for each hypothesis is then defined as $\mathcal{L}(h_j) = \sum_{i=1}^{N} \mathbf{1}\{h_j(q_{(i)}) \neq y_{(i)}\}$. In Appendix A, we show that optimizing for the zero-one loss $\arg\min_{0\le j\le N} \mathcal{L}(h_j)$ is equivalent to equation 1. Namely, with ties broken similar to above, $j^\star = \arg\min_{0\le j\le N} \mathcal{L}(h_j)$.

Multi-Class Classification. To generalize the above problem formulation to multi-class classification, we follow a similar strategy to Zhang et al. (2022). As shown in Figure 3, for each class k, we can view the problem of class-k v.s. others as a binary classification problem. The goal therefore becomes finding all K optimal separation thresholds, which is equivalent with solving K 1-D agnostic active learning problems. Moreover, let $\widetilde{p} : X \to \Delta^{(K-1)}$ denote the neural network prediction function, mapping examples to softmax scores. For each class k, we use the margin scores $\widehat{p}_i^k := [\widetilde{p}(x_i)]_k - \max_{k'} [\widetilde{p}(x_i)]_{k'}$ to sort the examples and break ties by their corresponding confidence scores $[\widetilde{p}(x_i)]_k$. Formally,

$$\left(q^k_{(1)} \le \cdots \le q^k_{(N)} \text{: sorted permutation of } \{\widehat{p}_i^k\}_{i=1}^N\right) \wedge \left(q^k_{(i)} = q^k_{(i+1)} \Rightarrow [\widetilde{p}(x_i)]_k \ge [\widetilde{p}(x_{i+1})]_k\right). \quad (2)$$

Note that sorting by margin scores is equivalent to sorting by sigmoid scores for binary classification.

4.2. One-Dimensional Agnostic Active Learning

The key insight of our approach is to leverage the well-established theory of agnostic active learning for threshold classifiers, which provides robust guarantees even under label noise. This allows us to robustly identify the optimal separation threshold $j^\star$ from the 1-D reduction above.

Problem Formulation and Noise Handling. In the agnostic setting, we have a sorted sequence $q_{(1)} \le q_{(2)} \le \cdots \le q_{(N)}$ with corresponding (possibly noisy) binary labels $y_{(1)}, y_{(2)}, \ldots, y_{(N)}$. The fundamental challenge is that no hypothesis $h_j \in \mathcal{H}$ may achieve zero empirical loss $\mathcal{L}(h_j)$ due to label noise or model misspecification.

To formalize how our algorithm handles label noise, we consider the underlying conditional probability function $P(y_i|x_i)$ for each data example x_i. Through the dimensionality reduction $x_i \to q_i$ (where q_i is the real-valued sigmoid/margin score), we obtain an ordered set of 1-dimensional features $\{q_i\}_{i=1}^N$. This mapping induces a distribution $P(y_i|q_i)$ over the reduced space, which naturally encodes any label noise present in the annotations.

Our objective is to find the threshold classifier $h_{j^\star}$ from the set of 1-dimensional threshold classifiers $\{h_j\}$ that minimizes the probability of error with respect to $P(y_i|q_i)$. Crucially, our agnostic approach makes no assumptions about the form of $P(y_i|x_i)$ or the induced $P(y_i|q_i)$, allowing the algorithm to handle arbitrary noise models without requiring prior knowledge of the noise distribution.

Version Space Reduction. The VReduce algorithm (Algorithm 1) implements a version space approach that maintains an interval $[I, J]$ representing plausible optimal thresholds. The key principle is that if we observe labeled examples $x_{(i)}$ with $i \le I$ mostly belong to class 1, and examples $x_{(j)}$ with $j \ge J$ mostly belong to class 2, then the optimal threshold $j^\star$ with high likelihood will lie within $[I, J]$.

The algorithm proceeds iteratively by: (1) maintaining and updating the version space interval $[I, J]$ based on observed labels, (2) sampling unlabeled examples uniformly within this interval to maximize information gain, (3) shrinking the version space based on empirical loss estimates that account for the noisy observations, and (4) repeating until the labeling budget is exhausted.

Theoretical Guarantees. The VReduce algorithm inherits robust theoretical properties from the agnostic active learning literature, achieving near-minimax and instance-optimal sample complexity bounds (Dasgupta et al., 2007; Hanneke et al., 2014). Unlike bisection-based methods that fail under label corruption, the disagreement-based framework (Balcan et al., 2006) provides natural robustness with high

Algorithm 1 VReduce: Version Space Reduction for Threshold Learning

Input: Labeled set L, budget b, class of interest k, parallel batch size B_{parallel}, sorted examples $\{x^k_{(i)}, y^k_{(i)}, q^k_{(i)}\}_{i=1}^N$ (note $y^k_{(i)}$ of unlabeled examples are hidden to the learner).
Initialize: Version space $[I, J]$ as the shortest interval such that: $\forall i \le I$ with $x_{(i)} \in L$: $y_{(i)} = k$, and $\forall j \ge J$ with $x_{(j)} \in L$: $y_{(j)} \neq k$.
Number of iterations $m \leftarrow b/B_{\text{parallel}}$. Shrinking factor $c \leftarrow (J-I)^{1/m}$.
for $t = 1, ..., m$ **do**
  Sample B_{parallel} unlabeled examples uniformly from $\{x^k_{(I)}, \dots, x^k_{(J)}\}$ and add to L.
  Compute empirical loss: $\widehat{\mathcal{L}}^k(s) = \sum_{r \le s : x_{(r)} \in L} \mathbf{1}\{y_{(r)} \neq k\} + \sum_{r > s : x_{(r)} \in L} \mathbf{1}\{y_{(r)} = k\}$.
  Update version space: $[I, J] \leftarrow \arg\min_{[i,j]: j-i=(J-I)/c} \max\{\widehat{\mathcal{L}}^k(i), \widehat{\mathcal{L}}^k(j)\}$.
end for
Return: Updated labeled set L.

probability guarantees regardless of the underlying data distribution or noise model. Building on the ACED framework (Katz-Samuels et al., 2021), our algorithm extends these classical guarantees to practical batch settings through the B_{parallel} parameter while preserving all theoretical properties, making it suitable for real-world annotation scenarios with multiple parallel annotators.

4.3. Algorithm

We are now ready to state our algorithm DIRECT as shown in Algorithm 2. Each round of DIRECT follows a two-phased procedure, where the first phase aims to identify the optimal separation threshold for each class using the agnostic active learning approach described above. The second phase then annotates examples closest to the estimated optimal separation thresholds for each class. We spend half each round's budget for both phases.

During the first phase, to identify the optimal separation threshold for all classes, we loop over each class k and run the agnostic active learning procedure VReduce for the corresponding 1-D class-k v.s. rest reduction. The second phase of DIRECT simply annotates examples closest to each optimal separation threshold, aiming to annotate a class-balanced and uncertain examples.

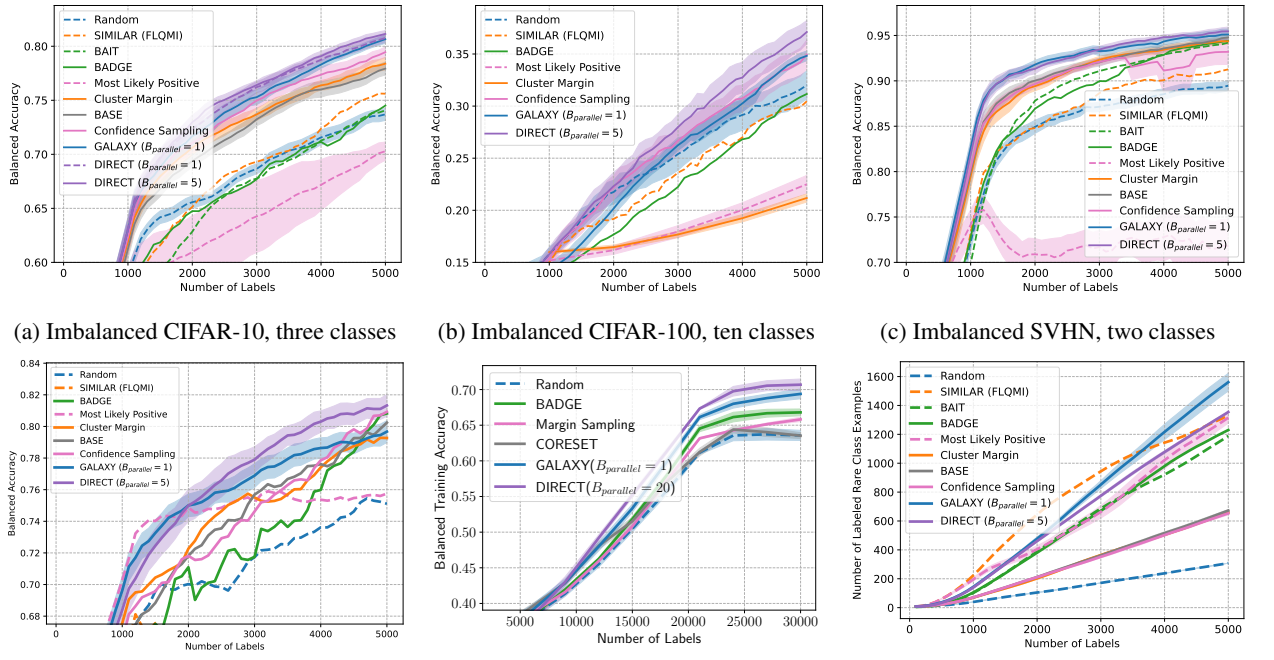
To address batch labeling, we let B_{train} denote the number of examples the algorithm collects before the neural network is retrained. In practice, this number is usually determined by the constraints of computational training cost. On the other hand, we let B_{parallel} denote the number of examples

Algorithm 2 DIRECT: DImension REduction for aCTive Learning under Imbalance and Label Noise

