ACTIVE LEARNING: SAMPLING IN THE LEAST PROBABLE DISAGREEMENT REGION

Anonymous authors
Paper under double-blind review

ABSTRACT

Active learning strategy to query samples closest to the decision boundary can be an effective strategy for sampling the most uncertain and thus informative samples. This strategy is valid only when the sample's "closeness" to the decision boundary can be estimated. As a measure for evaluating closeness to a given decision boundary of a given sample, this paper considers the least probable disagreement region (LPDR) which is a measure of the smallest perturbation on the decision boundary leading to altered prediction of the sample. Experimental results show that the proposed LPDR-based active learning algorithm consistently outperforms other high performing active learning algorithms and leads to state-of-the-art performance on various datasets and deep networks.

1 INTRODUCTION

Active learning (Cohn et al., 1996) is a subfield in machine learning for attaining sample efficiency by intelligently selecting a small subset of unlabeled samples for their labels to be used in training. In many real-world learning problems, a large collection of unlabeled samples is assumed available, and the labels of the most informative samples are iteratively queried for retraining the model based on various query strategies such as uncertainty sampling (Lewis & Gale, 1994; Scheffer et al., 2001; Culotta & McCallum, 2005; Wang et al., 2010; Nguyen et al., 2021), model change (Settles et al., 2008; Freytag et al., 2014; Ash et al., 2020), Bayesian active learning (Pinsler et al., 2019; Shi & Yu, 2019), core-set (Sener & Savarese, 2018), error reduction (Roy & McCallum, 2001; Yoo & Kweon, 2019), variance reduction (Schein & Ungar, 2007), discriminative sampling (Sinha et al., 2019; Gissin & Shalev-Shwartz, 2019; Zhang et al., 2020; Gu et al., 2020), feature matching between unlabeled and validation dataset (Gudovskiy et al., 2020), active Thomson sampling (Bouneffouf et al., 2014), minimize the redundancy by clustering (Yang et al., 2021), two-way exploration (Zhang et al., 2015), and adaptive batch mode (Chakraborty et al., 2014). Active learning attempts to achieve high accuracy using as few labeled samples as possible.

In uncertainty-based sampling–the most popular strategy, quantifying precisely the uncertainty remains an open question, and this paper is focused on this issue. This strategy is often considered for its simplicity and relatively low computational load, and it enhances the performance of the current model by utilizing the labels of the unlabeled samples whose predicted class is most vague (Settles, 2009; Yang et al., 2015; Sharma & Bilgic, 2017). It is generally understood that unlabeled sample closest to the decision boundary is the most informative as the sample is the most uncertain (Balcan et al., 2007; Kremer et al., 2014; Ducoffe & Precioso, 2018). Balcan et al. theoretically show that selecting unlabeled samples with the smallest margin to the decision boundary attains exponential improvement over random sampling in terms of sample complexity for binary classification with linear separators (Balcan et al., 2007). However, many uncertainty-based sampling for multiclass classification with deep network do not take into account the closeness of the sample to the decision boundary for the reason that in multiclass classification with deep network, it is difficult to identify samples closest to the decision boundary as the sample's closeness based on Euclidean distance is often not readily measurable (Ducoffe & Precioso, 2018; Mickisch et al., 2020).

This paper proposes a closeness measure that can be evaluated in multiclass classification with deep network as a measure of uncertainty. We assume that the most uncertain and thus the most informative samples will have their labels most "sensitive" to the smallest perturbation of the decision boundary, and that these samples are "closest" to the decision boundary. Here, the closeness are

defined as the sample's sensitiveness to the perturbation of the decision boundary, and it is based on the disagree metric between the decision boundary and its perturbation. The main contributions of this paper are summarized as follows.

1. This paper defines a measure of sample's closeness to the decision boundary referred to as the *least probable disagreement region* (LPDR) based on the disagree metric between hypotheses.
2. This paper introduces a hypothesis sampling method with a measure of disagreement in sampled hypotheses, referred to as the *disagree ratio*, for obtaining the sample order in terms of LPDR without evaluating the LPDR.
3. This paper proposes a high performing active learning algorithm of querying unlabeled samples closest to the decision boundary in terms of LPDR.

2 Related Work

Various forms of uncertainty measure have been studied. *Entropy* (Shannon, 1948) based uncertainty sampling strategy queries unlabeled samples yielding the maximum entropy from the predictive distribution, but it does not perform well for multiclass-classification tasks as entropy does not equate well with the closeness to the decision boundary (Joshi et al., 2009). *Mutual information* based strategy which includes the BALD (Houlsby et al., 2011), DBAL (Gal et al., 2017), and BatchBALD (Kirsch et al., 2019) queries unlabeled samples yielding the maximum mutual information between predictions and model parameters. The DBAL approximates the posterior of the model parameters of deep network by MC-dropout sampling, but each batch selection is independently conducted, and this leads to data inefficiency as correlations between data points in the batch are not taken into account (Kirsch et al., 2019). To address this deficiency, BatchBALD is introduced, but BatchBALD theoretically computes all possible mutual information between batch-wise predictions and model parameters, and for this reason, it is not appropriate for large query size. *Variation ratio* (Freeman, 1965) with ensemble method (Beluch et al., 2018) based on query by committee (QBC) strategy (Seung et al., 1992) queries unlabeled samples yielding the maximum variation ratio in labels predicted by the multiple networks, but it requires high computational load: each network belonging to the ensemble must be individually trained. *Gradient* based strategy (Ash et al., 2020) measures uncertainty as the gradient magnitude with respect to parameters in the final layer and queries unlabeled samples where these gradients span a diverse set of directions, but it requires high computational load when the dimension of parameters is large.

3 Least Probable Disagreement Region (LPDR)

This section theoretically defines LPDR and proposes "empirical LPDR" to approximate LPDR based on sampling from the instance and hypothesis spaces. In addition, a brute-force method for evaluating the empirical LPDR of each sample is described in determining the order of the closeness to the decision boundary.

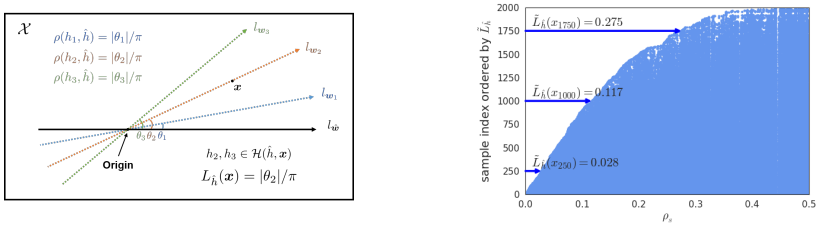
3.1 Definition of LPDR

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{D}$, and $\mathcal{H}$ be respectively the instance space, the label space, the joint distribution over $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$, and the hypothesis space of $h : \mathcal{X} \to \mathcal{Y}$. The (pseudo) metric between two hypotheses, referred to as *disagree metric*, is defined as the probability of the disagreement region in $\mathcal{X}$ where labels are predicted differently by the two hypotheses (Hanneke et al., 2014; Hsu, 2010). For $h_1, h_2 \in \mathcal{H}$, the disagree metric between h_1 and h_2 is defined as:

$$\rho(h_1, h_2) := \mathbb{P}_{X \sim \mathcal{D}}[h_1(X) \neq h_2(X)]$$

where X is random variable from $\mathcal{D}$. The LPDR of a sample for a given hypothesis is defined as the least probable disagreement region for the hypothesis that contains the sample. For given $\hat{h} \in \mathcal{H}$, let $\mathcal{H}(\hat{h}, \boldsymbol{x})$ be the set of hypotheses disagreed with $\hat{h}$ on $\boldsymbol{x} \in \mathcal{X}$:

$$\mathcal{H}(\hat{h}, \boldsymbol{x}) := \{h \in \mathcal{H} : h(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})\},$$



(a) (b)

Figure 1: Examples of LPDR and empirical LPDR. (a) LPDR of $\boldsymbol{x}$ for given $\hat{h}$ in binary classification with the linear classifier. Here $\boldsymbol{x}$ is uniformly distributed in $\mathbb{R}^2$. The h_2 and h_3 disagree with $\hat{h}$ in prediction on $\boldsymbol{x}$, and $|\theta_2|/\pi$ is the infimum for $\{\rho(h, \hat{h}) : h \in \mathcal{H}(\hat{h}, \boldsymbol{x})\}$. Thus, $L_{\hat{h}}(\boldsymbol{x}) = |\theta_2|/\pi$. (b) Empirical LPDR of $\boldsymbol{x}_i$ for $\hat{h}$ on MNIST dataset. The set of x-axis values of the blue dots for $\boldsymbol{x}_i$ are $\{\rho_S(h_c, \hat{h}) : h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x}_i)\}$, and $\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) = \min\{\rho_S(h_c, \hat{h}) : h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x}_i)\}$ (blue arrow).

then the LPDR of $\boldsymbol{x}$ for $\hat{h}$ is defined as follows:

$$L_{\hat{h}}(\boldsymbol{x}) := \inf_{h \in \mathcal{H}(\hat{h}, \boldsymbol{x})} \rho(h, \hat{h}).$$

The LPDR of $\boldsymbol{x}$ for $\hat{h}$ is the measure of closeness in terms of the disagree metric between $\hat{h}$ and h that can alter the predicted label of $\boldsymbol{x}$. The sample with the smallest LPDR is assumed to be the sample most sensitive to the small perturbation of the decision boundary and thus closest to the decision boundary.

Figure 1a shows an example of LPDR of $\boldsymbol{x}$ for given $\hat{h}$ in the binary classification with a set of linear classifier, $\mathcal{H} = \{h : h(\boldsymbol{x}) = \text{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}), \boldsymbol{w} \in \mathcal{W} = \mathbb{R}^2\}$ where $\boldsymbol{x}$ is uniformly distributed on $\mathcal{X} = \mathbb{R}^2$. In $\mathcal{X}$, $\boldsymbol{x}$ is a data point and $\boldsymbol{w}$ is represented as $l_{\boldsymbol{w}} = \{\boldsymbol{x} : \boldsymbol{x}^\mathsf{T}\boldsymbol{w} = 0\}$. Let θ_i be the angle between $l_{\boldsymbol{w}_i}$ and $l_{\hat{\boldsymbol{w}}}$, then the $\rho(h_i, \hat{h}) = |\theta_i|/\pi$ where the unit of θ_i is radian and $-\pi \leq \theta_i \leq \pi$ since $h_i(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})$ for all $\boldsymbol{x}$ between $l_{\boldsymbol{w}_i}$ and $l_{\hat{\boldsymbol{w}}}$. Here, $h_1(\boldsymbol{x}) = \hat{h}(\boldsymbol{x})$, while $h_2(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})$ and $h_3(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})$, thus $h_2, h_3 \in \mathcal{H}(\hat{h}, \boldsymbol{x})$. In this case, $|\theta_2|/\pi$ is the infimum for $\{\rho(h, \hat{h}) : h \in \mathcal{H}(\hat{h}, \boldsymbol{x})\}$, therefore $L_{\hat{h}}(\boldsymbol{x}) = |\theta_2|/\pi$.

Henceforth, unless otherwise stated, $\hat{h}$ will denote the hypothesis learned from labeled samples denoted by $\mathcal{L}$ and LPDR for $\hat{h}$ is evaluated on unlabeled samples denoted by $\mathcal{U}$ in this paper.

3.2 Empirical LPDR

In general, it is infeasible to explicitly evaluate LPDR for the following two reasons: 1) ρ cannot be explicitly evaluated when $\mathcal{D}$ is unknown, and 2) it is difficult to evaluate $\rho(h, \hat{h})$ for all $h \in \mathcal{H}$ especially when $|\mathcal{H}| = \infty$.

To address the first reason, $\rho(h, \hat{h})$ can be approximated as

$$\rho_S(h, \hat{h}) = \frac{1}{S} \sum_{i=1}^{S} \mathbb{I}\big[h(X_i) \neq \hat{h}(X_i)\big],$$

where $\mathbb{I}[\cdot]$ is an indicator function, and $X_1, \ldots, X_S$ are i.i.d. random variables from $\mathcal{D}$. By strong law of large number, $|\rho_S(h, \hat{h}) - \rho(h, \hat{h})| \to 0$ with probability 1 as $S \to \infty$.

To address the second issue, we consider $\mathcal{H}_C \subset \mathcal{H}$ of size C satisfying the following property.