Input: Pool X, #Rounds T, retraining batch size B_{train}, number of parallel annotations B_{parallel}.
Initialize: Uniformly sample B elements from X to form L_0. Let $U_0 \leftarrow X \backslash L_0$.
for $t = 1, ..., T-1$ **do**
  Train neural network on L_{t-1} and obtain f_{t-1}.
  Find optimal separation thresholds
  Initialize labeled set $L_t \leftarrow L_{t-1}$ and budget per class $b \leftarrow B_{\text{train}}/2K$.
  for k in RandPerm($\{1, ..., K\}$) **do**
    Sort margin scores $0 = q^k_{(0)} \le q^k_{(1)} \le \cdots \le q^k_{(N)}$ based on equation 2.
    Let $x^k_{(i)}, y^k_{(i)}$ denote the example and label corresponding to $q^k_{(i)}$.
    Identify threshold for class k: $L_t \leftarrow \text{VReduce}(L_t, b, k, B_{\text{parallel}}, \{x^k_{(i)}, y^k_{(i)}, q^k_{(i)}\}_{i=1}^N)$.
  end for
  Annotate examples around the identified threshold
  Compute budget per class $b \leftarrow (B_{\text{train}} - |L_t|)/K$.
  for k in RandPerm($\{1, ..., K\}$) **do**
    Estimate separation threshold (break ties by choosing the index closest to $\frac{N}{2}$): $\widehat{j}^k \leftarrow \arg\max_j (|\{y_{(i)} = k : x_{(i)} \in L_t \text{ and } i \le j\}| - |\{y_{(i)} \neq k : x_{(i)} \in L_t \text{ and } i \le j\}|)$.
    Annotate b unlabeled examples with sorted indices closest to $\widehat{j}^k$ and insert to L_t.
  end for
end for
Return: Train final classifier f_T based on L_T.

annotated in parallel. We note that, in practice, the number of examples collected before retraining is usually far greater than the number of annotators annotating in parallel, i.e., $B_{\text{parallel}} \ll B_{\text{train}}$. Lastly, as will be discussed in Section 6, our algorithm can also be modified for asynchronous labeling.

Theoretical Comparison with GALAXY. As mentioned in Section 3.2, GALAXY's graph-based approach aims to identify all *cuts* and sample examples around all of them equally. On the other hand, DIRECT aims to identify only the separation threshold and sample around it, which is superior as we have argued before and shown in our results. We now present a more theoretical comparison. As we show in Appendix A, the graph-based approach in GALAXY will identify and annotate around at least one more cut in addition to the optimal separation threshold, with probability at least $1 - \exp(-b \log(\frac{1}{1-\eta})/2)$. Here, b is the budget of a single round of annotation and η is the label noise ratio (see Appendix A for more details). This implies, when the budget b is large, GALAXY will likely annotate around unnecessary cuts. This is in contrast with the agnostic active



(a) Imbalanced CIFAR-10, three classes

(b) Imbalanced CIFAR-100, ten classes

(c) Imbalanced SVHN, two classes

(d) PathMNIST, two classes

(e) FMoW, Balanced Pool Accuracy

(f) Imbalanced SVHN, two classes, # of minority labeled examples

Figure 4: Performance comparison of DIRECT against other baselines algorithms in the noiseless but imbalanced setting. (a)-(d) are balanced accuracy on ResNet-18 experiments while (e) shows experiment under the LabelBench framework. B_{parallel} indicates the number of parallel annotations as mentioned in Section 4.3. $B_{\text{parallel}} = 1$ is equivalent with sychornous labeling. Results are averaged over four trials and the shaded areas represent standard errors around the mean.

learning approach we take in DIRECT, where as shown by Katz-Samuels et al. (2021), the probability of misidentifying the optimal separation threshold decays exponentially w.r.t. budget b. In other words, with a large budget b, with high likelihood, DIRECT will focus its annotation around the optimal separation threshold. Lastly, time complexity analysis in Appendix C shows DIRECT's superior speed compared to BADGE and GALAXY.

5. Experiments

We conduct experiments under two primary setups:

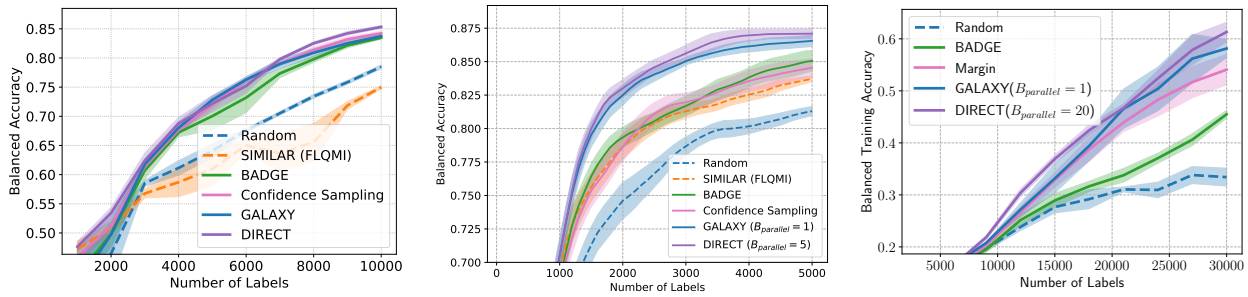
1. Supervised fine-tuning of ResNet-18 on imbalanced datasets similar to Zhang et al. (2022).
2. Fine-tuning large pretrained model (CLIP ViT-B32) with semi-supervised training strategies under the LabelBench framework (Zhang et al., 2024a).

For both evaluation setups, we first evaluate the performance of DIRECT under the noiseless setting in Section 5.1, showing its superior label-efficiency and ability to accommodate batch labeling. In Section 5.2, we evaluate deep active learning algorithms under a novel setting with both class imbalance and noisy labels. Under this setting, we also include an ablation study of the performance of DIRECT on various levels of label noises. While we highlight many results in this section, see Appendix E for complete results. Our implementation is publicly available at: `https://github.com/EfficientTraining/LabelBench/blob/main/LabelBench/strategy/strategy_impl/direct.py`

Experiment Setups. Our experiments utilize 14 imbalanced datasets derived from popular computer vision datasets. For the ResNet experiments, we utilize imbalanced and/or long-tail versions of CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011) and PathMNIST (Yang et al., 2021) datasets. For the LabelBench experiments, we utilize the FMoW (Christie et al., 2018), iWildcam (Beery et al., 2021), iNaturalist (Van Horn et al., 2018) and ImageNet-LT (Deng et al., 2009) datasets. We refer the readers to Appendix D for more details on our experiment setups.

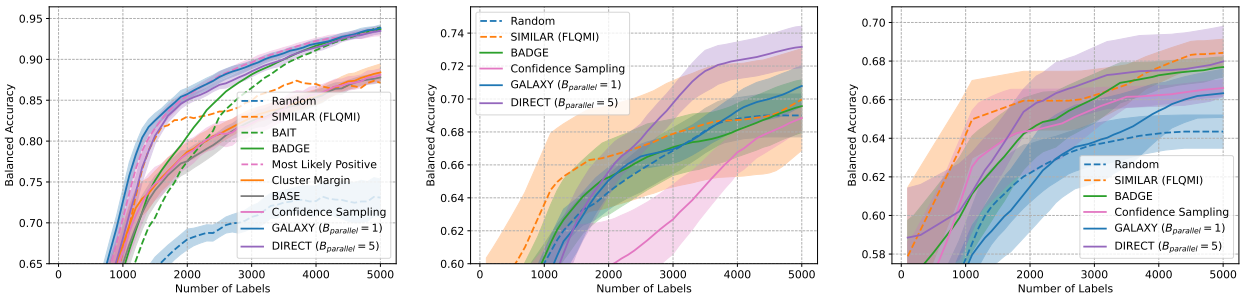
5.1. Experiments under Imbalance, without Label Noise

For the noiseless experiment on ResNet-18, we compare against nine baselines: GALAXY (Zhang et al., 2022), SIMILAR (Kothawade et al., 2021), BADGE (Ash et al., 2019), BASE (Emam et al., 2021), BAIT (Ash et al., 2021), Cluster Margin (Citovsky et al., 2021), Confidence Sampling (Set-



(a) CIFAR-10LT, 10 classes, 15% label noise (b) Imbalanced SVHN, two classes, 10% label noise (c) iWildcam Balanced Pool Accuracy, 10% label noise

Figure 5: Performance of DIRECT against baseline algorithms under label noise. (a)-(b) are balanced accuracy on ResNet-18 experiments while (c) shows results under the LabelBench framework. Results are averaged over four trials and the shaded areas represent standard errors around the mean.



(a) Imbalanced CIFAR-100, two classes, no label noise (b) Imbalanced CIFAR-100, two classes, 10% label noise (c) Imbalanced CIFAR-100, two classes, 15% label noise

Figure 6: Performance of DIRECT against baseline algorithms under different levels of label noise. Results are averaged over four trials and the shaded areas represent standard errors around the mean.

tles, 2009), Most Likely Positive (Jiang et al., 2018; Warmuth et al., 2001; 2003) and Random Sampling. We briefly distinguish the algorithms into two categories. In particular, both SIMILAR and Most Likely Positive annotate examples that are similar to existing labeled minority examples, thus can significantly annotate a large quantity of minority examples. The rest of the algorithms primarily optimizes for different notions of informativeness such as diversity and uncertainty.

For the LabelBench experiments, due to the large dataset and model embedding sizes, we choose a subset of the algorithms that are computationally efficient and among top performers in the ResNet-18 results, including BADGE, Margin Sampling, CORESET and GALAXY. As highlighted in Figures 1(a) and 4(a)-(d), DIRECT consistently and significantly outperforms existing algorithms on the ResNet-18 experiments. In Figures 1(c) and 4(e), we demonstrate the increased label-efficiency is also consistently shown in the LabelBench experiments. Compared to random sampling, DIRECT can save more than 80% of the annotation cost on imbalanced SVHN experiment of Figure 4(c). In terms of class-balancedness, we consistently observe that both Most Likely Positive and SIMILAR annotating greater number of minority class examples, but significantly underperforms in terms of balanced accuracy (an example showin in Figure 4(f)). While Zhang et al. (2022) has already observed this phenomenon, we can further see that DIRECT collects slightly less minority class examples than GALAXY, but outperforms in terms of balanced accuracy. While it is crucial to optimize class-balancedness for better model performance, we see that both extremes of annotating too few and too many minority examples could lead to worse generalization performances. When too few examples are from minority class, the performance of the minority classes could be significantly hindered. When optimized to annotate as many examples from minority class as possible, the algorithm has to tradeoff annotating informative examples to examples it is more certain to be in the minority class. Together, this suggests an intricate balance between the two objectives, generalization performance and class-balancedness.