Property 1 *For any given $\boldsymbol{x} \in \mathcal{U}$ and any $h \in \mathcal{H}$ with $\rho(h, \hat{h}) \leq \max_{\boldsymbol{x} \in \mathcal{U}} L_{\hat{h}}(\boldsymbol{x})$, there exists $h_c \in \mathcal{H}_C \subset \mathcal{H}$ satisfying that $h_c(\boldsymbol{x}) = h(\boldsymbol{x})$ and that for any $\epsilon > 0$,*

$$|\rho(h, \hat{h}) - \rho(h_c, \hat{h})| < \epsilon$$

as $C \to \infty$.

For $\mathcal{H}_C$ satisfying Property 1, we define empirical LPDR as follows:

$$\tilde{L}_{\hat{h}}(\boldsymbol{x}) := \inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho_S(h_c, \hat{h}). \tag{1}$$

Here for given $\boldsymbol{x} \in \mathcal{U}$ and $\hat{h}$, $\mathcal{H}_C(\hat{h}, \boldsymbol{x}) = \{h_c \in \mathcal{H}_C : h_c(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})\}$. The empirical LPDR converges to LPDR and the rank-orders based on the empirical LPDR and LPDR are equivalent when

$$\frac{\log C}{S} \to 0 \tag{2}$$

as $\min(S, C) \to \infty$ by the following theorem.

Theorem 1 *Suppose that $\mathcal{H}_C$ is the set of hypotheses satisfying Property 1 where $\mathcal{U} = \{\boldsymbol{x}_i\}_{i=1}^M$ and Eq. 2 holds, then the following two statements can be made.*

1. *(convergence) For any $\boldsymbol{x} \in \mathcal{U}$,*

$$|\tilde{L}_{\hat{h}}(\boldsymbol{x}) - L_{\hat{h}}(\boldsymbol{x})| \to 0.$$

2. *(rank-order consistency) If $L_{\hat{h}}(\boldsymbol{x}_i) \neq L_{\hat{h}}(\boldsymbol{x}_j)$ for $i \neq j$, then*

$$\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) < \tilde{L}_{\hat{h}}(\boldsymbol{x}_j) \Longrightarrow L_{\hat{h}}(\boldsymbol{x}_i) < L_{\hat{h}}(\boldsymbol{x}_j)$$

with probability tending to 1 as $\min(S, C) \to \infty$, where S is the number of i.i.d. random variables for approximating ρ and $C = |\mathcal{H}_C|$.

The proof of Theorem 1 is deferred to Appendix A.1.

3.3 Brute-force Search for Empirical LPDR

To construct $\mathcal{H}_C$ satisfying Property 1, we consider Gaussian sampling of parameters corresponding to hypotheses with gradually increasing variance based on the following conjecture:

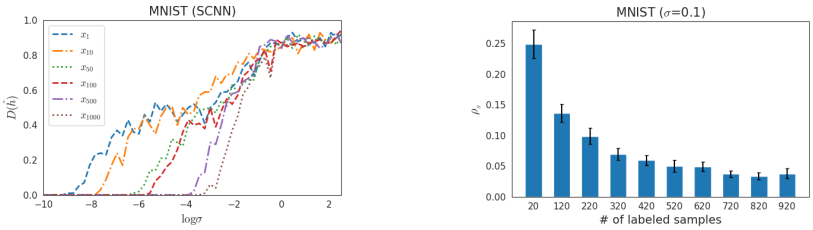
Conjecture 1 *Suppose that h is sampled with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ where $\boldsymbol{w}, \hat{\boldsymbol{w}} \in \mathcal{W}$ are parameters of $h, \hat{h}$ respectively, then $\mathbb{E}[\rho(h, \hat{h})]$ is continuous and strictly increasing with σ.*

The theoretical and empirical verifications of Conjecture 1 are presented in Appendix C.1. At first, many $h^{(k)}$ are sampled with $\boldsymbol{w}^{(k)} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma_k^2)$, and a set of hypotheses $\mathcal{H}^{(k)}$ is constructed as $\mathcal{H}^{(k)} = \{h_n^{(k)} : (k-1)\epsilon < \rho_S(h_n^{(k)}, \hat{h}) < k\epsilon\}_{n=1}^N$ by the selection of sampled hypotheses. Then, $\mathcal{H}_C$ can be constructed as:

$$\mathcal{H}_C = \bigcup_{k=1}^K \mathcal{H}^{(k)}$$

with $\{\sigma_k\}_{k=1}^K$ such that $\sigma_k < \sigma_{k+1}$ where K is the smallest value satisfying that $\mathcal{H}_C(\hat{h}, \boldsymbol{x}) \neq \emptyset$ for all $\boldsymbol{x} \in \mathcal{U}$. Here, for any $h \in \mathcal{H}$ with $\rho(h, \hat{h}) \leq \max_{\boldsymbol{x} \in \mathcal{U}} L_{\hat{h}}(\boldsymbol{x})$, there exists $k \in [K]$ such that $(k-1)\epsilon \leq \rho(h, \hat{h}) \leq k\epsilon$, and thus $|\rho_S(h_n^{(k)}, \hat{h}) - \rho(h, \hat{h})| < \epsilon$ for all $h_n^{(k)} \in \mathcal{H}^{(k)}$. Sampling large N hypotheses from each σ_k provides a high probability that there exists $h_n^{(k)}$ such that $h_n^{(k)}(\boldsymbol{x}) = h(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{U}$, so that Property 1 is stochastically satisfied. Also, by setting $S \gg \log C$, Eq. 2 is satisfied. Consequently, the empirical LPDR of $\boldsymbol{x}$ for $\hat{h}$ is obtained by Eq. 1.

Figure 1b depicts an example of $\tilde{L}_{\hat{h}}(\boldsymbol{x}_i)$ for $\boldsymbol{x}_i \in \mathcal{U}$ on MNIST dataset. The $\boldsymbol{x}_i$ is the i^{th} sample ordered by empirical LPDR. The set of x-axis values of the blue dots on the horizontal line, whose y-axis value is i, are $\{\rho_S(h_c, \hat{h}) : h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x}_i)\}$. Thus, $\tilde{L}_{\hat{h}}(\boldsymbol{x}_i)$ is the x-axis value of the leftmost blue dot for $\boldsymbol{x}_i$ (indicated by the blue arrow).



(a) (b)

Figure 2: The need for regulating σ (MNIST). (a) The disagree ratios of the samples goes to 0 as σ decreases and goes to 0.9 as σ increases. (b) The $\mathbb{E}[\rho_S(h, \hat{h})]$ for $h \in \mathcal{H}_B$ decreases as the number of labeled samples increases when $\mathcal{H}_B$ is constructed with a static σ for all acquisition steps.

3.4 Rank of LPDR

Samples with small LPDR are more sensitive to perturbation of the decision boundary, thus the label of the samples are more likely to be altered by the hypothesis perturbation. That is the disagreement in sampled hypotheses with $\hat{h}$ may have an inverse relation with empirical LPDR. This paper defines a measure of the disagreement in sampled hypotheses with $\hat{h}$ on $\boldsymbol{x}$, referred to as *disagree ratio*:

$$D(\hat{h}, \boldsymbol{x}) := \frac{|\mathcal{H}_B(\hat{h}, \boldsymbol{x})|}{|\mathcal{H}_B|} = \frac{1}{N} \sum_{n=1}^{N} \mathbb{I}[h_n(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})]$$

where $\mathcal{H}_B(\hat{h}, \boldsymbol{x}) = \{h_n \in \mathcal{H}_B : h_n(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})\}, \mathcal{H}_B = \{h_n\}_{n=1}^N$, and h_n is sampled with $\boldsymbol{w}_n \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$. Then, we formulate the following conjecture:

Conjecture 2 *Suppose that $\mathcal{H}_B$ is constructed with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ and $0 < \sigma < \infty$. Then,*

$$D(\hat{h}, \boldsymbol{x}_1) > D(\hat{h}, \boldsymbol{x}_2) \iff L_{\hat{h}}(\boldsymbol{x}_1) < L_{\hat{h}}(\boldsymbol{x}_2) \tag{3}$$

with probability tending to 1 as $|\mathcal{H}_B| \to \infty$ where $D(\hat{h}, \boldsymbol{x}) = |\mathcal{H}_B(\hat{h}, \boldsymbol{x})|/|\mathcal{H}_B|$.

The theoretical and empirical verifications of Conjecture 2 are presented in Appendix C.2. The Eq. 3 of Conjecture 2 implies that we can use $D(\hat{h}, \boldsymbol{x})$ to identify the order of $L_{\hat{h}}(\boldsymbol{x})$, required to query closest samples to hypothesis. Our motivation is to find a measure to identify the order of LPDR without evaluating the empirical LPDR. We would like the measure evaluation to be computationally lighter than that of the empirical LPDR.

Obtaining the order of LPDR by the disagree ratio can reduce time over the brute-force search for empirical LPDR in Section 3.3. Let M, K, N, and S be the number of unlabeled samples, grid for σ_k, sampled hypotheses for each σ_k, and i.i.d. random variables from $\mathcal{D}$ for approximating ρ, respectively. The brute-force search requires time complexity of $\mathcal{O}(M \times K \times N \times S)$. While, using the disagree ratio requires the time complexity of $\mathcal{O}(M \times N)$. In the results of the active learning performance comparison between evaluating empirical LPDR by brute-force search and evaluating the order of empirical LPDR by using disagree ratio on various datasets, there is no significant difference in the performance between the two methods (See Appendix D).

4 LPDR-based Active Learning

This section introduces the proposed LPDR-based active learning algorithm with the disagree ratio referred to as 'DRAL' and its variation with the weighted disagree ratio referred to as 'DRAL$^+$'. This paper considers a pool-based active learning that queries q most informative samples from randomly sampled pool data $\mathcal{P} \subset \mathcal{U}$ of size m where $\mathcal{U}$ is unlabeled samples.

4.1 Variance for Disagree Ratio

When constructing $\mathcal{H}_B$, setting σ is an important issue as shown in the following theorem:

Algorithm 1 DRAL

Input:
$\mathcal{L}_0, \mathcal{U}_0$: Initial labeled and unlabeled samples
σ^2 : Initial variance for sampling
ρ^* : Target disagree metric ($= q/m$)

Procedure:
for step $t = 0$ **to** $T-1$ **do**
  Obtain $\hat{\boldsymbol{w}}_t$ by training with $\mathcal{L}_t$
  for $n = 1$ **to** N **do**
    $\boldsymbol{w}_n \sim \mathcal{N}(\hat{\boldsymbol{w}}_t, \mathbf{I}\sigma^2)$
    $\rho' = \rho_S(h_n, \hat{h}_t)$
    $\sigma \leftarrow \sigma e^{-\beta(\rho' - \rho^*)}$ where $\beta > 0$
  end for
  $D(\hat{h}_t, \boldsymbol{x}_i)$ for $i \in \mathcal{I}_{\mathcal{P}_t} = \{j : \boldsymbol{x}_j \in \mathcal{P}_t \subset \mathcal{U}_t\}$
  $\mathcal{I}^* = \arg\max_{\mathcal{I} \subset \mathcal{I}_{\mathcal{P}_t}, |\mathcal{I}|=q} \sum_{i \in \mathcal{I}} D(\hat{h}_t, \boldsymbol{x}_i)$
  $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{I}^*}, \mathcal{U}_{t+1} = \mathcal{U}_t \backslash \{\boldsymbol{x}_i\}_{i \in \mathcal{I}^*}$
end for

Algorithm 2 DRAL$^+$

Input:
$\mathcal{L}_0, \mathcal{U}_0, \sigma^2, \rho^*$

Procedure:
for step $t = 0$ **to** $T-1$ **do**
  Obtain $\hat{\boldsymbol{w}}_t$ by training with $\mathcal{L}_t$
  Evaluate empirical error $\hat{\varepsilon}_t$ of $\hat{h}_t$
  for $n = 1$ **to** N **do**
    $\boldsymbol{w}_n \sim \mathcal{N}(\hat{\boldsymbol{w}}_t, \mathbf{I}\sigma^2)$
    Evaluate empirical error ε_n of h_n
    $\gamma_n = e^{-(\varepsilon_n - \hat{\varepsilon}_t)}$
    $\rho' = \rho_S(h_n, \hat{h}_t)$, then update σ
  end for
  $D_w(\hat{h}_t, \boldsymbol{x}_i)$ for $i \in \mathcal{I}_{\mathcal{P}_t}$
  $\mathcal{I}^* = \arg\max_{\mathcal{I} \subset \mathcal{I}_{\mathcal{P}_t}, |\mathcal{I}|=q} \sum_{i \in \mathcal{I}} D_w(\hat{h}_t, \boldsymbol{x}_i)$
  Update $\mathcal{L}_{t+1}$ and $\mathcal{U}_{t+1}$
end for

Theorem 2 *Consider the binary classification with the linear classifier on bounded $\mathcal{X}$, i.e., $\mathcal{H} = \{h : h(\boldsymbol{x}) = \mathrm{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w})\}$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|_\infty < \infty$. Suppose that $\mathcal{H}_B = \{h_n\}_{n=1}^N$ is constructed with $\boldsymbol{w}_n \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ for given $\hat{h} \in \mathcal{H}$. Then, for all $\boldsymbol{x} \in \mathcal{X}$, 1) $D(\hat{h}, \boldsymbol{x}) \to 0$ when $\sigma \to 0$ and 2) $D(\hat{h}, \boldsymbol{x}) \to 1/2$ when $\sigma \to \infty$ in probability as $N \to \infty$, where $D(\hat{h}, \boldsymbol{x}) = |\mathcal{H}_B(\hat{h}, \boldsymbol{x})|/|\mathcal{H}_B|$.*

The proof of Theorem 2 is deferred to Appendix A.2. The implication of Theorem 2 is that when σ is too small or too large, it would be difficult to obtain the order of LPDR by comparing the disagree ratios. In practice, N is finite, thus setting σ for $\mathcal{H}_B$ is more important. Figure 2a shows that the disagree ratios of samples converge to 0 as $\sigma \to 0$ and converge to 0.9 as σ increases when $|\mathcal{H}_B| = 100$. Thus, it is required to set an appropriate σ, but finding an appropriate σ is computationally prohibitive. Additionally, the behavior of $D(\hat{h}, \boldsymbol{x})$ with the fixed σ is not same with respect to the number of labeled samples. Figure 2b shows that $\mathbb{E}[\rho_S(h_n, \hat{h})]$ for $h_n \in \mathcal{H}_B$ decreases as the number of labeled samples increases on MNIST dataset when $\mathcal{H}_B$ is constructed with a static $\sigma = 0.1$ for all acquisition steps in active learning. Decrease in $\mathbb{E}[\rho_S(h_n, \hat{h})]$ leads to decrease in $|\mathcal{H}_B(\hat{h}, \boldsymbol{x})|$ for each $\boldsymbol{x}$, eventually $D(\hat{h}, \boldsymbol{x})$ converges to zero.

To address these problems, σ needs to be regulated at each acquisition step to keep $\mathbb{E}[\rho_S(h_n, \hat{h})]$ static. In Figure 6c of Appendix C.1, $\mathbb{E}[\rho_S(h_n, \hat{h})]$ is almost a linear function of $\log \sigma$ in the ascension, that is, $\sigma \propto e^{\beta \mathbb{E}[\rho_S(h_n, \hat{h})]}$ for some $\beta > 0$. Based on this observation, we can keep $\mathbb{E}[\rho_S(h_n, \hat{h})]$ static by updating σ as follows:

$$\sigma \leftarrow \sigma e^{-\beta(\rho_S - \rho^*)}$$

where ρ^* is the target disagree metric (see Appendix E: $\mathbb{E}[\rho_S(h_n, \hat{h})]$ is securely guided towards the target value). The remaining question is how to determine an appropriate ρ^*. To identify q most informative samples from unlabeled sample set of size m, we set $\rho^* = q/m$ and achieved high performance (see Appendix F: the proposed algorithm shows high performance when $\rho^* = q/m$).

4.2 Algorithm with Disagree Ratio (DRAL)

The proposed LPDR-based active learning algorithm with the disagree ratio referred to as 'DRAL' is provided in Algorithm 1. Let $\mathcal{L}_t$, $\mathcal{U}_t$ and $\mathcal{P}_t \subset \mathcal{U}_t$ be labeled samples, unlabeled samples, and pool data of size m at step t respectively. At step t, $\hat{\boldsymbol{w}}_t$ is obtained by training with $\mathcal{L}_t$, then h_n is sampled with $\boldsymbol{w}_n \sim \mathcal{N}(\hat{\boldsymbol{w}}_t, \mathbf{I}\sigma^2)$ for $n = 1, \ldots, N$. Here, σ is updated so that the $\mathbb{E}[\rho_S(h_n, \hat{h}_t)]$ achieves the target value ρ^*. Then, DRAL queries the top q unlabeled samples having the highest disagree ratio from $\mathcal{P}_t$.

4.3 Algorithm with Weighted Disagree Ratio (DRAL$^+$)

When calculating the disagree ratio, each sampled hypothesis is given equal weight. However, we observe performance variation among the sampled hypotheses (See Appendix G). For this reason,

Table 1: Settings for data and acquisition size. Acquisition size denotes the number of initial labeled samples + query size for each step (the size of pool data) $\rightarrow$ the number of final labeled samples.

Dataset	Model	# of parameters sampled / total	Data size train / validation / test	Acquisition size		
MNIST	S-CNN	1.3K/1.2M	55,000 / 5,000 / 10,000	20	+20 (2,000)	$\rightarrow$ 1,020
CIFAR10	K-CNN	5.1K/2.2M	45,000 / 5,000 / 10,000	200	+400 (4,000)	$\rightarrow$ 9,800
SVHN	K-CNN	5.1K/2.2M	68,257 / 5,000 / 26,032	200	+100 (2,000)	$\rightarrow$ 10,200
CIFAR100	WRN-16-8	51.3K/11.0M	45,000 / 5,000 / 10,000	5,000	+2,000 (10,000)	$\rightarrow$ 25,000
Tiny ImageNet	WRN-16-8	409.8K/11.4M	90,000 / 10,000 / 10,000	10,000	+5,000 (20,000)	$\rightarrow$ 50,000
HAM10000	WRN-16-8	14.3K/11.0M	7,015 / 1,500 / 1,500	500	+300 (3,000)	$\rightarrow$ 3,500

this paper introduces weighting factor γ_n on h_n such that more/less weight is placed on h_n that performs better/worse than $\hat{h}$ on labeled samples. The weighting factor is given below as

$$\gamma_n = e^{-(\varepsilon_n - \hat{\varepsilon}_t)},$$

where ε_n and $\hat{\varepsilon}_t$ are the empirical errors of h_n and $\hat{h}_t$ respectively. Then, the following weighted disagree ratio can be defined as shown below as

$$D_w(\hat{h}_t, \boldsymbol{x}) := \frac{\sum_{n=1}^{N} \gamma_n \mathbb{I}[h_n(\boldsymbol{x}) \neq \hat{h}_t(\boldsymbol{x})]}{\sum_{n=1}^{N} \gamma_n}.$$

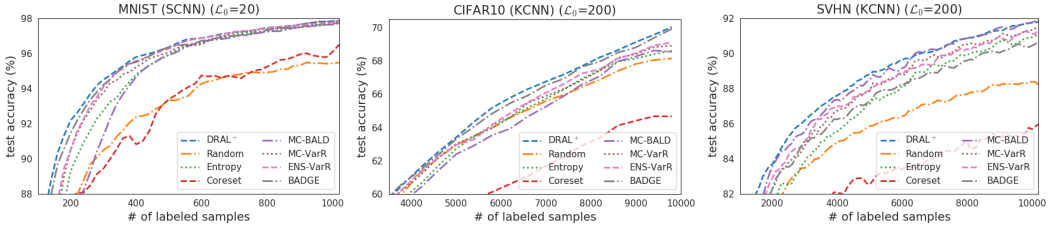
The details of the algorithm and the framework with the weighted disagree ratio referred to as 'DRAL$^+$' are provided in Algorithm 2 and Appendix M. In the results of the performance comparison between DRAL$^+$ and DRAL on various datasets, DRAL$^+$ consistently either performs better than or comparable with DRAL (See Appendix H).

5 Experiments

This section discusses experimental results for performance comparison with the baseline active learning algorithms on benchmark datasets in deep learning. A total of 6 benchmark datasets are used for experiments: MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2019), CIFAR100 (Krizhevsky et al., 2009), Tiny ImageNet (subset of the ILSVRC dataset containing 200 categories; Russakovsky et al., 2015), and HAM10000 (Tschandl et al., 2018) datasets. Three CNN networks are used as deep networks: S-CNN, K-CNN (Chollet et al., 2015) and Wide-ResNet (WRN-16-8; Zagoruyko & Komodakis, 2016). Results are averaged over 5 repetitions. The settings for data, initial, and query sizes are summarized in Table 1. We use 100 forward passes for MC-dropout sampling, and ensemble consists of 5 networks of identical architecture but different random initialization and random batches. There is no significant performance difference between when parameter sampling is performed in the entire layer and in the last layer, thus parameters are sampled in the last layer for computational efficiency. We set hyperparameters for DRAL$^+$ as $\sigma_0 = 0.01, \beta = 1$, and $N = 100$ in convenience as DRAL$^+$ is robust against hyperparameters (see Appendix I). The details of datasets, networks, and training settings are presented in Appendix B. Figure 3–5 show plots of test accuracy enlarged in appropriate to accentuate the performance difference among different methods: initial labeled sample sizes are not shown in the figures. Figures that include initial labeled sample size are presented in Appendix K.

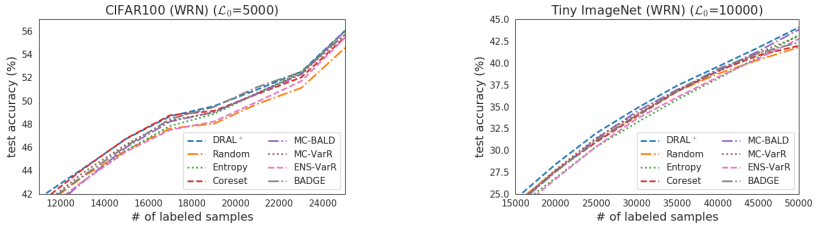
5.1 Results for MNIST, CIFAR10 and SVHN

A number of experiments are conducted to compare performance of DRAL$^+$ with the baseline active learning algorithms including state-of-the-art algorithms. Figure 3 shows the test accuracy with respect to the number of labeled samples on MNIST, CIFAR10 and SVHN datasets. Each algorithm is denoted as follows 'Entropy': entropy-based uncertainty sampling (Shannon, 1948), 'Coreset': core-set selection (Sener & Savarese, 2017), 'MC-BALD': MC-dropout sampling with BALD (Gal et al., 2017), 'MC-VarR': MC-dropout sampling with variation ratio (Ducoffe & Precioso, 2015), 'ENS-VarR': ensemble method with variation ratio (Beluch et al., 2018), and 'BADGE': batch active learning by diverse gradient embeddings (Ash et al., 2020). Overall, DRAL$^+$ either consistently performs best or comparable with other algorithms on both datasets. Entropy shows the poor performance compared to other uncertainty-based algorithms in all results. Coreset shows the worst



(a) (b) (c)

Figure 3: The performance comparison of DRAL$^+$ with the baseline active learning algorithms on MNIST (a), CIFAR10 (b) and SVHN (c) datasets. Overall, DRAL$^+$ consistently either performs best or comparable with all other algorithms regardless of dataset.



(a) (b)

Figure 4: The performance comparison of DRAL$^+$ with the baseline active learning algorithms on on CIFAR100 (a) and Tiny ImageNet (b) datasets with WRN-16-8 networks. These datasets are more difficult. DRAL$^+$ outperforms all other algorithms on more difficult tasks.

performance compared to all other algorithms including Random. MC-BALD performs comparable with DRAL$^+$ on SVHN, but it shows the poor performance on MNIST and CIFAR10. MC-VarR and ENS-VarR show a significant performance drop compared to DRAL$^+$ on all datasets. BADGE performs comparable with DRAL$^+$ on MNIST and CIFAR10 datasets, but is shows a significant performance drop compared to DRAL$^+$ on SVHN dataset. It is observed that the performance of other algorithms has a relatively strong dataset or network dependency compared to DRAL$^+$.