We would also like to highlight the ability to handle batch labeling. Across our experiments, we see DIRECT out-

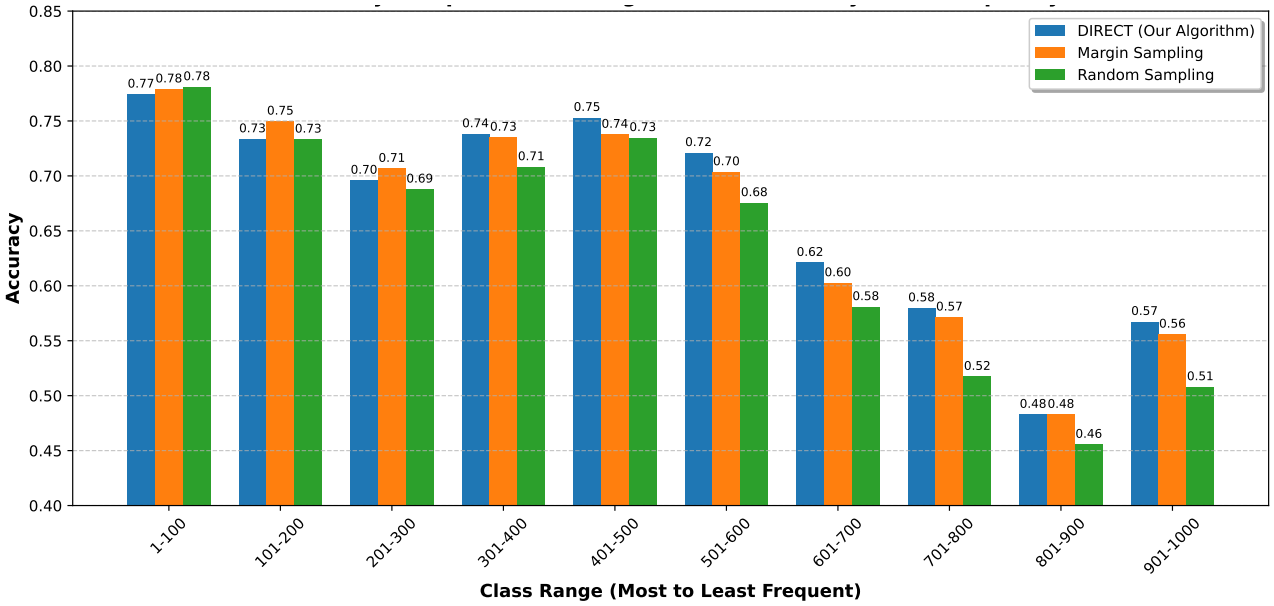


Figure 7: Performance of different active learning algorithm across different classes for the ImageNet-LT dataset.

performs with different amounts of parallel annotation ($B_{\text{parallel}} = 1, 5$ and 20), indicating its general effectiveness. This is in comparison to the synchoronous nature of GALAXY, where it is always using $B_{\text{parallel}} = 1$. On Figures 1(a) and 5(a), we see that DIRECT outperforms GALAXY with synchronous labeling. Furthermore, in these experiments we also see using $B_{\text{parallel}} = 5$ only affects algorithm performances minimally for DIRECT.

5.2. Experiments under Imbalance and Label Noise

We conduct novel sets of experiments under both class imbalance and label noise. Here, for both ResNet-18 and LabelBench experiments, we evaluate against all of the algorithms that performed well under the imbalance but noiseless setting above. For all of our experiments, we introduce a fixed percentage of label noise, where the given fraction of the examples' labels are corrupted to a different class uniformly at random. For most of our experiments with 10% label noise shown in Figure 5, we observe again that DIRECT consistently improves over all baselines including GALAXY. The results are consistent on ResNet-18 and LabelBench setups, and with different B_{parallel} values, showing DIRECT's robustness under label noise.

Different Levels of Label Noise As shown in Figures 6(a)-(c) and 1(b), we observe the results on imbalanced CIFAR-100 with two classes across numerous levels of label noise, with 0%, 10%, 15% and 20% respectively. In fact, the noiseless experiment in Figure 6(a) is the only setting DIRECT slightly underperforms GALAXY in terms of generalization accuracy. However, we see DIRECT becomes more label-efficient under label noise. It is also worth noting that with high label noise of 20%, we observe in Figure 1(b) that existing algorithms underperform random sampling. In contrast, DIRECT significantly outperforms random sampling, saving more than 60% of the annotation cost.

5.3. Qualitative Analysis

The ImageNet-LT dataset is constructed with class frequencies that gradually decrease according to class index. In Figure 7, we organize classes into bins ordered from most frequent to least frequent. The results show that DIRECT significantly outperforms baseline algorithms on less frequent classes (indices 301-1000), which accounts for DIRECT's superior overall performance despite modest accuracy reductions on more frequent classes. This finding aligns with our balancedness analysis, where DIRECT demonstrates improved labeling of samples from rare classes.

6. Conclusion and Future Work

In this paper, we conducted the first study of deep active learning under both class imbalance and label noise. We proposed an algorithm DIRECT that significantly and consistently outperforms existing literature. In this work, we also addressed the batch sampling problem of Zhang et al. (2022), by annotating multiple examples in parallel. Studying asynchronous labeling could be a natural extension of our work. A potential solution is to utilize an asynchronous variant of one-dimensional active learning algorithm. In addition, one can further batch the labeling process across different classes to further accommodate an even larger number of parallel annotators.

Impact Statement

In the rapidly evolving landscape of machine learning, the efficacy of active learning in addressing data imbalance and label noise is a significant stride towards more robust and equitable AI systems. This research explores how active learning can effectively mitigate the challenges posed by imbalanced datasets and erroneous labels, prevalent in real-world scenarios.

The positive impacts of this research are multifaceted. It enhances the accessibility and utility of machine learning in domains where data imbalance is a common challenge, such as healthcare, finance, and social media analytics. By improving class-balancedness in annotated sets, models trained on these datasets are less biased and more representative of real-world distributions, leading to fairer and more accurate outcomes. Additionally, this research contributes to reducing the time and cost associated with data annotation, which is particularly beneficial in fields where expert annotation is expensive or scarce.

However, if not carefully implemented, active learning strategies could inadvertently introduce new biases or amplify existing ones, particularly in scenarios where the initial data is severely imbalanced or contains deeply ingrained biases. Furthermore, the advanced nature of these techniques may widen the gap between organizations with access to state-of-the-art technology and those without, potentially exacerbating existing inequalities in technology deployment.

Acknowledgements

This work has been supported in part by NSF Award 2112471.

References

Aggarwal, U., Popescu, A., and Hudelot, C. Active learning for imbalanced datasets. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1428–1437, 2020.

Ash, J., Goel, S., Krishnamurthy, A., and Kakade, S. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34: 8927–8939, 2021.

Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J., and Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 65–72, 2006.

Beery, S., Agarwal, A., Cole, E., and Birodkar, V. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9368–9377, 2018.

Cai, X. Active learning for imbalanced data: The difficulty and proportions of class matter. *Wireless Communications and Mobile Computing*, 2022, 2022.

Chen, Y., Sankararaman, K., Lazaric, A., Pirotta, M., Karamshuk, D., Wang, Q., Mandyam, K., Wang, S., and Fang, H. Improved adaptive algorithm for scalable active learning with weak labeler. *arXiv preprint arXiv:2211.02233*, 2022.

Christie, G., Fendley, N., Wilson, J., and Mukherjee, R. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6172–6180, 2018.

Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., and Kumar, S. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Coleman, C., Chou, E., Katz-Samuels, J., Culatana, S., Bailis, P., Berg, A. C., Nowak, R., Sumbaly, R., Zaharia, M., and Yalniz, I. Z. Similarity search for efficient active learning and search of rare concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6402–6410, 2022.

Dasgupta, S., Hsu, D. J., and Monteleoni, C. A general agnostic active learning algorithm. *Advances in neural information processing systems*, 20, 2007.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Elenter, J., NaderiAlizadeh, N., and Ribeiro, A. A lagrangian duality approach to active learning. *arXiv preprint arXiv:2202.04108*, 2022.

Emam, Z. A. S., Chu, H.-M., Chiang, P.-Y., Czaja, W., Leapman, R., Goldblum, M., and Goldstein, T. Active learning at the imagenet scale. *arXiv preprint arXiv:2111.12880*, 2021.

Gal, Y., Islam, R., and Ghahramani, Z. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pp. 1183–1192. PMLR, 2017.

Geifman, Y. and El-Yaniv, R. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Hanneke, S. et al. Theory of disagreement-based active learning. *Foundations and Trends® in Machine Learning*, 7(2-3):131–309, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, T.-K., Agarwal, A., Hsu, D. J., Langford, J., and Schapire, R. E. Efficient and parsimonious agnostic active learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Jain, L. and Jamieson, K. G. A new perspective on pool-based active classification and false-discovery control. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiang, S., Malkomes, G., Abbott, M., Moseley, B., and Garnett, R. Efficient nonmyopic batch active search. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, 2018.

Jin, Q., Yuan, M., Wang, H., Wang, M., and Song, Z. Deep active learning models for imbalanced image classification. *Knowledge-Based Systems*, 257:109817, 2022.

Katz-Samuels, J., Jain, L., Jamieson, K. G., et al. An empirical process approach to the union bound: Practical algorithms for combinatorial and linear bandits. *Advances in Neural Information Processing Systems*, 33:10371–10382, 2020.

Katz-Samuels, J., Zhang, J., Jain, L., and Jamieson, K. Improved algorithms for agnostic pool-based active classification. In *International Conference on Machine Learning*, pp. 5334–5344. PMLR, 2021.

Khosla, S., Whye, C. K., Ash, J. T., Zhang, C., Kawaguchi, K., and Lamb, A. Neural active learning on heteroskedastic distributions. *arXiv preprint arXiv:2211.00928*, 2022.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kothawade, S., Beck, N., Killamsetty, K., and Iyer, R. Similar: Submodular information measures based active learning in realistic scenarios. *Advances in Neural Information Processing Systems*, 34:18685–18697, 2021.