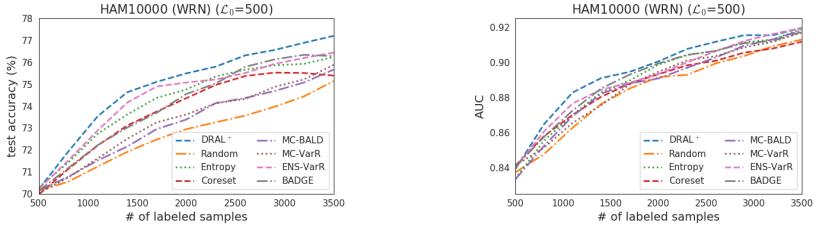
Furthermore, the running time of DRAL$^+$ for active learning is comparable to Entropy, MC-BALD, MC-VarR and Coreset. ENS-VarR requires about 5 times more computational load than DRAL$^+$, and BADGE requires several times more computational load than DRAL$^+$ when the parameter dimension and query size are very large such as Tiny ImageNet with WRN-16-8. The details of the running time is presented in Table 4 of Appendix J.

5.2 Results for CIFAR100 and Tiny ImageNet

Experiments on more difficult task are conducted. Figure 4 shows test accuracy with respect to the number of labeled samples on Tiny ImageNet dataset with WRN-16-8. CIFAR100 and Tiny ImageNet are considered to be more difficult task than other benchmark datasets. Even on more difficult tasks, DRAL$^+$ outperforms all other algorithms.

5.3 Results for HAM10000

Additional experiments are conducted to compare the performance of the algorithms on imbalanced HAM10000 dataset with WRN-16-8. Figure 5a shows the results of the test accuracy with respect to the number of labeled samples. DRAL$^+$ outperforms all other algorithms compared. Figure 5b shows the results of AUC with respect to the number of labeled samples. DRAL$^+$ performs better than or comparable with all other algorithms.



(a) (b)

Figure 5: The performance comparison on imbalanced HAM10000 dataset with WRN-16-8 in terms of the test accuracy (a) and AUC (b). DRAL$^+$ performs better than or comparable with all other algorithms.

Table 2: The mean $\pm$ standard deviation of the performance differences (%) relative to DRAL$^+$ for each algorithm and each dataset. The negative value indicates lower performance compared to DRAL$^+$, and the asterisk (*) indicates that the p-value ls less than 0.05 in one-sample t-test for the performance differences. DRAL$^+$ consistently and significantly outperforms other algorithms for all datasets, while the performance of the algorithms except DRAL$^+$ vary depending on datasets.

	MNIST	CIFAR10	SVHN	CIFAR100	T. ImageNet	HAM10000
DRAL$^+$[ours]	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
Random	-3.26±0.44*	-1.05±0.26*	-2.94±0.15*	-1.16±0.49*	-0.99±0.45*	-2.24±0.48*
Entropy[41]	-1.01±0.26*	-0.97±0.35*	-1.75±0.13*	-0.69±0.32*	-1.40±0.36*	-0.69±0.42*
Coreset[36]	-3.34±0.78*	-4.77±0.19*	-5.93±0.35*	-0.28±0.33	-0.90±0.32*	-1.13±0.33*
MC-BALD[14]	-1.70±0.48*	-1.25±0.35*	-0.29±0.04*	-0.67±0.22*	-0.58±0.39*	-1.75±0.37*
MC-VarR[10]	-0.67±0.31*	-0.75±0.46*	-0.90±0.10*	-0.42±0.24*	-0.64±0.49*	-1.63±0.38*
ENS-VarR[3]	-0.47±0.36*	-0.65±0.48*	-0.82±0.07*	-1.07±0.21*	-1.43±0.30*	-0.52±0.26*
BADGE[1]	-0.26±0.10*	-0.22±0.44	-1.46±0.05*	-0.31±0.77	-0.79±0.51*	-0.86±0.29*

5.4 Summary

Each cell of Table 2 presents the mean and standard deviation of five $\bar{\Delta}$s for each algorithm and dataset, where $\bar{\Delta}$ is the average of performance differences relative to DRAL$^+$ over all steps for each repetition. The negative value indicates lower performance compared to DRAL$^+$, and the asterisk (*) indicates the p-value is less than 0.05 in one-sample t-test for the null of no difference versus the alternative that the DRAL$^+$ is better. DRAL$^+$ consistently outperforms other algorithms on all datasets, while the performance of the algorithms except DRAL$^+$ vary depending on datasets. Furthermore, DRAL$^+$ shows significant performance improvement in most cases (39 out of 42).

6 Conclusion

This paper defines a measure of sample's closeness to the decision boundary of the current network referred to as the least probable disagreement region (LPDR) based on the disagree metric between hypotheses. In addition, this paper introduces a hypothesis sampling method with a measure of disagreement in sampled hypotheses referred to as the disagree ratio for obtaining the order of LPDR without explicit or empirical evaluation of LPDR for computational efficiency. Based on the order of LPDR, this paper proposes an uncertainty-based active learning algorithm of querying unlabeled samples closest to the current decision boundary in terms of LPDR.

The proposed LPDR-based active learning algorithm consistently outperforms all high performing active learning algorithms and leads to state-of-the-art active learning performance on all datasets in this paper. In addition, the proposed algorithm is simple enough to perform only parameter perturbation and can be applied to a variety of classification tasks with both shallow and deep networks. Furthermore, the proposed algorithm runs fast enough to comparable to entropy-based uncertainty sampling for it requires a low computational load. In conclusion, LPDR-based sampling with the disagree ratio by parameter perturbation is an effective active learning algorithm based on the uncertainty of the current network.

REFERENCES

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In International Conference on Learning Representations, 2020. URL https://openreview.net/forum?id=ryghZJBKPS.

Maria-Florina Balcan, Andrei Broder, and Tong Zhang. Margin based active learning. In International Conference on Computational Learning Theory, pp. 35–50. Springer, 2007.

William H Beluch, Tim Genewein, Andreas Nürnberger, and Jan M Köhler. 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.

Djallel Bouneffouf, Romain Laroche, Tanguy Urvoy, Raphael Féraud, and Robin Allesiardo. Contextual bandit for active learning: Active thompson sampling. In International Conference on Neural Information Processing, pp. 405–412. Springer, 2014.

George EP Box. A note on the generation of random normal deviates. Ann. Math. Statist., 29: 610–611, 1958.

Shayok Chakraborty, Vineeth Balasubramanian, and Sethuraman Panchanathan. Adaptive batch mode active learning. IEEE transactions on neural networks and learning systems, 26(8):1747–1760, 2014.

François Chollet et al. Keras. https://keras.io, 2015.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. Journal of artificial intelligence research, 4:129–145, 1996.

Aron Culotta and Andrew McCallum. Reducing labeling effort for structured prediction tasks. In AAAI, volume 5, pp. 746–751, 2005.

Melanie Ducoffe and Frederic Precioso. Qbdc: query by dropout committee for training deep supervised architecture. arXiv preprint arXiv:1511.06412, 2015.

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

Linton C Freeman. Elementary applied statistics: for students in behavioral science. John Wiley & Sons, 1965.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In European Conference on Computer Vision, pp. 562–577. Springer, 2014.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1183–1192. JMLR. org, 2017.

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

Bin Gu, Zhou Zhai, Cheng Deng, and Heng Huang. Efficient active learning by querying discriminative and representative samples and fully exploiting unlabeled data. IEEE Transactions on Neural Networks and Learning Systems, 2020.

Denis Gudovskiy, Alec Hodgkinson, Takuya Yamaguchi, and Sotaro Tsukizawa. Deep active learning for biased datasets via fisher kernel self-supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9041–9049, 2020.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision, pp. 1026–1034, 2015.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.

Daniel Joseph Hsu. Algorithms for active learning. PhD thesis, UC San Diego, 2010.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2372–2379. IEEE, 2009.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep Bayesian active learning. In Advances in Neural Information Processing Systems, pp. 7024–7035, 2019.

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

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. https://www.cs.toronto.edu/~kriz/cifar.html, 2009.

Yann LeCun, Corinna Cortes, and Christopher JC Burges. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist, 1998.

David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In SIGIR'94, pp. 3–12. Springer, 1994.

David Mickisch, Felix Assion, Florens Greßner, Wiebke Günther, and Mariele Motta. Understanding the decision boundary of deep neural networks: An empirical study. arXiv preprint arXiv:2002.01810, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and A Ng. The street view house numbers (svhn) dataset, 2019.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. How to measure uncertainty in uncertainty sampling for active learning. Machine Learning, pp. 1–34, 2021.

Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In Advances in Neural Information Processing Systems, pp. 6359–6370, 2019.

Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. ICML, Williamstown, pp. 441–448, 2001.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In International Symposium on Intelligent Data Analysis, pp. 309–318. Springer, 2001.

Andrew I Schein and Lyle H Ungar. Active learning for logistic regression: an evaluation. Machine Learning, 68(3):235–265, 2007.

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

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. In Advances in neural information processing systems, pp. 1289–1296, 2008.

H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In Proceedings of the fifth annual workshop on Computational learning theory, pp. 287–294, 1992.

Claude E Shannon. A mathematical theory of communication. Bell system technical journal, 27(3): 379–423, 1948.

Manali Sharma and Mustafa Bilgic. Evidence-based uncertainty sampling for active learning. Data Mining and Knowledge Discovery, 31(1):164–202, 2017.

Weishi Shi and Qi Yu. Integrating bayesian and discriminative sparse kernel machines for multi-class active learning. In Advances in Neural Information Processing Systems, pp. 2285–2294, 2019.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In Proceedings of the IEEE International Conference on Computer Vision, pp. 5972–5981, 2019.

Charles Spearman. "General Intelligence" objectively determined and measured. American Journal of Psychology, 15:201–293, 1904.

Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In Proceedings of the ninth ACM international conference on Multimedia, pp. 107–118, 2001.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data, 5: 180161, 2018.

Shuo Wang, Jian-Jian Wang, Xiang-Hui Gao, and Xue-Zheng Wang. Pool-based active learning based on incremental decision tree. In 2010 International Conference on Machine Learning and Cybernetics, volume 1, pp. 274–278. IEEE, 2010.

Yazhou Yang, Xiaoqing Yin, Yang Zhao, Jun Lei, Weili Li, and Zhe Shu. Batch mode active learning based on multi-set clustering. IEEE Access, 9:51452–51463, 2021.

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, 113(2):113–127, 2015.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 93–102, 2019.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8756–8765, 2020.

Xiao-Yu Zhang, Shupeng Wang, and Xiaochun Yun. Bidirectional active learning: A two-way exploration into unlabeled and labeled data set. IEEE transactions on neural networks and learning systems, 26(12):3034–3044, 2015.

A PROOFS OF THEOREMS

A.1 PROOF OF THEOREM 1

Since $M < \infty$ and $L_{\hat{h}}(\boldsymbol{x}_i) \neq L_{\hat{h}}(\boldsymbol{x}_j)$ for $i \neq j$, there exists $\delta > 0$ such that

$$\delta = \min_{i \neq j} |L_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j)|,$$

and our strategy is to show that $\sup_i |\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_i)| < \delta/2$ in probability.

At first, we prove the convergence of empirical LPDR to LPDR. Since $|\mathbb{I}[h(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})]| \leq 1$ for any $h \in \mathcal{H}$, Hoeffding's inequality implies that for any $\epsilon > 0$,

$$\mathbb{P}\left[|\rho_S(h, \hat{h}) - \rho(h, \hat{h})| \geq \epsilon\right] \leq 2e^{-c_1 \epsilon^2 S}$$

with $c_1 > 0$, and

$$\mathbb{P}\left[\sup_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} |\rho_S(h_c, \hat{h}) - \rho(h_c, \hat{h})| \geq \epsilon\right] \leq 2Ce^{-c_1 \epsilon^2 S} \tag{4}$$

because $|\mathcal{H}_C(\hat{h}, \boldsymbol{x})| \leq C$. Furthermore, Property 1 implies that for any $\epsilon > 0$, the following holds: For all $h \in \mathcal{H}(\hat{h}, \boldsymbol{x})$,

$$|\rho(h_c, \hat{h}) - \rho(h, \hat{h})| < \epsilon \tag{5}$$

as $C \to \infty$ because of Property 1. Additionally, by Eq. 4 and 5, for any $\epsilon > 0$, we have that

$$\bigcap_{h \in \mathcal{H}(\hat{h}, \boldsymbol{x})} \left\{\exists h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x}) \quad \text{s.t.} \quad |\rho_S(h_c, \hat{h}) - \rho(h, \hat{h})| < 2\epsilon\right\} \tag{6}$$

with probability equal to or greater than $1 - 2Ce^{-c_1 \epsilon^2 S}$ when C is sufficiently large. It is because the $|\rho_S(h_c, \hat{h}) - \rho(h, \hat{h})| \leq |\rho_S(h_c, \hat{h}) - \rho(h_c, \hat{h})| + |\rho(h_c, \hat{h}) - \rho(h, \hat{h})|$ and Eq. 6 implies that $\sup_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} |\rho_S(h, \hat{h}) - \rho(h, \hat{h})| < \epsilon$ with probability equal to or greater than $1 - 2Ce^{-c_1 \epsilon^2 S}$.

Then,

$$\inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho_S(h_c, \hat{h}) - 2\epsilon \leq \inf_{h \in \mathcal{H}(\hat{h}, \boldsymbol{x})} \rho(h, \hat{h}) \leq \inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho_S(h_c, \hat{h}) + 2\epsilon \tag{7}$$

with probability equal to or greater than $1 - 2Ce^{-c_1 \epsilon^2 S}$. We'll prove the lower bound and upper bound of Eq. 7 separately.

Let the set $\mathcal{H}_C^*(\hat{h}, \boldsymbol{x}) \subset \mathcal{H}_C(\hat{h}, \boldsymbol{x})$ be a smallest subset satisfying the Property 1, i.e. all elements of $\mathcal{H}_C^*(\hat{h}, \boldsymbol{x})$ are used to approximate $\rho(h, \hat{h})$ for all $h \in \mathcal{H}(\hat{h}, \boldsymbol{x})$. Then, Eq. 6 implies that

$$\inf_{h_c \in \mathcal{H}_C^*(\hat{h}, \boldsymbol{x})} \rho_S(\hat{h}, \boldsymbol{x}) - 2\epsilon \leq \inf_{h \in \mathcal{H}(\hat{h}, \boldsymbol{x})} \rho(\hat{h}, \boldsymbol{x})$$

with probability equal to or greater than $1 - 2Ce^{-c_1 \epsilon^2 S}$ when S is sufficiently large. In addition, the property of infimum implies that $\inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho_S(\hat{h}, \boldsymbol{x}) - 2\epsilon \leq \inf_{h_c \in \mathcal{H}_C^*(\hat{h}, \boldsymbol{x})} \rho_S(\hat{h}, \boldsymbol{x}) - 2\epsilon$, which proves the lower bound of Eq. 7. Also, Eq 4 implies that $\bigcap_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \{|\rho_S(h_c, \hat{h}) - \rho(h_c, \hat{h})| < \epsilon\}$ with probability equal to or greater than $1 - 2Ce^{-c_1 \epsilon^2 S}$ when S is sufficiently large. Thus we have

$$\inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho(\hat{h}, \boldsymbol{x}) \leq \inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho_S(\hat{h}, \boldsymbol{x}) + 2\epsilon,$$

and by the property of infimum, $\inf_{h \in \mathcal{H}(\hat{h}, \boldsymbol{x})} \rho(\hat{h}, \boldsymbol{x}) \leq \inf_{h_c \in \mathcal{H}_C(\hat{h}, \boldsymbol{x})} \rho(\hat{h}, \boldsymbol{x})$, which proves the upper bound of Eq. 7.

Consequently, by the definition of LPDR and empirical LPDR,

$$\mathbb{P}\left[|\tilde{L}_{\hat{h}}(\boldsymbol{x}) - L_{\hat{h}}(\boldsymbol{x})| < 2\epsilon\right] \geq 1 - 2Ce^{-c_1 \epsilon^2 S}, \tag{8}$$

which goes to 1 as $\min(S, C) \to \infty$ and $\log C/S \to 0$ by Eq. 2. This implies the convergence of empirical LPDR to LPDR. Next, we prove the rank-order consistency between the empirical LPDR and LPDR of $\{\boldsymbol{x}_i\}_{i=1}^M$. It is trivial that

$$\mathbb{P}\left[\max_{i\in[M]} |\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_i)| \geq 2\epsilon\right]$$

is equal to or less than

$$\sum_{i=1}^{M} \mathbb{P}\left[|\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_i)| \geq 2\epsilon\right] = \sum_{i=1}^{M} \left\{1 - \mathbb{P}\left[|\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_i)| < 2\epsilon\right]\right\},$$

and it has the upper bound of $2MCe^{-c_1\epsilon^2 S}$ by the Eq. 8. Consequently,

$$\max_{i\in[M]} |\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_i)| < 2\epsilon \tag{9}$$

with probability tending to 1 as $\min(S, C) \to \infty$ by Eq. 2, and it holds for any ϵ such that $0 < 2\epsilon < \delta/2$. Then the uniform convergence of empirical LPDR on $\{\boldsymbol{x}_i\}_{i=1}^M$, denoted by the Eq. 9., implies that the maximum difference between all pairs of $L_{\hat{h}}(\boldsymbol{x}_i)$ and $\tilde{L}_{\hat{h}}(\boldsymbol{x}_i)$ is less than the minimum of pair-wise differences of $L_{\hat{h}}(\boldsymbol{x}_i)$s with probability tending to 1 as $\min(S, C) \to \infty$. Therefore, it implies the rank-order consistency between the empirical LPDR and LPDR:

$$\bigcap_{i\neq j} \left\{\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - \tilde{L}_{\hat{h}}(\boldsymbol{x}_j) > 0 \Longrightarrow L_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j) > 0\right\}$$

with probability tending to 1 as as $\min(S, C) \to \infty$ because the contra-positive such that if $L_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j) \leq 0$, then

$$L_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j) < -\delta < 0$$

and

$$\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - \tilde{L}_{\hat{h}}(\boldsymbol{x}_j) < L_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j) + 4\epsilon < -\delta + 4\epsilon \leq 0$$

holds uniformly on $i \neq j$ with probability tending to 1 as $\min(S, C) \to \infty$ by the Eq. 9 implying that

$$\max_{i\neq j} |\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j)| < 4\epsilon = \bigcap_{i\neq j} \left\{|\tilde{L}_{\hat{h}}(\boldsymbol{x}_i) - L_{\hat{h}}(\boldsymbol{x}_j)| < 4\epsilon\right\}$$

with probability tending to 1 as $\min(S, C) \to \infty$.

A.2 Proof of Theorem 2

The disagree ratio is

$$D(\hat{h}, \boldsymbol{x}) = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}\left[h_n(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})\right],$$

and $h_n(\boldsymbol{x})$ disagrees with $\hat{h}(\boldsymbol{x})$ if $\operatorname{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}_n) \neq \operatorname{sgn}(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}})$, here, $\operatorname{sgn}(0) = 1$. Let $\|\hat{\boldsymbol{w}}\| = 1$ without the loss of generality, $\|\boldsymbol{x}\| \neq 0$ to avoid the null, and note that

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{w}_n = \boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}} + \sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n$$

where $\mathbf{e}_n = (Z_{n1}, \ldots, Z_{n|\boldsymbol{w}|})^\mathsf{T}$ and $Z_{nk} \sim \mathcal{N}(0, 1^2)$. Then, when $N \to \infty$, $\forall \boldsymbol{x}$,

$$D(\hat{h}, \boldsymbol{x}) \to \begin{cases} \mathbb{P}\left[\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n \geq -\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}\right], & \boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}} < 0 \\ \mathbb{P}\left[\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n < -\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}\right], & \boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}} \geq 0 \end{cases}$$

in probability.

In the first fold of $\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}} < 0$,

$$\mathbb{P}\left[\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n \geq -\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}\right] = \mathbb{P}\left[\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n \geq |\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}|\right] = \mathbb{P}\left[\sigma\|\boldsymbol{x}\|Z \geq |\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}|\right] = 1 - \Phi\left(\frac{a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}\right)}{\sigma}\right)$$

where $Z \sim \mathcal{N}(0, 1^2)$, Φ is the cumulative distribution function of the normal distribution, and $a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}\right) = |\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}|/\|\boldsymbol{x}\|$. Note that $\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}_n \sim \mathcal{N}(0, \sigma^2\|\boldsymbol{x}\|^2)$.

Table 3: Settings for training.

Dataset	Model	Epochs	Batch size	Optimizer	Learning Rate	Learning Rate Schedule ×decay [epoch schedule]
MNIST	S-CNN	50	32	Adam	0.001	-
CIFAR10	K-CNN	150	64	Adam	0.0001	-
SVHN	K-CNN	150	64	Adam	0.0001	-
CIFAR100	WRN-16-8	100	128	Nesterov	0.05	×0.2 [60, 80]
Tiny ImageNet	WRN-16-8	200	128	Nesterov	0.1	×0.2 [60, 120, 160]
HAM10000	WRN-16-8	100	64	Nesterov	0.05	×0.2 [60, 80]

Next, in the second fold of $\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{w}} \geq 0$,

$$\mathbb{P}\left[\sigma \boldsymbol{x}^{\mathsf{T}}\mathbf{e}_n < -\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{w}}\right] = \mathbb{P}\left[\sigma\|\boldsymbol{x}\|Z > |\boldsymbol{x}^{\mathsf{T}}\hat{\boldsymbol{w}}|\right] = 1 - \Phi\left(\frac{a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}\right)}{\sigma}\right).$$

Here, $\Phi(\infty) = 1$ and $\Phi(0) = 1/2$ by the smoothness of Φ. Consequently, in both folds,

$$D(\hat{h}, \boldsymbol{x}) \to 1 - \Phi\left(\frac{a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}\right)}{\sigma}\right) = \begin{cases} 0, & \sigma \to 0 \\ 1/2, & \sigma \to \infty \end{cases}$$

in probability as $N \to \infty$.

B Datasets, Networks and Experimental Settings

B.1 Benchmark Datasets

MNIST (LeCun et al., 1998) is a handwritten digit dataset which has $60,000$ training samples and $10,000$ test samples in 10 classes. Each sample is a black and white image and 28×28 in size.

CIFAR10 and **CIFAR100** (Krizhevsky et al., 2009) are tiny image datasets which has $50,000$ training samples and $10,000$ test samples in 10 and 100 classes respectively. Each sample is a color image and 32×32 in size.

SVHN (Netzer et al., 2019) is a real-world digit dataset which has $73,257$ training samples and $26,032$ test samples in 10 classes. Each sample is a color image and 32×32 in size.

Tiny ImageNet is a subset of the ILSVRC (Russakovsky et al., 2015) dataset which has $100,000$ samples in 200 classes. Each sample is a color image and 64×64 in size. In experiments, Tiny ImageNet is split into two parts: $90,000$ samples for training and $10,000$ samples for test.

HAM10000 (Tschandl et al., 2018) is a imbalanced dermatoscopic image dataset which has $10,015$ samples in 7 classes. Each sample is a color image and resized to 75×75. In experiments, HAM10000 is split into two parts: $8,515$ samples for training and $1,500$ samples for test.

All datasets are used without any preprocessing of images.

B.2 Deep Networks

S-CNN (Chollet et al., 2015) consists of [3×3×32 conv $-$ 3×3× 64 conv $-$ 2×2 maxpool $-$ dropout (0.25) $-$ 128 dense $-$ dropout (0.5) $-$ # class dense $-$ softmax] layers, and it is used for MNIST.

K-CNN (Chollet et al., 2015) consists of [two 3×3×32 conv $-$ 2×2 maxpool - dropout (0.25) $-$ two 3×3×64 conv $-$ 2×2 maxpool - dropout (0.25) $-$ 512 dense $-$ dropout (0.5) $-$ # class dense - softmax] layers, and it is used for CIFAR10, SVHN, and CIFAR100.

WRN-16-8 (Zagoruyko & Komodakis, 2016) is a wide residual network that has 16 convolutional layers and a widening factor 8, and it is used for CIFAR100 and Tiny ImageNet.

B.3 EXPERIMENTAL SETTINGS

Training settings regarding number of epochs, batch size, optimizer, learning rate, and learning rate schedule are summarized in Table 3. The model parameters are initialized with He normal initialization (He et al., 2015) for all experimental settings. For all experiments, the initial labeled samples for each repetition are randomly sampled according to the distribution of the training set.