Kremer, J., Steenstrup Pedersen, K., and Igel, C. Active learning with support vector machines. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(4):313–326, 2014.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lin, C., Mausam, M., and Weld, D. Re-active learning: Active learning with relabeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Mohamadi, M. A., Bae, W., and Sutherland, D. J. Making look-ahead active learning strategies feasible with neural tangent kernels. *arXiv preprint arXiv:2206.12569*, 2022.

Müller, R., Kornblith, S., and Hinton, G. E. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Settles, B. Active learning literature survey. 2009.

Soltani, N., Zhang, J., Salehi, B., Roy, D., Nowak, R., and Chowdhury, K. Learning from the best: Active learning for wireless communications. *arXiv preprint arXiv:2402.04896*, 2024.

Tong, S. and Koller, D. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2(Nov):45–66, 2001.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Wang, H., Huang, W., Margenot, A., Tong, H., and He, J. Deep active learning by leveraging training dynamics. *arXiv preprint arXiv:2110.08611*, 2021.

Warmuth, M. K., Rätsch, G., Mathieson, M., Liao, J., and Lemmen, C. Active learning in the drug discovery process. In *NIPS*, pp. 1449–1456, 2001.

Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., and Lemmen, C. Active learning with support vector machines in the drug discovery process. *Journal of chemical information and computer sciences*, 43(2):667–673, 2003.

Xie, T., Zhang, J., Bai, H., and Nowak, R. Deep active learning in the open world. *arXiv preprint arXiv:2411.06353*, 2024.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021.

Younesian, T., Zhao, Z., Ghiassi, A., Birke, R., and Chen, L. Y. Qactor: Active learning on noisy labels. In *Asian Conference on Machine Learning*, pp. 548–563. PMLR, 2021.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., and Shinozaki, T. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34: 18408–18419, 2021.

Zhang, C. and Chaudhuri, K. Active learning from weak and strong labelers. *Advances in Neural Information Processing Systems*, 28, 2015.

Zhang, J., Katz-Samuels, J., and Nowak, R. Galaxy: Graph-based active learning at the extreme. *arXiv preprint arXiv:2202.01402*, 2022.

Zhang, J., Chen, Y., Canal, G., Das, A. M., Bhatt, G., Mussmann, S., Zhu, Y., Bilmes, J., Du, S. S., Jamieson, K., et al. Labelbench: A comprehensive framework for benchmarking adaptive label-efficient learning. *Journal of Data-centric Machine Learning Research*, 2024a.

Zhang, J., Shao, S., Verma, S., and Nowak, R. Algorithm selection for deep active learning with imbalanced datasets. *Advances in Neural Information Processing Systems*, 36, 2024b.

A. Equivalent Objective

Lemma A.1. *The agnostic active learning reduction is equivalently finding the optimal separation threshold. Namely,*

$$\arg\min_j \mathcal{L}(h_j) = \arg\max_j \left(|\{y_{(i)} = 1 : 1 \le i \le j\}| - |\{y_{(i)} = 2 : 1 \le i \le j\}|\right)$$

Proof. Recall the definitions: $h_j(q) = \begin{cases} 1 & \text{if } q \le q_{(j)} \\ 2 & \text{if } q > q_{(j)} \end{cases}$ and $\mathcal{L}(h_j) = \sum_{i=1}^N \mathbf{1}\{h_j(q_{(i)}) \neq y_{(i)}\}$, we can expand the loss as follows

$$\begin{aligned}
\arg\min_j \mathcal{L}(h_j) &= \arg\min_j \sum_{i=1}^N \mathbf{1}\{h_j(q_{(i)}) \neq y_{(i)}\} \\
&= \arg\min_j N - \sum_{i=1}^N \mathbf{1}\{h_j(q_{(i)}) = y_{(i)}\} \\
&= \arg\max_j \sum_{i=1}^N \mathbf{1}\{h_j(q_{(i)}) = y_{(i)}\} \\
&= \arg\max_j \left(\sum_{i=1}^j \mathbf{1}\{y_{(i)} = 1\}\right) + \left(\sum_{i=j+1}^N \mathbf{1}\{y_{(i)} = 2\}\right) \\
&= \arg\max_j \left(\sum_{i=1}^j \mathbf{1}\{y_{(i)} = 1\}\right) + \left(\sum_{i=j+1}^N \mathbf{1}\{y_{(i)} = 2\}\right) - \left(\sum_{i=1}^N \mathbf{1}\{y_{(i)} = 2\}\right) \\
&= \arg\max_j \sum_{i=1}^j \left(\mathbf{1}\{y_{(i)} = 1\} - \mathbf{1}\{y_{(i)} = 2\}\right)
\end{aligned}$$

□

B. Theoretical Analysis

In this section, we analyze the performance of GALAXY under random label noise and show the probability of identifying and sampling around additional cuts increases as more examples are labeled. This is in contrast to the DIRECT's agnostic active learning approach, where the probability of identifying and sampling around only the optimal separation threshold decays exponentially in the number of labeling budget.

Specifically, under the binary classification scenario, one is given a sorted list of N examples $\{x_{(i)}\}_{i=1}^N$, with ground truth labels $y^\star_{(1)} = y^\star_{(2)} = ... = y^\star_{(N_1)} = 1$ and $y^\star_{(N_1+1)} = ... = y^\star_{(N_1+N_2)} = 2$, where $N_1 + N_2 = N$. Under uniform i.i.d. label noise with noise ratio $\eta > 0$, the *observed labels* are denoted as $\{y_{(i)}\}_{i=1}^N$, where $\mathbb{P}(y_{(i)} \neq y^\star_{(i)}) = \eta$. In other words, the observed label is flipped with probability η.

Theorem B.1. *Given a budget of $b > 2\log N$, let M_b be the random variable denoting number of identified cuts in addition to the optimal separation threshold by one round of GALAXY. We must have $\mathbb{P}(M_b \ge 1) \ge 1 - \exp(-b\log(\frac{1}{1-\eta})/2)$, implying GALAXY samples around at least one more cut in addition to the optimal separation threshold with high probability.*

Proof. In the perfect scenario where GALAXY does not receive any corrupted labels, it would use $\log N$ budget with bisection to find the optimal separation threshold and annotate around it. However, within the first $\frac{b}{2}$ annotations, whenever GALAXY receives a corrupted label, it will identify a cut in addition to the optimal separation threshold, i.e., $M_b \ge 1$. Therefore, the probability of $M_b \ge 1$ is greater than the probability of receiving at least one corrupted labels in the first $\frac{b}{2}$ annotations. With simple probability bound, we can show that

$$\mathbb{P}(M_b \ge 1) > 1 - (1-\eta)^{b/2} = 1 - \exp(b\log(1-\eta)/2) = 1 - \exp(-b\log(\frac{1}{1-\eta})/2).$$

□

Name	K	N	Imb Ratio $\gamma = \frac{\min_k N_k}{\max_{k'} N_{k'}}$
Imb CIFAR-10	2	50000	.1111
Imb CIFAR-10	3	50000	.1250
Imb CIFAR-100	2	50000	.0101
Imb CIFAR-100	3	50000	.0102
Imb CIFAR-100	10	50000	.0110
Imb SVHN	2	73257	.0724
Imb SVHN	3	54448	.2546
PathMNIST	2	89996	.1166
FMoW	62	76863	.0049
iWildCam	14	129809	$4.57 \cdot 10^{-5}$

Table 1: Dataset settings for our experiments. N denotes the total number of examples in our dataset. γ is the class imbalance ratio defined in Section 3.1.

As the theorem suggests, when b is large, GALAXY will identify and annotate around at least one additional cut with high probability.

C. Time Complexity

The computation complexity for each batch of DIRECT is $O(KN\log(N) + B_{\text{train}}N)$ for data selection plus the training and inference costs of the neural network. $O(KN\log(N))$ comes from sorting examples by their margin scores for each class and $O(B_{\text{train}}N)$ is the cost for running Algorithm 2 for $O(B_{\text{train}})$ iterations. Each iteration of Algorithm 2 only costs $O(N)$ time as we can efficiently solve the objective by cumulative sums. We note that the cost associated with neural network training and inference is always the dominating factor.

For comparisons, BADGE has time complexity $O(B_{\text{train}}N(K+D))$, significantly more expensive than DIRECT, with D denotes the dimensionality of the penultimate layer features. In addition, GALAXY has computational complexity of $O(KN\log(N)) + B_{\text{train}}KN$, also more expensive than DIRECT. In all of our experiments, both BADGE and GALAXY indeed is slower than DIRECT. We further note that the time complexity factor of K in DIRECT can be easily parallelized by conducting the K sorting procedures on different CPU cores.

Below, we also provide a comprehensive list of computational complexity of different algorithms we consider in our implementation: As for data selection algorithms, let K be the number of classes, N be the pool size and B_{train} be the batch size, D be the penultimate layer embedding dimension and T be the number of batches. Below, we detail the computation cost of data selection of each algorithm we consider.

- DIRECT: $O(T(KN\log N + B_{train}N))$.
- GALAXY: $O(T(KN\log N + B_{train}KN))$
- BADGE: $O(TB_{train}N(K+D))$
- Margin sampling/most likely positive/confidence sampling: $O(TKN)$
- Coreset: $O(T^2B_{train}ND)$
- SIMILAR: $O(TB_{train}ND)$
- Cluster margin: $O(N^2\log N + TN(K+\log N))$
- BASE: $O(TN(D+B_{train}))$

Overall, our experiments are conducted on NVIDIA 3090 ti GPUs. Each trial of the ResNet-18 experiment takes less than two hours while each trial of the LabelBench experiments takes roughly 12 hours.