C VERIFICATION OF CONJECTURES

C.1 VERIFICATION OF CONJECTURE 1

Theoretical verification:

Consider the binary classification with a set of linear classifiers,

$$\mathcal{H} = \{h : h(x) = \text{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}), \boldsymbol{w} \in \mathcal{W} = \mathbb{R}^2\}$$

where $\boldsymbol{x}$ is uniformly distributed on $\mathcal{X} = \mathbb{R}^2$. By the duality between $\boldsymbol{w}$ and $\boldsymbol{x}$ (Tong & Chang, 2001), in $\mathcal{W}$, $\boldsymbol{w}$ is a point and $\boldsymbol{x}$ is represented by the hyperplane, $l_{\boldsymbol{x}} = \{\boldsymbol{w} \in \mathcal{W} : \text{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}) = 0\}$. Let h be a sampled hypothesis with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$, $\hat{\theta}$ be the angle of $\hat{\boldsymbol{w}} = (\hat{w}_1, \hat{w}_2)^\mathsf{T}$, i.e., $\tan\hat{\theta} = \hat{w}_2/\hat{w}_1$, θ be the angle of $\boldsymbol{w} = (w_1, w_2)^\mathsf{T}$, i.e., $\tan\theta = w_2/w_1$, and $\theta_{\boldsymbol{x}}$ be the angle between $l_{\boldsymbol{x}}$ and positive x-axis. Here, $\theta, \theta_{\boldsymbol{x}} \in [-\pi + \hat{\theta}, \pi + \hat{\theta}]$ in convenience. When $\theta_{\boldsymbol{x}}$ or $\pi + \theta_{\boldsymbol{x}}$ is between θ and $\hat{\theta}$, $h(\boldsymbol{x}) \neq \hat{h}(\boldsymbol{x})$, otherwise $h(\boldsymbol{x}) = \hat{h}(\boldsymbol{x})$. Thus, $\rho(h, \hat{h}) = |\theta - \hat{\theta}|/\pi$

Using Box-Muller transform (Box, 1958), $\boldsymbol{w}$ can be generated by

$$w_1 = \hat{w}_1 + \sigma\sqrt{-2\log u}\cos(2\pi v), \quad w_2 = \hat{w}_2 + \sigma\sqrt{-2\log u}\sin(2\pi v)$$

where u and v are independent uniform random variables on $[0, 1]$. Then, $\|\boldsymbol{w} - \hat{\boldsymbol{w}}\| = \sigma\sqrt{-2\log u}$ and $(w_2 - \hat{w}_2)/(w_1 - \hat{w}_1) = \tan(2\pi v)$, i.e., the angle of $\boldsymbol{w} - \hat{\boldsymbol{w}}$ is $2\pi v$. Here,

$$\|\hat{\boldsymbol{w}}\|\sin(\theta - \hat{\theta}) = \sigma\sqrt{-2\log u}\sin(2\pi v - \theta) \tag{10}$$

by using the perpendicular line from $\hat{\boldsymbol{w}}$ to the line passing through the origin and $\boldsymbol{w}$ (see the Figure 6a–6b for its geometry), and Eq. 10 is satisfied for all θ. For given u and v, θ is continuous and the derivative of θ with respect to σ is

$$\frac{d\theta}{d\sigma} = \frac{\sqrt{-2\log u}\sin^2(2\pi v - \theta)}{\|\hat{\boldsymbol{w}}\|\sin(2\pi v - \hat{\theta})}, \quad \text{thus} \quad \begin{cases} \frac{d\theta}{d\sigma} > 0, & v \in (\frac{\hat{\theta}}{2\pi}, \frac{\pi+\hat{\theta}}{2\pi}) \\ \frac{d\theta}{d\sigma} < 0, & v \in [0, \mathbf{1}] \setminus [\frac{\hat{\theta}}{2\pi}, \frac{\pi+\hat{\theta}}{2\pi}] \end{cases}.$$

Then,

$$\frac{d\rho(h, \hat{h})}{d\sigma} = \text{sgn}(\theta - \hat{\theta})\frac{d\theta}{d\sigma} > 0 \quad \text{where} \quad v \notin \left\{\frac{\hat{\theta}}{2\pi}, \frac{\pi + \hat{\theta}}{2\pi}\right\}.$$

Thus, $\rho(h, \hat{h})$ is continuous and strictly increasing with σ when $v \neq \hat{\theta}/2\pi$ or $v \neq (\pi + \hat{\theta})/2\pi$. Let $\rho(h, \hat{h}) = g(\sigma, u, v)$, then

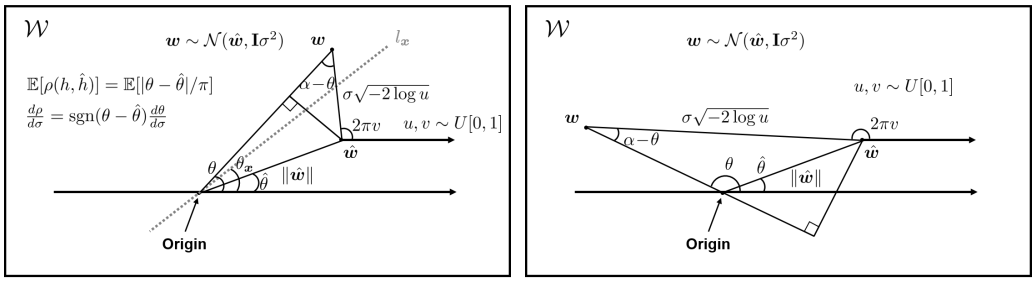
$$\mathbb{E}[\rho(h, \hat{h})] = \mathbb{E}[g(\sigma, u, v)] = \int g(\sigma, u, v)h(u)h(v)dudv$$

where $h(u) = \mathbb{I}[0 < u < 1]$ and $h(v) = \mathbb{I}[0 < v < 1]$. For $0 < \sigma_1 < \sigma_2$,

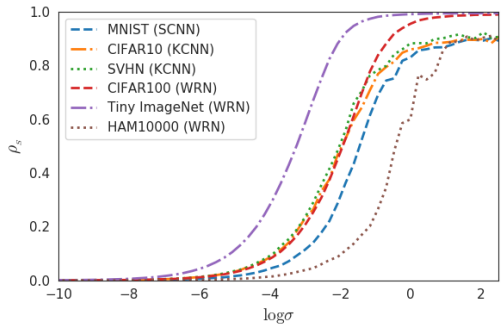
$$\mathbb{E}[g(\sigma_2, u, v)] - \mathbb{E}[g(\sigma_1, u, v)] = \int g(\sigma_2, u, v)h(u)h(v)dudv - \int g(\sigma_1, u, v)h(u)h(v)dudv > 0$$

Empirical verification:

Figure 6c shows the empirical results for various datasets with deep networks. The $\mathbb{E}[\rho_s(h, \hat{h})]$ is mostly continuous and strictly increasing with $\log\sigma$ when h is sampled with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ on MNIST, CIFAR10, SVHN, CIFAR100, Tiny ImageNet, and HAM10000 datasets. Therefore, these results empirically substantiates Conjecture 1.



(a) (b)



(c)

Figure 6: The verification of Conjecture 1. (a)–(b) Theoretical verification in binary classification with the linear classifier $h(x) = \mathrm{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w})$ on uniformly distributed $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^2$. Let h be a sampled hypothesis with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma)$, then $\rho(h,\hat{h}) = |\theta - \hat{\theta}|/\pi$ where $-\pi + \hat{\theta} \leq \theta \leq \pi + \hat{\theta}$. Here, $\rho(h,\hat{h})$ is continuous and strictly increasing with σ, thus $\mathbb{E}[\rho(h,\hat{h})]$ is continuous and strictly increasing with σ. (c) Empirical verification for various datasets with deep networks. The $\mathbb{E}[\rho_S(h,\hat{h})]$ is mostly continuous and strictly increasing with $\log\sigma$ when h is sampled with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$.

C.2 Verification of Conjecture 2

Theoretical verification:

Consider the binary classification with a set of linear classifiers,

$$\mathcal{H} = \{h : h(x) = \mathrm{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}), \boldsymbol{w} \in \mathcal{W} = \mathbb{R}^2\}$$

where $\boldsymbol{x}$ is uniformly distributed on $\mathcal{X} = \mathbb{R}^2$. By the duality between $\boldsymbol{w}$ and $\boldsymbol{x}$ (Tong & Chang, 2001), in $\mathcal{W}$, $\boldsymbol{w}$ is a point and $\boldsymbol{x}$ is represented by the hyperplane, $l_{\boldsymbol{x}} = \{\boldsymbol{w} \in \mathcal{W} : \mathrm{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}) = 0\}$.

Let $\mathcal{H}_B$ be the set of h sampled with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$, $\hat{\theta}$ be the angle of $\hat{\boldsymbol{w}}$, θ be the angle of $\boldsymbol{w}$, and $\theta_{\boldsymbol{x}}$ be the angle between $l_{\boldsymbol{x}}$ and positive x-axis as in Figure 7a. Here, the lines with angles $\theta_{\boldsymbol{x}}$ and $\pi + \theta_{\boldsymbol{x}}$ are same, thus we consider $\theta_{\boldsymbol{x}} \in [-\frac{\pi}{2} + \hat{\theta}, \frac{\pi}{2} + \hat{\theta}]$. Let

$$\mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}) = \begin{cases} \{\boldsymbol{w} : \theta \in (\theta_{\boldsymbol{x}}, \pi + \theta_{\boldsymbol{x}})\} & \theta_{\boldsymbol{x}} > \hat{\theta} \\ \{\boldsymbol{w} : \theta \in (-\pi + \theta_{\boldsymbol{x}}, \theta_{\boldsymbol{x}})\} & \theta_{\boldsymbol{x}} < \hat{\theta} \end{cases},$$

then $\mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x})$ is corresponding with $\mathcal{H}(\hat{h}, \boldsymbol{x})$, and thus $D(\hat{h}, \boldsymbol{x}) = |\mathcal{H}_B(\hat{h}, \boldsymbol{x})|/|\mathcal{H}_B| \to \mathbb{P}[\boldsymbol{w} \in \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x})]$ with probability tending to 1 as $|\mathcal{H}_B| \to \infty$. Let d_1, d_2 be the distances between $\hat{\boldsymbol{w}}$ and $l_{\boldsymbol{x}_1}, l_{\boldsymbol{x}_2}$ respectively, and

$$\mathcal{W}_1 = \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_1) \setminus \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_2), \quad \mathcal{W}_2 = \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_2) \setminus \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_1)$$

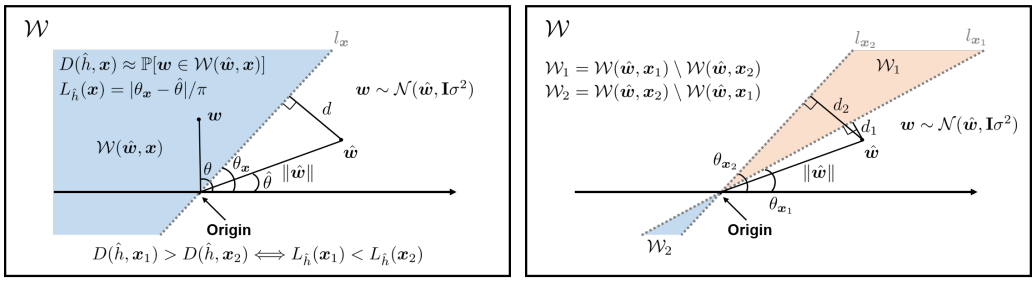


Figure 7: The verification of Conjecture 2. (a)–(b) Theoretical verification in binary classification with the linear classifier $h(x) = \text{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w})$ on uniformly distributed $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^2$. Let $\mathcal{H}_B$ be the set of h sampled with $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$. The $D(\hat{h}, \boldsymbol{x}) \to \mathbb{P}[\boldsymbol{w} \in \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x})]$ decreases as d increases with probability tending to 1 as $|\mathcal{H}_B| \to \infty$. While, $L_{\hat{h}}(\boldsymbol{x}) = |\theta_{\boldsymbol{x}} - \hat{\theta}|/\pi$ increases as d increases. Thus, $D(\hat{h}, \boldsymbol{x}_1) > D(\hat{h}, \boldsymbol{x}_2) \Leftrightarrow L_{\hat{h}}(\boldsymbol{x}_1) < L_{\hat{h}}(\boldsymbol{x}_2)$. (c) Empirical verification for various datasets with deep networks the strong negative rank correlation coefficients of from -0.95 to -0.94 between the empirical LPDR and the disagree ratio are observed for $\log\sigma \in (-6, -2)$.

as in Figure 7b. Suppose that $d_1 < d_2$, then

$$\mathbb{P}[\boldsymbol{w} \in \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_1)] = \mathbb{P}[\boldsymbol{w} \in \mathcal{W}(\hat{\boldsymbol{w}}, \boldsymbol{x}_2)] + \mathbb{P}[\boldsymbol{w} \in \mathcal{W}_1] - \mathbb{P}[\boldsymbol{w} \in \mathcal{W}_2] > 0$$

since $|\mathcal{W}_1| = |\mathcal{W}_2|$ and $\phi(\boldsymbol{w}_1|\hat{\boldsymbol{w}}, \sigma^2) > \phi(\boldsymbol{w}_2|\hat{\boldsymbol{w}}, \sigma^2)$ for all pairs of $\boldsymbol{w}_1 \in \mathcal{W}_1, \boldsymbol{w}_2 \in \mathcal{W}_2$ that are symmetric at the origin of $\hat{\boldsymbol{w}}$, where $\phi(\cdot|\hat{\boldsymbol{w}}, \sigma^2)$ is the probability density function of bivariate normal distribution with mean $\hat{\boldsymbol{w}}$ and covariance matrix $\mathbf{I}\sigma^2$. Thus,

$$d_1 < d_2 \iff D(\hat{h}, \boldsymbol{x}_1) > D(\hat{h}, \boldsymbol{x}_2)$$

with probability tending to 1 as $|\mathcal{H}_B| \to \infty$.