D. Experiment Setup

ResNet-18 Experiments. ResNet-18 with passive training has been the standard evaluation in existing deep active literature (Ash et al., 2019; Zhang et al., 2022). Our experiment setup utilizes the CIFAR-10, CIFAR-100 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2011) and PathMNIST (Yang et al., 2021) image classification datasets. The original forms of these datasets are roughly balanced across 9, 10 or 100 classes. We construct an extremely imbalanced dataset by grouping a

large number of classes into one majority class. For example, given a balanced dataset above with M classes. We generate an imbalanced dataset with K classes ($K < M$) by the first $K-1$ classes from the original dataset and combining the rest of the classes $K, ..., M$ into a single majority class K. Imbalance ratios are shown in Table 1. In addition, we also utilize the standard CIFAR-10LT and CIFAR-100LT variants in our experiments for noisy label setting.

For neural network training, we utilize the standard passive training on labeled examples with cross entropy loss and Adam optimizer (Kingma & Ba, 2014). The ResNet-18 model (He et al., 2016) is pretrained on ImageNet (Deng et al., 2009) from the PyTorch library. To address data imbalance, for all algorithms, we utilize a reweighted cross entropy loss by the inverse frequency of the number of labeled examples in each class. For experiments with label noise, we further add a 10% label smoothing during training (Müller et al., 2019) for all algorithms.

LabelBench Experiments. Proposed by Zhang et al. (2024a), LabelBench evaluates active learning performance in a more comprehensive framework. Here, we fine-tune the large pretrained model from CLIP's ViT-B32 model (Radford et al., 2021). The framework also utilizes semi-supervised learning method FlexMatch (Zhang et al., 2021) to further leverage the unlabeled examples in the pool for training. We conduct experiments on the two imbalanced datasets in LabelBench, with FMoW (Christie et al., 2018) and iWildcam (Beery et al., 2021). Similar to the ResNet-18 experiments, for all algorithms, we use a 10% label smoothing in the loss function to improve training under label noise. We did find FlexMatch to perform poorly under the combination of imbalance and label noise, so we used the passive training method for label noise experiments.

E. All Results

E.1. Noiseless Results under Imbalance

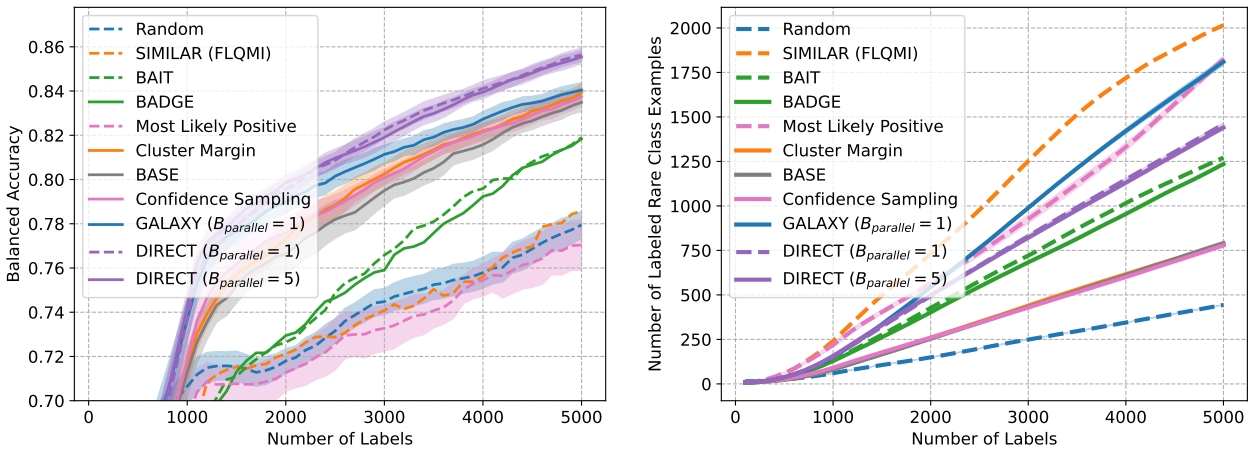


Figure 8: Imbalanced CIFAR-10, two classes.

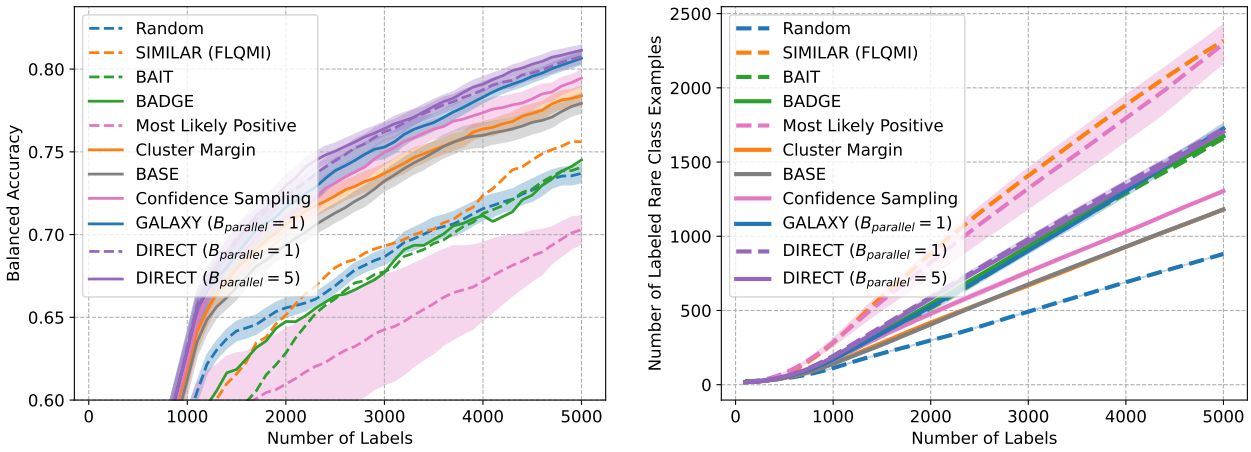


Figure 9: Imbalanced CIFAR-10, three classes.

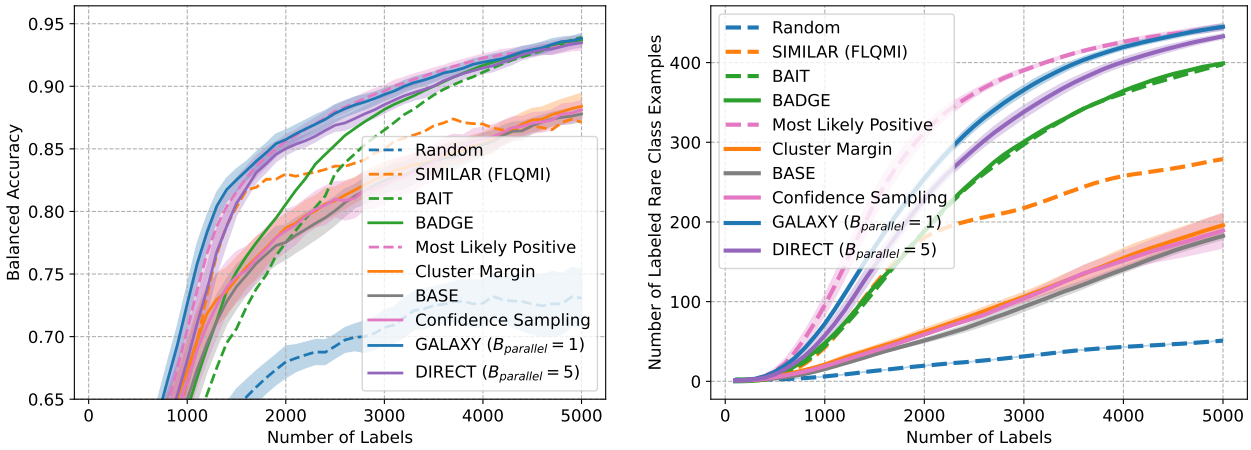


Figure 10: Imbalanced CIFAR-100, two classes.

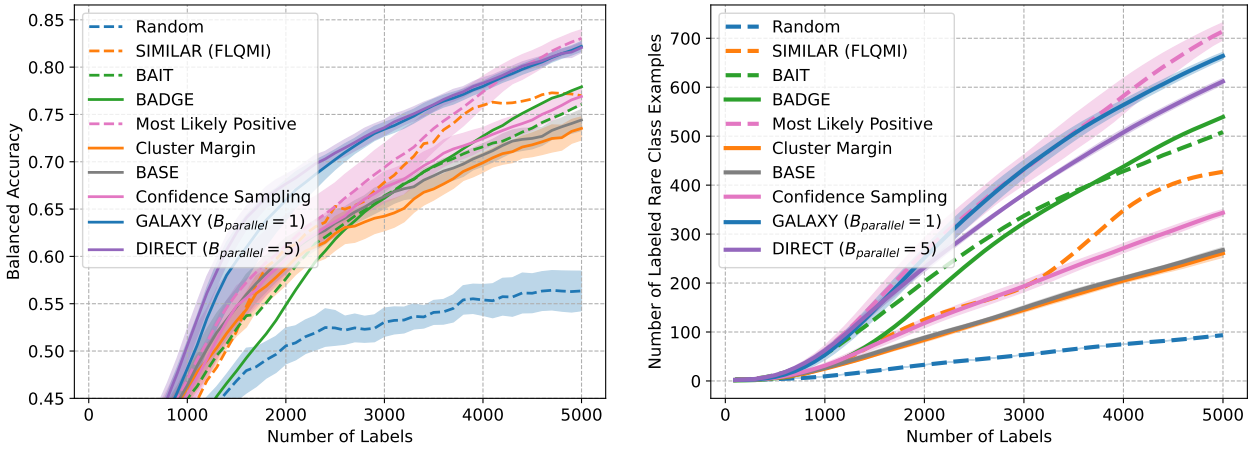


Figure 11: Imbalanced CIFAR-100, three classes.

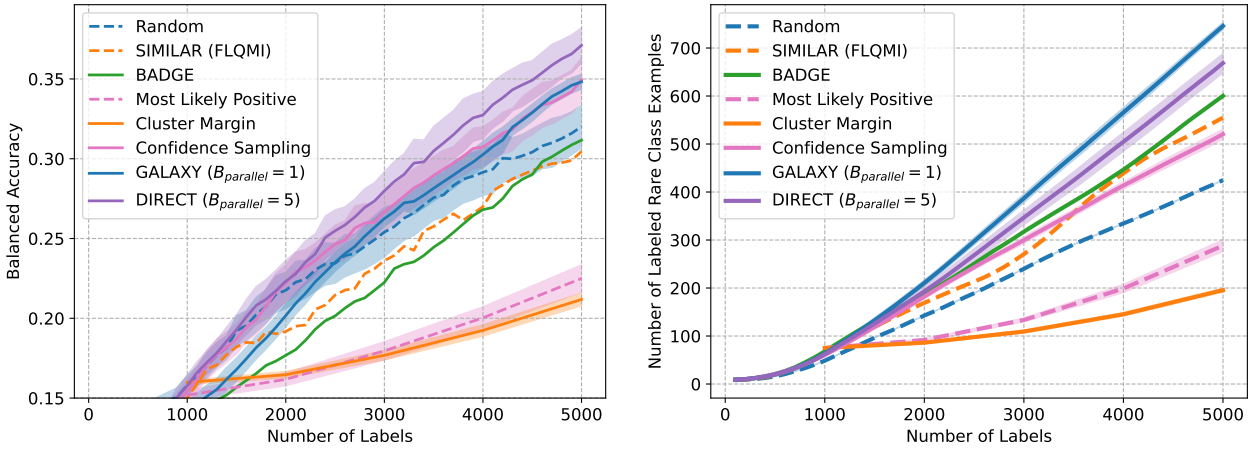


Figure 12: Imbalanced CIFAR-100, 10 classes.

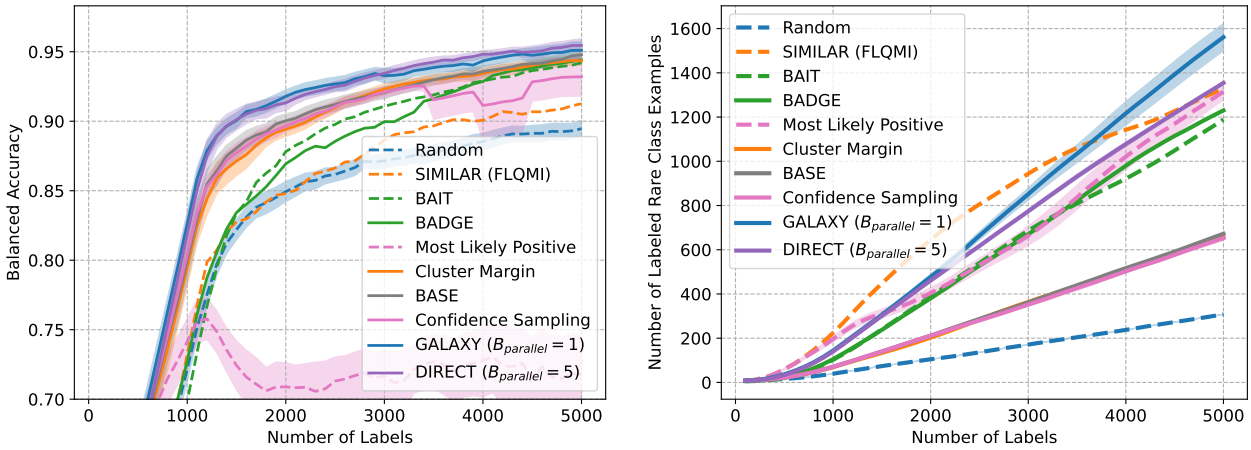


Figure 13: Imbalanced SVHN, two classes.

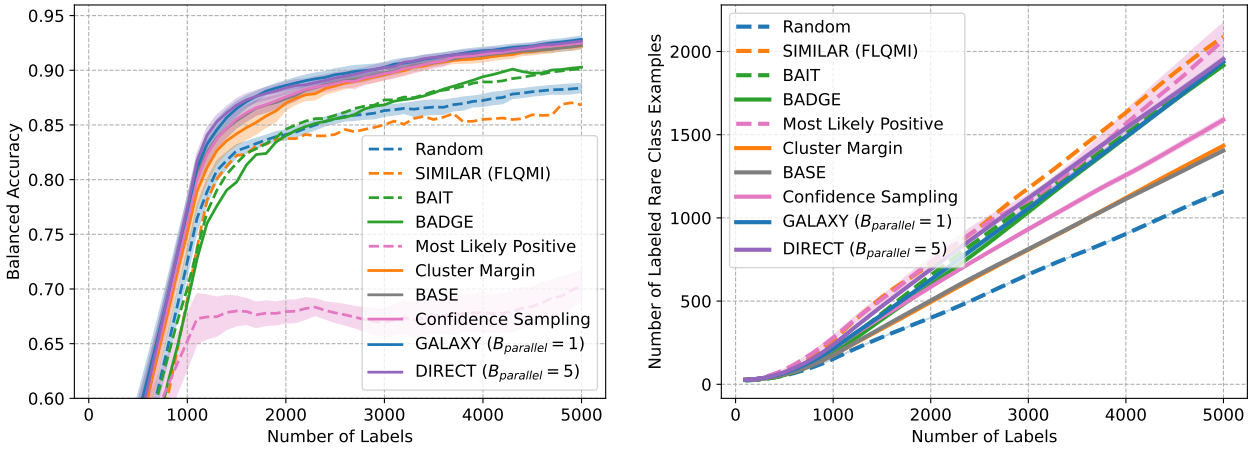


Figure 14: Imbalanced SVHN, three classes.

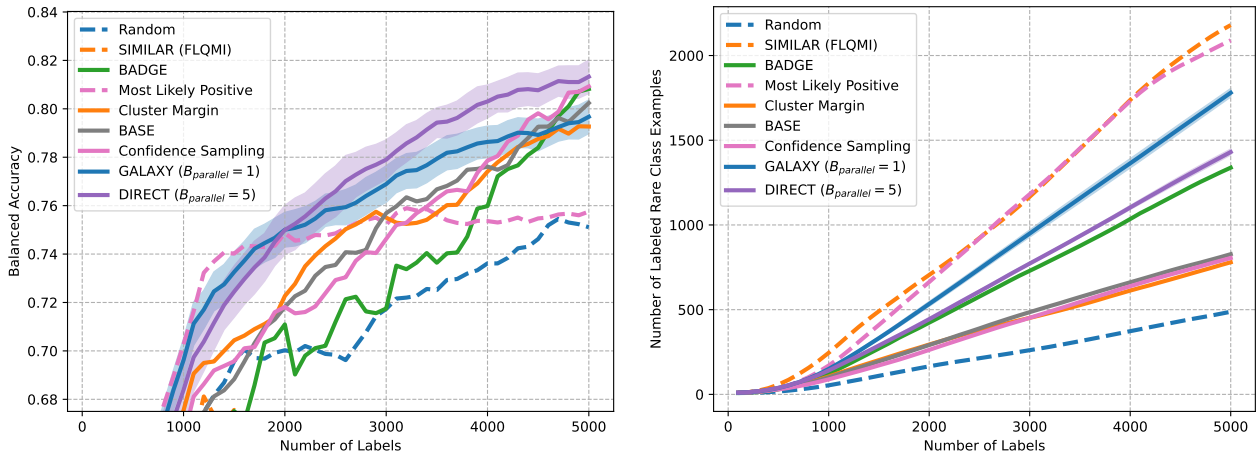
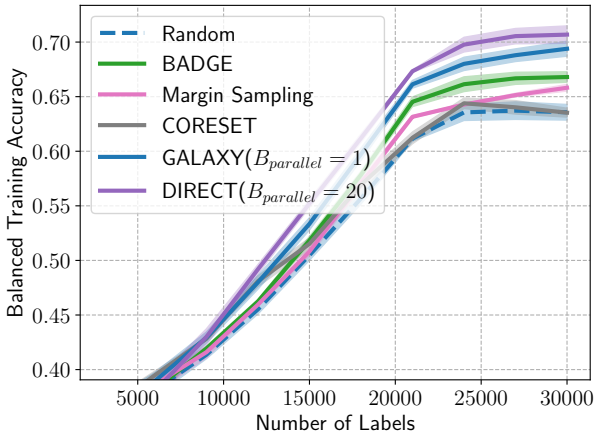
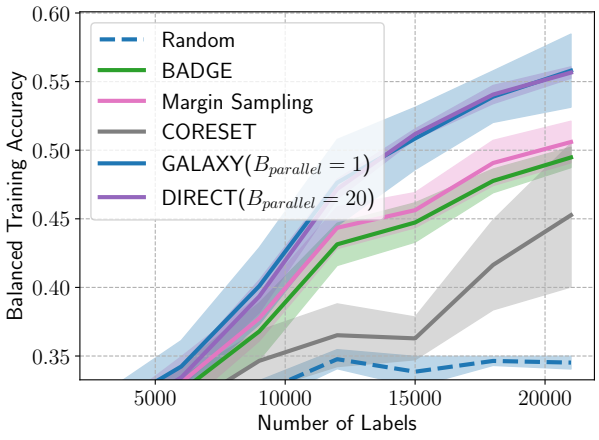


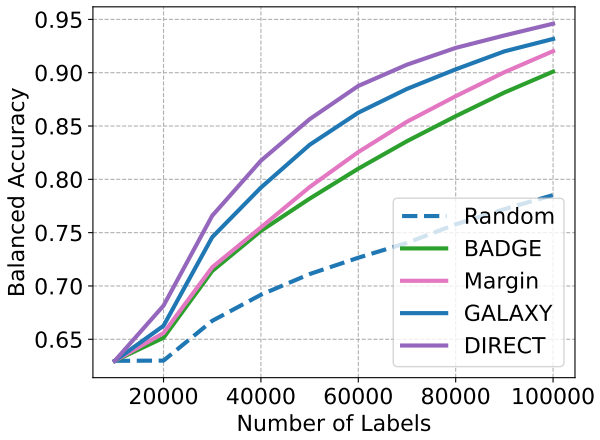
Figure 15: PathMNIST, two classes.