Meanwhile, $L_{\hat{h}}(\boldsymbol{x}) = |\theta_{\boldsymbol{x}} - \hat{\theta}|/\pi$ and $d_i = \|\hat{\boldsymbol{w}}\| \sin|\theta_{\boldsymbol{x}_i} - \hat{\theta}|$, then

$$d_1 < d_2 \iff |\theta_{\boldsymbol{x}_1} - \hat{\theta}| < |\theta_{\boldsymbol{x}_2} - \hat{\theta}| \iff L_{\hat{h}}(\boldsymbol{x}_1) < L_{\hat{h}}(\boldsymbol{x}_2).$$

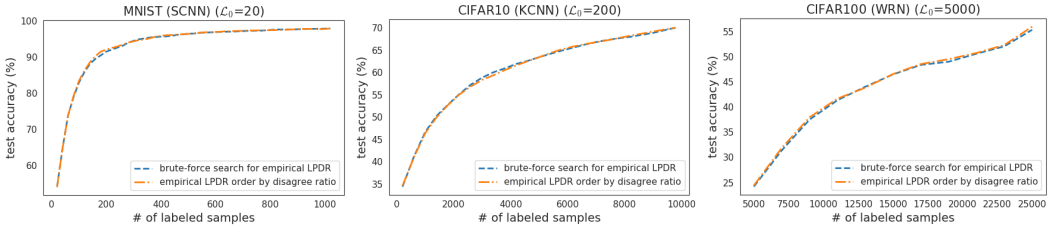
Therefore,

$$D(\hat{h}, \boldsymbol{x}_1) > D(\hat{h}, \boldsymbol{x}_2) \iff L_{\hat{h}}(\boldsymbol{x}_1) < L_{\hat{h}}(\boldsymbol{x}_2)$$

with probability tending to 1 as $|\mathcal{H}_B| \to \infty$.

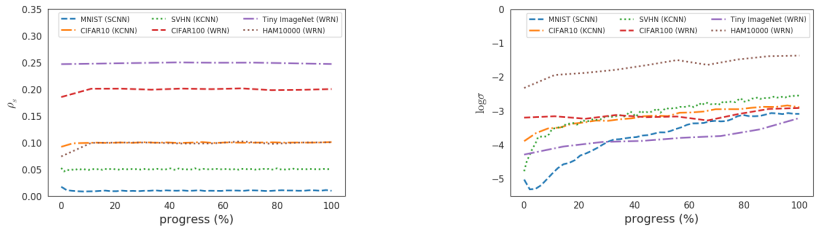
Empirical verification:

Figure 7c shows the empirical results for various datasets with deep networks. Spearman's rank correlation coefficient (Spearman, 1904) between the empirical LPDR and the disagree ratio is close to -1 in certain range of σ in all experimental settings–strong negative correlation coefficients of from -0.95 to -0.94 are observed for $\log\sigma \in (-6, -2)$. Therefore, these results empirically substantiates Conjecture 2.



(a) (b) (c)

Figure 8: The performance comparison between using empirical LPDR and using the disagree ratio on MNIST with S-CNN (a), CIFAR10 with K-CNN (b), and CIFAR100 with WRN-16-8 (c). There is no significant difference in the performance between the two methods. Thus, LPDR-based active learning can be performed by using the disagree ratio.



(a) (b)

Figure 9: The $\mathbb{E}[\rho_S(h_n, \hat{h})]$ and $\log \sigma$ with respect to the labeling proceeds for all experimental settings. The proposed algorithm reliably guides the $\mathbb{E}[\rho_S(h_n, \hat{h})]$ to be the target value by increasing the variance of sampling as the number of labeled samples increases.

D Empirical LPDR vs Disagree Ratio

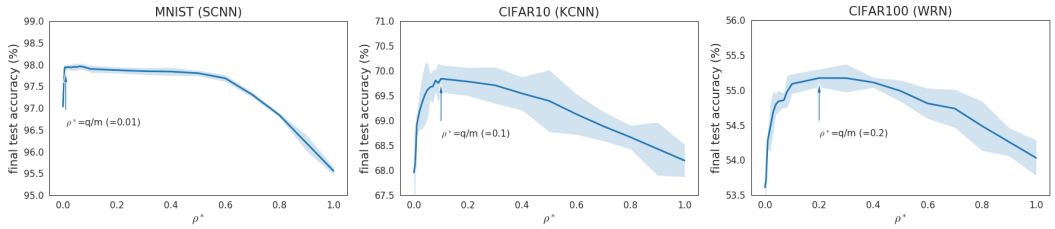
Figure 8 compares the active learning performance between evaluating empirical LPDR by brute-force search and evaluating the order of empirical LPDR by using disagree ratio on MNIST with S-CNN, CIFAR10 with K-CNN, and CIFAR100 with WRN-16-8. In all cases, there is no significant difference in the performance between the two methods. As a result, LPDR-based active learning can be conducted without evaluating empirical LPDR instead empirical LPDR order is obtained through disagree ratio.

E Regulating σ to Keep $\mathbb{E}[\rho_S(h_n, \hat{h})]$ Static at ρ^*

Figure 9a shows $\mathbb{E}[\rho_S(h_n, \hat{h})]$ in Algorithm 1 with respect to the active learning progress. For all experiments, the proposed algorithm reliably guides the $\mathbb{E}[\rho_S(h_n, \hat{h})]$ to be $\rho^* = q/m$ (MNIST: 0.01, CIFAR10: 0.1, SVHN: 0.05, CIFAR100: 0.2, Tiny ImageNet: 0.25, HAM10000: 0.1). Figure 9b shows $\log \sigma$ with respect to the active learning progress. For all experiments, the σ increases as the labeling proceeds. As the number of labeled samples increases, the larger variance is required to keep $\mathbb{E}[\rho_S(h_n, \hat{h})]$ static at ρ^* for unlabeled samples move away from the decision boundary of $\hat{h}$ due to an increase in the network confidence.

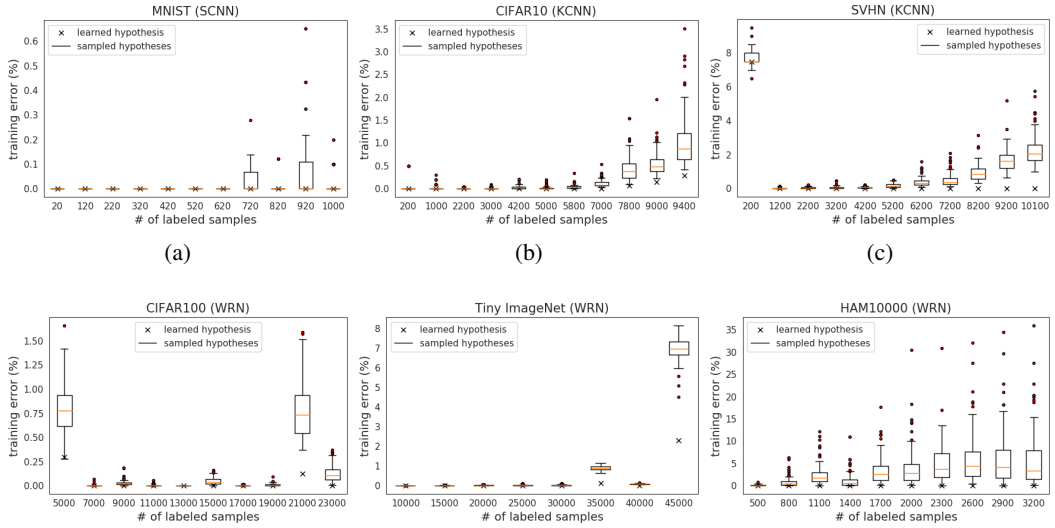
F Final Test Accuracy vs ρ^*

Figure 10 shows the final test accuracy with respect to ρ^* on MNIST, CIFAR10, and CIFAR100 datasets. The results show that the proposed algorithm performs well at around $\rho^* = q/m$ (MNIST: 0.01, CIFAR10: 0.1, CIFAR100: 0.2). In addition, the range of ρ^*, associated with the high performance, is wide; thus, DRAL is robust against the ρ^* in the wide range.



(a) (b) (c)

Figure 10: The final accuracy with respect to the ρ^* on MNIST (a), CIFAR10 (b), and CIFAR100 (c) datasets. The proposed algorithm shows high performance when $\rho^* = q/m$.



(a) (b) (c)

(d) (e) (f)

Figure 11: The empirical errors of the learned and the sampled hypotheses with respect to the number of labeled samples for MNIST (a), CIFAR10 (b), SVHN (c), CIFAR100 (d), Tiny ImageNet (e), and HAM10000 (f) datasets. It is observed that the empirical errors of the sampled hypotheses have various values.

G Empirical Errors of Sampled Hypotheses

Figure 11 shows the empirical errors on $\mathcal{L}$ of the sampled hypotheses with respect to the number of labeled samples for MNIST, CIFAR10, SVHN, CIFAR100, Tiny ImageNet, and HAM10000 datasets. The empirical errors of sampled hypotheses have various values.

H DRAL$^+$ vs DRAL

Figure 12 shows the performance comparison between DRAL$^+$ and DRAL with respect to the number of labeled samples on MNIST, CIFAR10, SVHN, CIFAR100, Tiny ImageNet, and HAM10000 datasets. Overall, DRAL$^+$ consistently either performs better or comparable with DRAL regardless of the experimental settings. When the empirical errors of the sampled hypotheses are mostly zero such as the results of MNIST, CIFAR100, and Tiny ImageNet, there is no significant difference in performance between DRAL$^+$ and DRAL. However, when the mean empirical error of the sampled hypotheses is larger than the empirical error of the learned hypothesis such as the results of CIFAR10, SVHN, and HAM10000, DRAL$^+$ brings a significant performance improvement compared

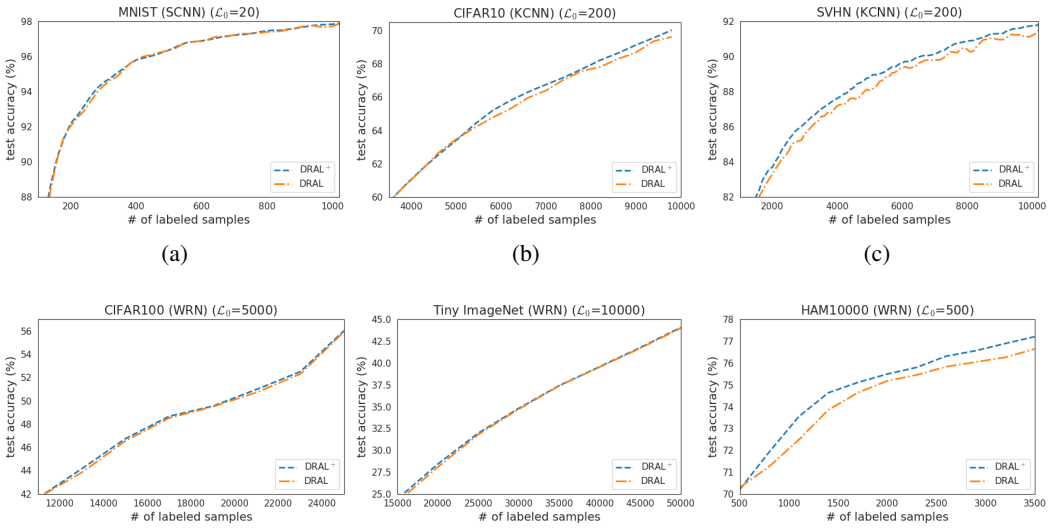


Figure 12: The performance comparison between DRAL$^+$ and DRAL on MNIST (a), CIFAR10 (b), SVHN (c), CIFAR100 (d), Tiny ImageNet (e), and HAM10000 (f) datasets. Overall, DRAL$^+$ consistently either performs better than or comparable with DRAL regardless of the experimental settings.

to DRAL. The larger the variance in the empirical errors of the sampled hypotheses, the larger the performance gap between DRAL$^+$ and DRAL tends to be.

I Robustness of DRAL$^+$ against Hyperparameters

DRAL$^+$ has four hyperparameters: 1) the initial $\sigma = \sigma_0$, 2) the positive hyperparameter β, 3) the number of sampled hypotheses N, and 4) the layer of the network to which sampling is applied. The σ_0 has no significant effect on the performance of DRAL$^+$ for σ is adaptively regulated to make $\rho' = \rho^*$ while hypothesis sampling. Figure 13 shows the performance comparison with respect to β, N, and sampling layer on MNIST, CIFAR10, and CIFAR100 datasets. Figure 13a–13c show that there is no significant performance difference for various $\beta \in \{0.1, 1, 10\}$ on all datasets. The robustness against β is based on the sufficient buffer for regulating σ since the range of ρ^* associated with the best performance is wide. Figure 13d–13f show that there is no significant performance difference for various $N \in \{5, 10, 20, 50, 100, 200\}$ on all datasets. Figure 13g–13i show that there is no significant performance difference whether parameter sampling is applied to the entire layers or on the last layer of the network. The robustness against N or sampling layer is based on the sufficient discrimination in the disagree ratio for identifying q most informative unlabeled samples with a small number of sampled hypotheses by setting $\rho^* = q/m$.

J Running Time

In Table 4, the mean of running time for active learning are given for all algorithms and datasets. The unit is minutes, and the value in parentheses is the ratio to Entropy. The running time of DRAL$^+$ increased by only 1-7% compared to Entropy on all datasets except MNIST, and it is comparable to MC-BALD, MC-VarR, and Coreset. The relatively large running time in MNIST is because it takes a very short time to train the model compared to that of the acquisition. Ens-VarR requires about 5 times more computational load than DRAL$^+$ on all datasets, and BADGE requires twice and more than eight times the computational load on CIFAR100 and TinyImageNet datasets, respectively.

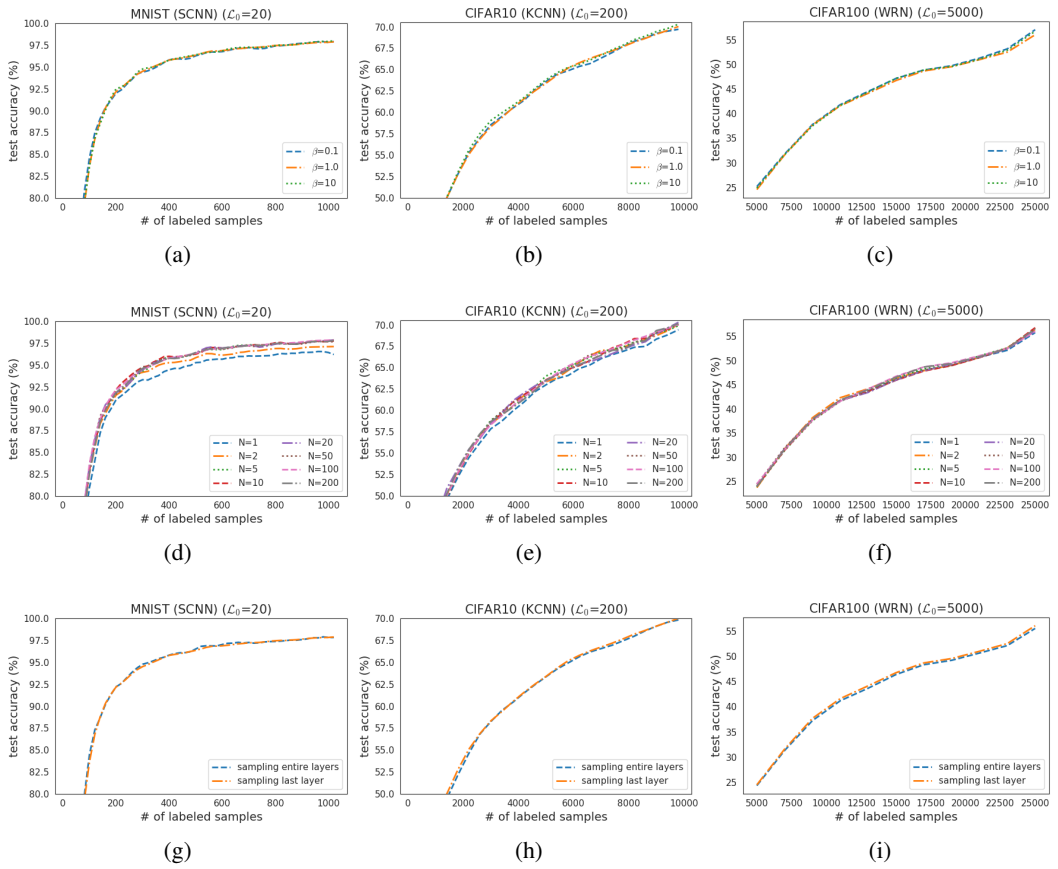


Figure 13: The performance comparison with respect to the hyperparameters of DRAL$^+$ on MNIST, CIFAR10, and CIFAR100 datasets. (a) – (c) β. (d) – (f) N. (g) – (i) sampling layer. DRAL$^+$ is robust against β, N, and sampling layer.

Table 4: The mean of running time (minutes) for active learning are given for all algorithms and datasets. The value in parentheses is the ratio to Entropy. We observe that DRAL$^+$ operates as fast as Entropy, and that ENS-VarR or BADGE require a large computational load.

	MNIST	CIFAR10	SVHN	CIFAR100	T. ImageNet	HAM10000
DRAL$^+$[ours]	7.9 (139)	100 (107)	192 (104)	396 (103)	4,589 (101)	311 (101)
Entropy[41]	5.7 (100)	93 (100)	186 (100)	385 (100)	4,547 (100)	307 (100)
Coreset[36]	14.4 (254)	139 (149)	237 (128)	439 (114)	4,722 (104)	321 (104)
MC-BALD[14]	6.8 (119)	99 (106)	180 (97)	441 (115)	4,828 (106)	403 (131)
MC-VarR[10]	6.8 (119)	101 (108)	187 (101)	443 (115)	4,861 (107)	403 (131)
ENS-VarR[3]	24.9 (438)	575 (616)	885 (477)	2,274 (591)	23,394 (515)	1,744 (568)
BADGE[1]	11.0 (193)	141 (151)	238 (128)	864 (224)	39,265 (864)	326 (106)

K Results for Test Accuracy

Figure 14 shows the test accuracy with respect to the number of labeled samples from initial to final step for all experimental settings.

L Robustness of DRAL$^+$ against Initial Labeled Size

Figure 15 shows the performance comparison with respect to the number of initial labeled samples on MNIST, CIFAR10, and CIFAR100 datasets. There is no significant performance difference ac-

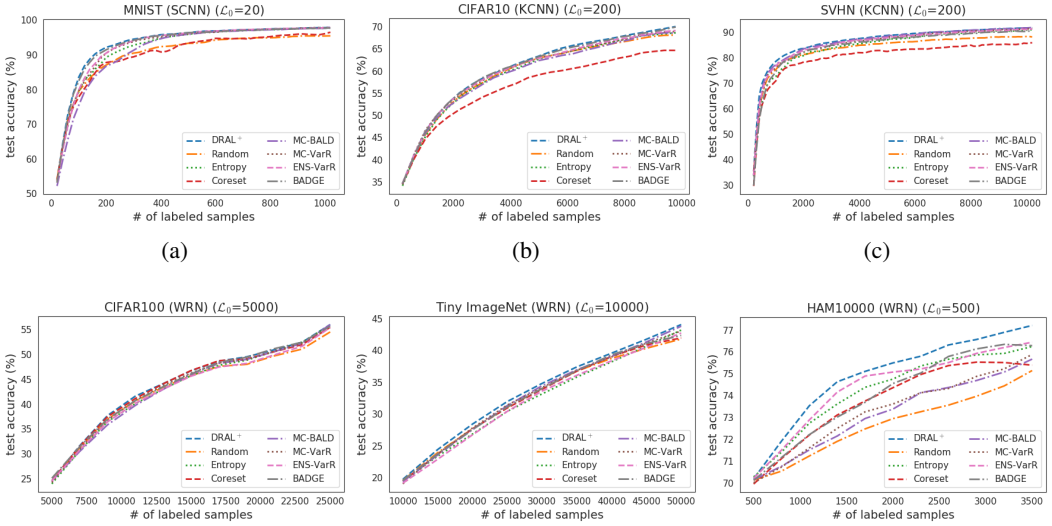


Figure 14: The test accuracy with respect to the number of labeled samples from initial to final step for all experimental settings.

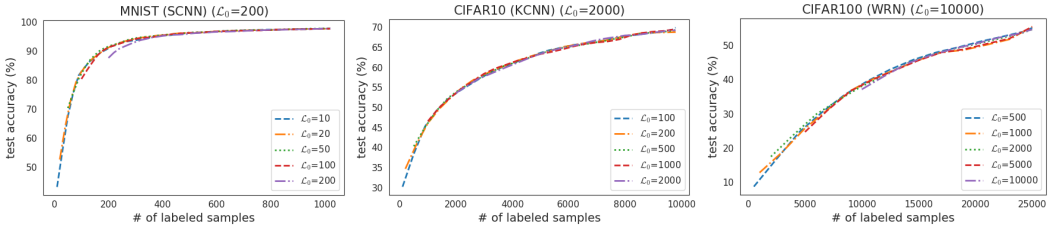


Figure 15: The performance comparison with respect to the number of initial labeled samples on MNIST (a), CIFAR10 (b), and CIFAR100 (c) datasets. The proposed algorithm is robust against to the number of initial labeled samples and performs well even when the initial size is much smaller.

cording to the number of initial labeled samples. The proposed algorithm is robust against to the number of initial labeled samples and performs well even when the initial size is much smaller.

M FRAMEWORK OF DRAL$^+$

Figure 16 shows the framework of the proposed algorithm.

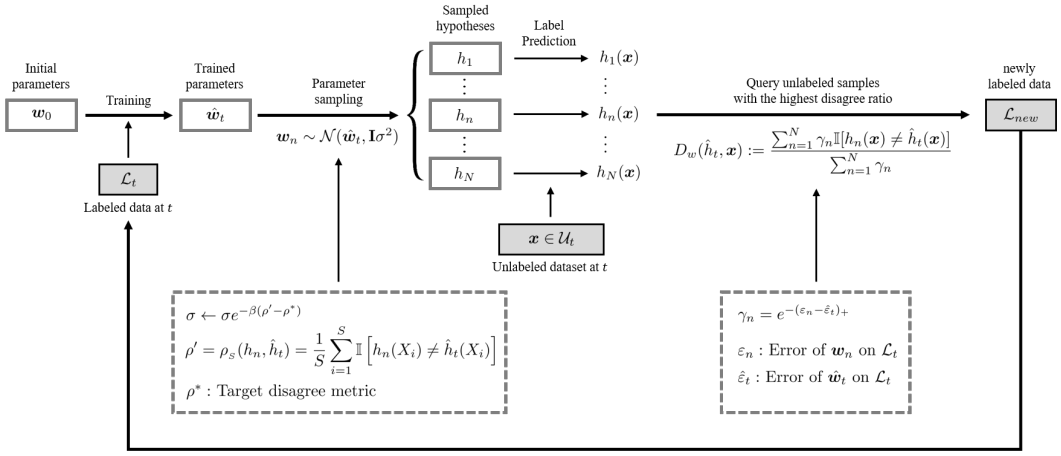


Figure 16: The framework of the proposed algorithm.