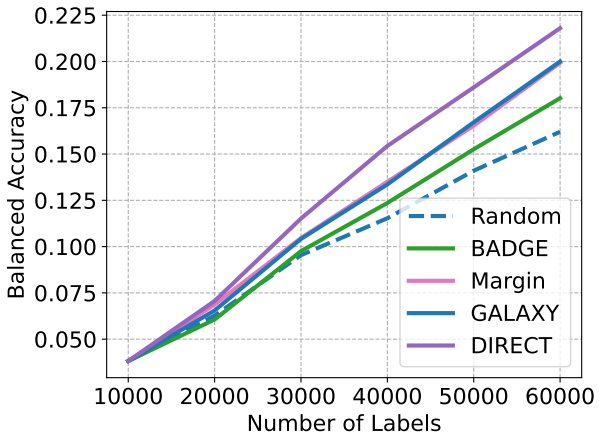
(a) FMoW Balanced Pool Accuracy



(b) iWildcam Balanced Pool Accuracy



(c) ImageNet-LT Balanced Pool Accuracy



(d) iNaturalist Balanced Pool Accuracy

Figure 16: LabelBench results in the noiseless setting.

E.2. Label Noise Results under Imbalance

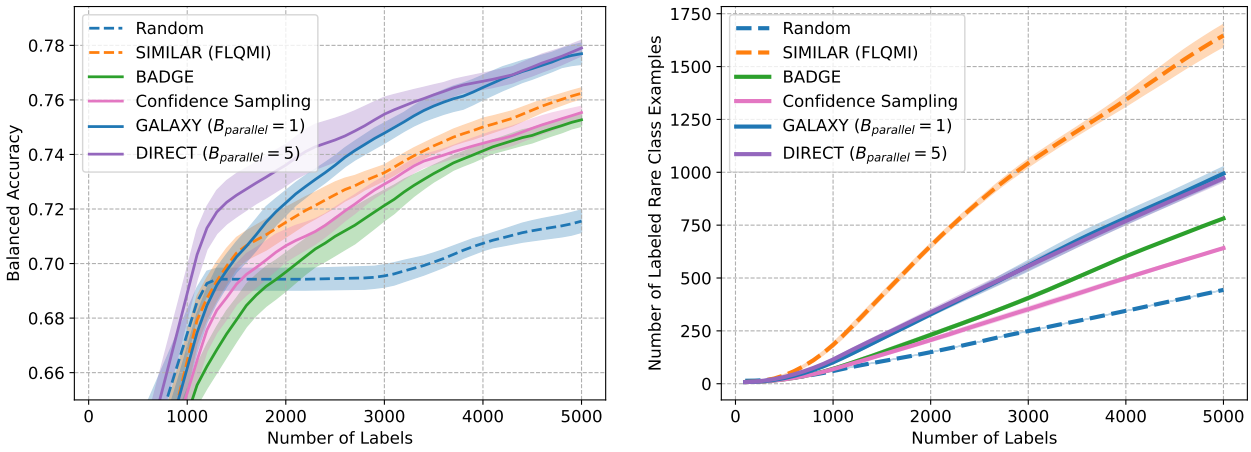


Figure 17: Imbalanced CIFAR-10, two classes, 10% label noise.

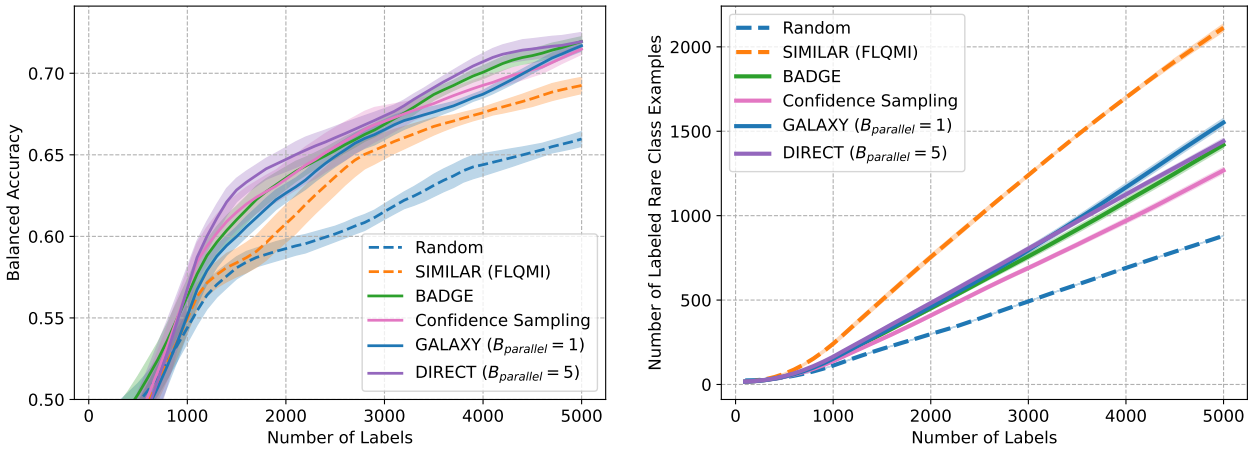


Figure 18: Imbalanced CIFAR-10, three classes, 10% label noise.

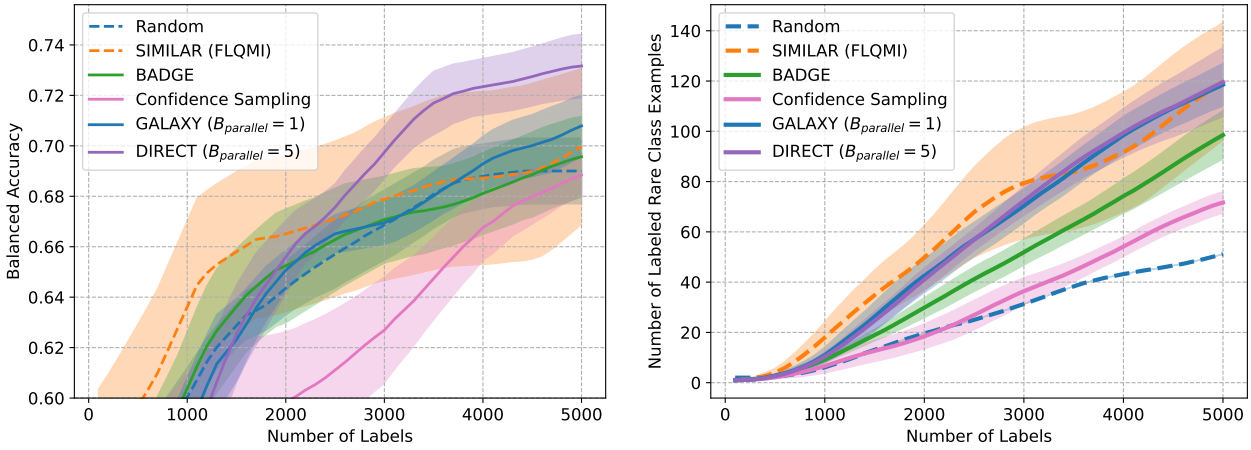


Figure 19: Imbalanced CIFAR-100, two classes, 10% label noise.

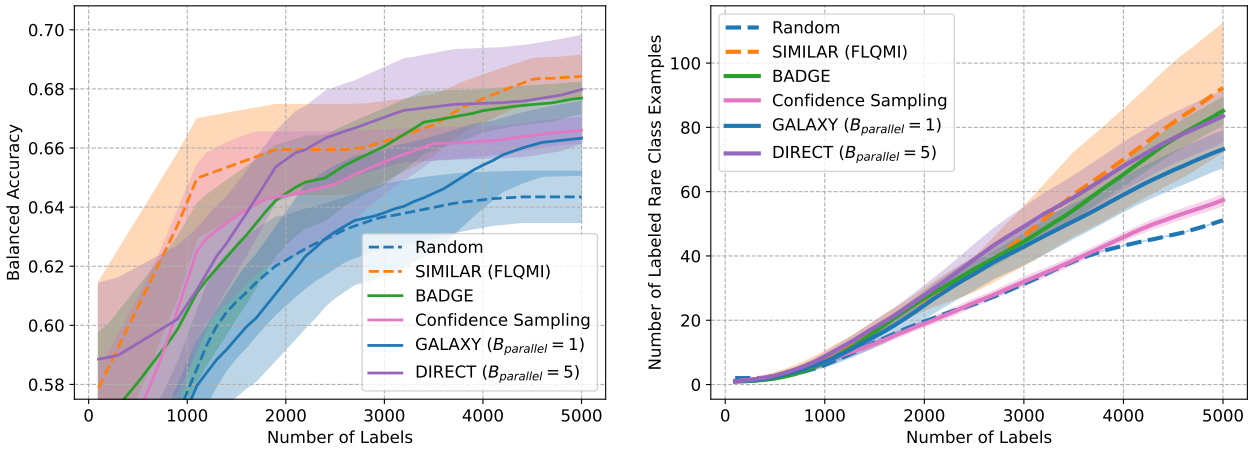


Figure 20: Imbalanced CIFAR-100, two classes, 15% label noise.

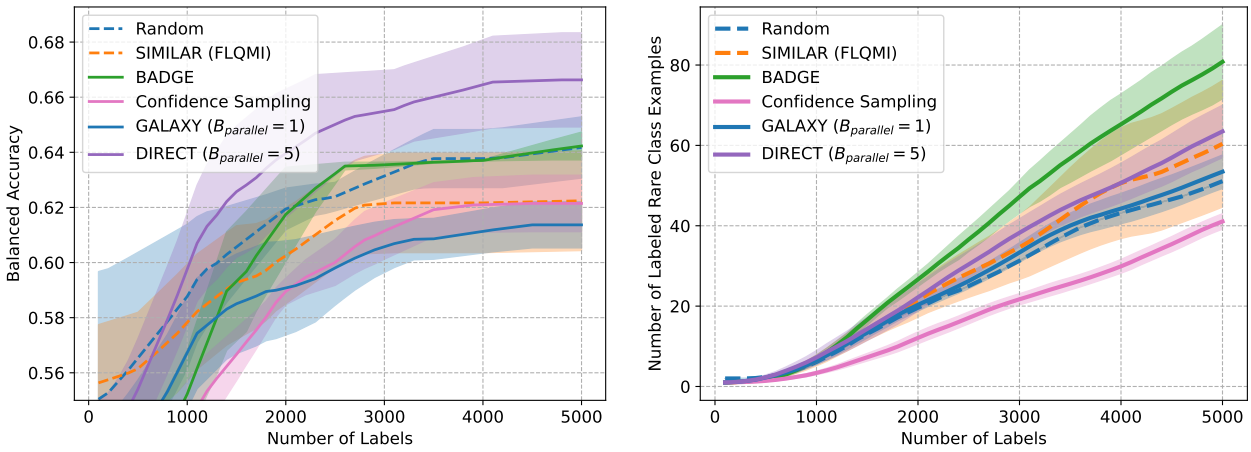


Figure 21: Imbalanced CIFAR-100, two classes, 20% label noise.

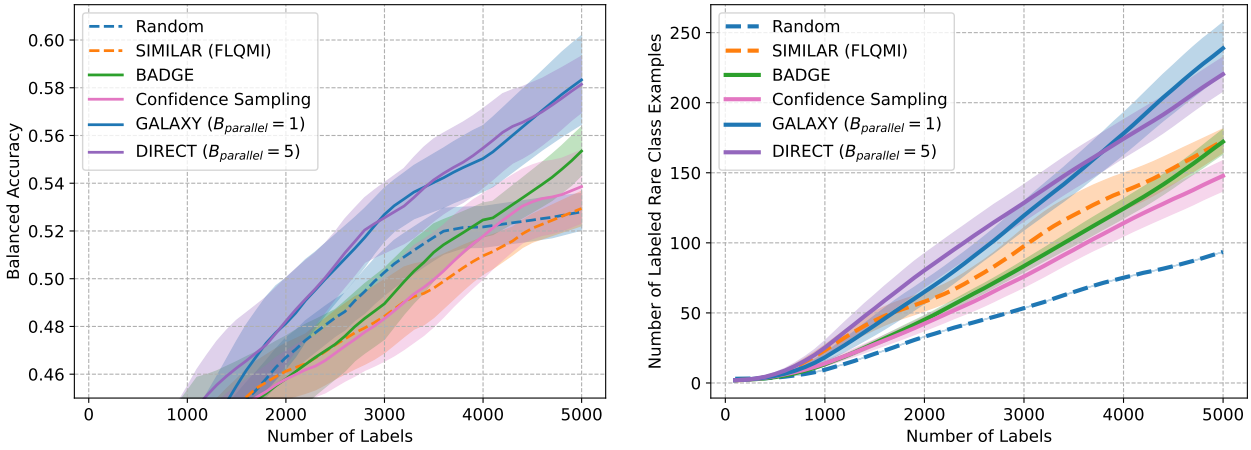


Figure 22: Imbalanced CIFAR-100, three classes, 10% label noise.

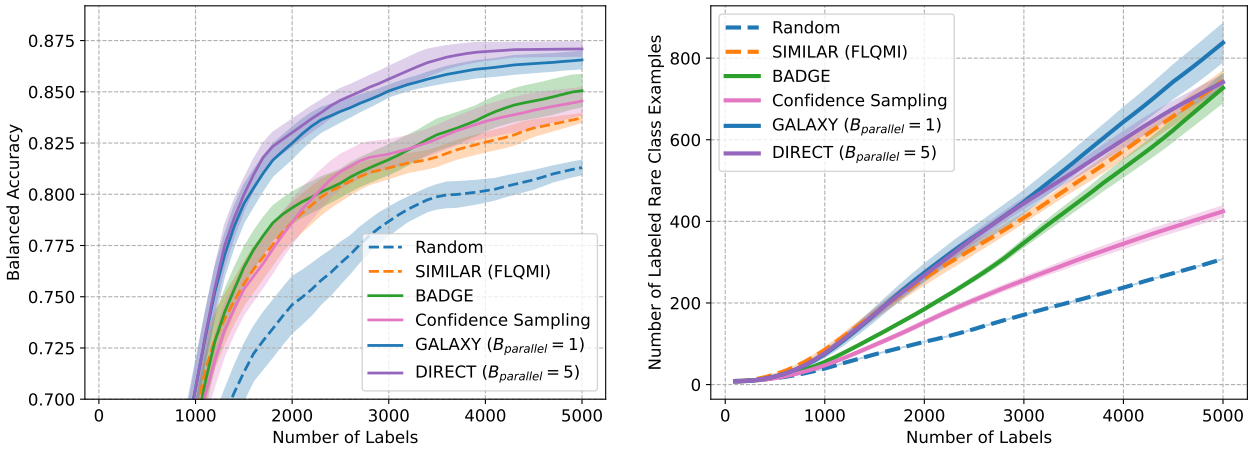


Figure 23: Imbalanced SVHN, two classes, 10% label noise.

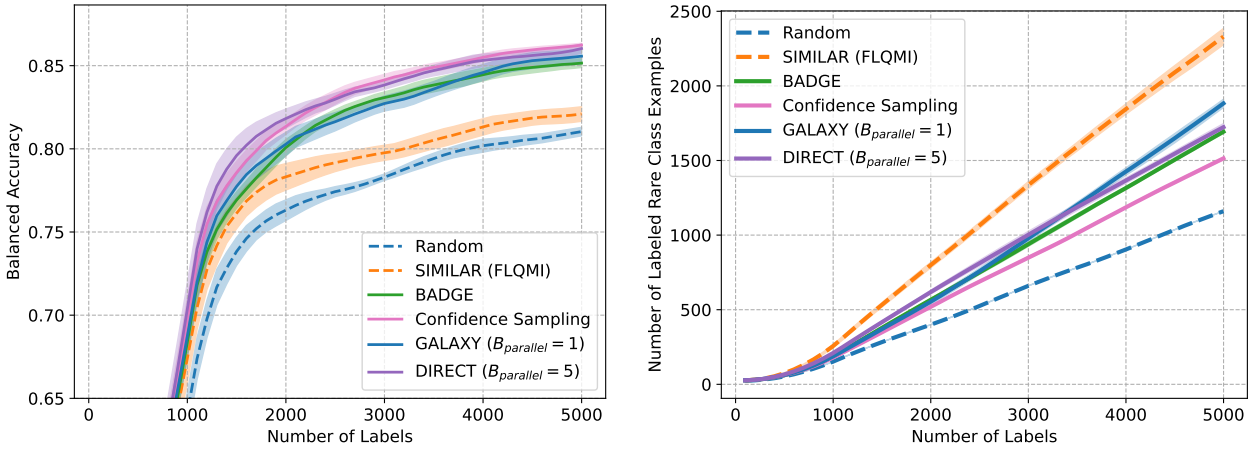
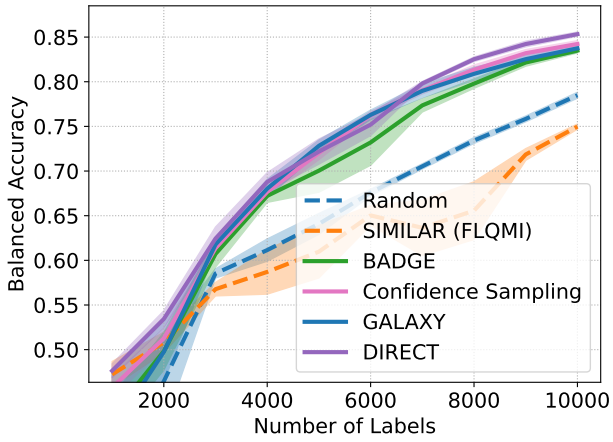
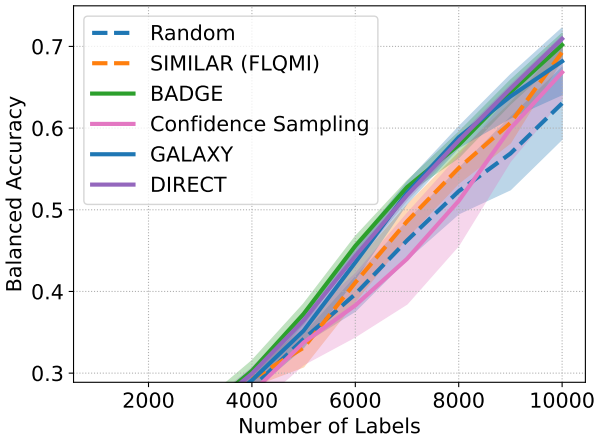


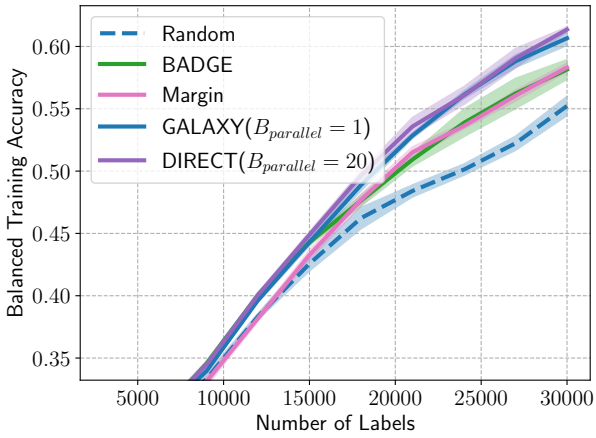
Figure 24: Imbalanced SVHN, three classes, 10% label noise.



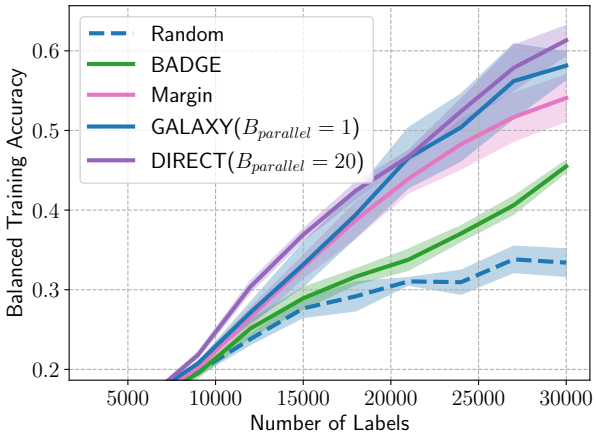
(a) CIFAR-10LT, 10 classes, 15% label noise.



(b) CIFAR-100LT, 100 classes, 15% label noise.



(a) FMoW Balanced Pool Accuracy



(b) iWildcam Balanced Pool Accuracy

Figure 26: LabelBench results in the 10% label noise setting.