LEAST PROBABLE DISAGREEMENT REGION FOR ACTIVE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Active learning strategy to query unlabeled samples nearer the estimated decision boundary at each step has been known to be effective when the distance from the sample data to the decision boundary can be explicitly evaluated; however, in numerous cases in machine learning, especially when it involves deep learning, conventional distance such as the ℓ_p from sample to decision boundary is not readily measurable. This paper defines a theoretical distance of unlabeled sample to the decision boundary as the least probable disagreement region (LPDR) containing the unlabeled sample, and it discusses how this theoretical distance can be empirically evaluated with a lower order of time complexity. Monte Carlo sampling of the hypothesis is performed in approximating the theoretically defined distance. Experimental results on various datasets show that the proposed algorithm consistently outperforms all other high performing uncertainty based active learning algorithms and leads to state-of-the-art active learning performance on CIFAR10, CIFAR100, Tiny ImageNet and Food101 datasets. Only the proposed algorithm outperforms random sampling on CIFAR100 dataset using K-CNN while all other algorithms fail to do so.

1 INTRODUCTION

Active learning (Cohn et al., 1996) is a subfield of machine learning to attain data efficiency with fewer labeled training data when it is allowed to choose the training data from which to learn. For many real-world learning problems, large collections of unlabeled samples is assumed available, and based on a certain query strategy, the label of the most informative data is iteratively queried to an oracle to be used in retraining the model (Bouneffouf et al., 2014; Roy & McCallum, 2001; Sener & Savarese, 2017b; Settles et al., 2008; Sinha et al., 2019; Sener & Savarese, 2017a; Pinsler et al., 2019; Shi & Yu, 2019; Gudovskiy et al., 2020). Active learning attempts to achieve high accuracy using as few labeled samples as possible (Settles, 2009).

Of the possible query strategies, uncertainty-based sampling (Culotta & McCallum, 2005; Scheffer et al., 2001; Mussmann & Liang, 2018), which enhances the current model by labeling unlabeled samples that are difficult for the model to predict, is a simple strategy commonly used in pool-based active learning (Lewis & Gale, 1994). Nevertheless, many existing uncertainty-based algorithms have their own limitations. Entropy (Shannon, 1948) based uncertainty sampling can query unlabeled samples near the decision boundary for binary classification, but it does not perform well in multiclass classification as entropy does not equate well with the distance to a complex decision boundary (Joshi et al., 2009). Another approach based on MC-dropout sampling (Gal et al., 2017) which uses a mutual information based BALD (Houlsby et al., 2011) as an uncertainty measure identifies unlabeled samples that are individually informative. This approach, however, is not necessarily informative when it is jointly considered with other samples for label acquisition. To address this problem, BatchBALD (Kirsch et al., 2019) is introduced. However, BatchBALD computes, theoretically, all possible joint mutual information of batch, and is infeasible for large query size. The ensemble method (Beluch et al., 2018), one of the query by committee (QBC) algorithm (Seung et al., 1992), has been shown to perform well in many cases. The fundamental premise behind the QBC is minimizing the version space (Mitchell, 1982), which is the set of hypotheses that are consistent with labeled samples. However, the ensemble method requires high computation load because all networks that make up the ensemble must be trained.

This paper defines a theoretical distance referred to as the least probably disagreement region (LPDR) from sample to the estimated decision boundary, and in each step of active learning, labels of unlabeled samples nearest to the decision boundary in terms of LPDR are obtained to be used for retraining the classifier to improve accuracy of the estimated decision boundary. It is generally understood that labels to samples near the decision boundary are the most informative as the samples are uncertain. Indeed in Balcan et al. (2007), selecting unlabeled samples with the smallest margin to the linear decision boundary and thereby minimal certainty attains exponential improvement over random sampling in terms of sample complexity. In deep learning, it is difficult to identify samples nearest to the decision boundary as sample distance to decision boundary is difficult to evaluate. An adversarial approach (Ducoffe & Precioso, 2018) to approximate the sample distance to decision boundary has been studied but this method does not show preservation of the order of the sample distance and requires considerable computation in obtaining the distance.

2 Distance: Least Probable Disagreement Region (LPDR)

Let $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{H}$ and $\mathcal{D}$ be the instance space, the label space, the set of hypotheses $h : \boldsymbol{x} \to y$ and the joint distribution over $\mathcal{X} \times \mathcal{Y}$ where $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathcal{Y}$ respectively. The distance between two hypotheses $\hat{h}$ and h is defined as the probability of the disagreement region for $\hat{h}$ and h:

$$\rho(\hat{h}, h) := \mathbb{P}_{\mathcal{D}}[\hat{h}(X) \neq h(X)]. \tag{1}$$

This paper defines the sample distance d of $\boldsymbol{x}$ to the hypothesis $\hat{h} \in \mathcal{H}$ based on ρ as the least probable disagreement region that contains $\boldsymbol{x}$:

$$d(\boldsymbol{x}, \hat{h}) := \inf_{h \in \mathcal{H}_{\boldsymbol{x}}} \rho(\hat{h}, h) \tag{2}$$

where $\mathcal{H}_{\boldsymbol{x}} = \{h \in \mathcal{H} : \hat{h}(\boldsymbol{x}) \neq h(\boldsymbol{x})\}$. Here, the sample distribution $\mathcal{D}$ is unknown, and $\mathcal{H}_{\boldsymbol{x}}$ may be uncountably infinite. Therefore, a systematic and empirical method for evaluating the distance is required. One might be as follows: Let $B(\hat{h}, r) := \{h \in \mathcal{H} : \rho(\hat{h}, h) \leq r\}$ be a ball centered at $\hat{h} \in \mathcal{H}$ of radius $r \geq 0$. Determine ρ_1 such that $\hat{h}(\boldsymbol{x}) = h_1(\boldsymbol{x})$ for all $h_1 \in B(\hat{h}, \rho_1)$ and also determine ρ_2 such that there exists $h_2 \in B(\hat{h}, \rho_2)$ satisfying $\hat{h}(\boldsymbol{x}) \neq h_2(\boldsymbol{x})$ then perform grid search between ρ_1 and ρ_2 to narrow the gap between ρ_1 and ρ_2. Sampling the hypotheses within the ball can be performed by sampling the corresponding parameters with the assumption that small perturbation in model parameter leads to small perturbation in hypothesis (see Assumption 1). This scheme is based on performing grid search on each unlabeled sample and is therefore computationally inefficient. By introducing a hypothesis set $\mathcal{H}'$ that meets Assumption 2, unlabeled samples can be ordered according to d without grid search, and $\mathcal{H}'$ can be constructed by Monte-Carlo sampling from a symmetric distribution with mean of $\hat{h}$.

Assumption 1. *If the parameters of two hypotheses are close, the corresponding hypotheses are also close, i.e.,* $\|\hat{\boldsymbol{w}} - \boldsymbol{w}_1\| \leq \|\hat{\boldsymbol{w}} - \boldsymbol{w}_2\|$ *implies that* $\rho(\hat{h}, h_1) \leq \rho(\hat{h}, h_2)$ *where* $\hat{\boldsymbol{w}}, \boldsymbol{w}_1$ *and* $\boldsymbol{w}_2$ *are the parameters pertaining to* $\hat{h}, h_1$ *and* h_2 *respectively.*

Assumption 2. *There exists a set,* $\mathcal{H}'$*, of hypotheses sampled around* $\hat{h}$ *with the property that large variation ratio for a given sample data implies small sample distance to* $\hat{h}$*, i.e., there exists* $\mathcal{H}'$ *such that* $V(\boldsymbol{x}_1) \geq V(\boldsymbol{x}_2)$ *implies that* $d(\boldsymbol{x}_1, \hat{h}) \leq d(\boldsymbol{x}_2, \hat{h})$ *where* $V(\boldsymbol{x}) = 1 - f_m^{(\boldsymbol{x})}/|\mathcal{H}'|$ *and* $f_m^{(\boldsymbol{x})}$ *is the frequency of the modal class predicted by* $\mathcal{H}'$ *on* $\boldsymbol{x}$*.*

3 Empirical Studies of LPDR

3.1 Hypotheses and Parameters In Deep Networks: Assumption 1

The distance between two hypotheses can be approximated by vectors of the predicted labels on random samples by the hypotheses:

$$\rho(\hat{h}, h) \approx \rho_e(\hat{h}, h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\left[\hat{h}(\boldsymbol{x}^{(i)}) \neq h(\boldsymbol{x}^{(i)})\right] \tag{3}$$

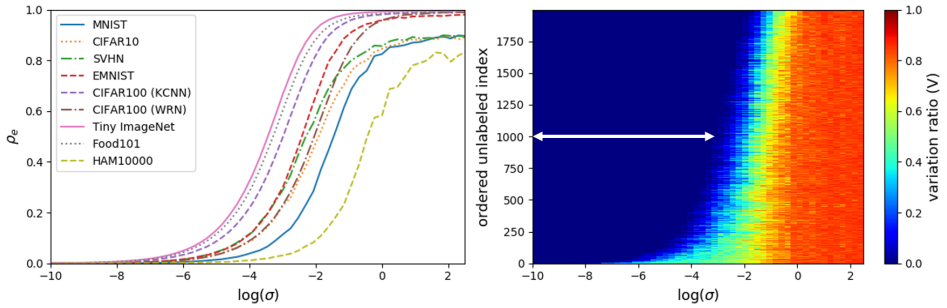

Figure 1: Empirical validation of Assumption 1. Left figure: Relationship between approximated hypothesis distance and σ at step $t = 0$. Hypothesis distance is almost linearly proportional to $\log(\sigma)$ in the ascension. Right figure: Relationship between variation ratio and σ. Sample distance to the decision boundary can be expressed as σ at which the variation ratio is not zero for the first time (MNIST). The unlabeled samples are ordered in terms of LPDR.

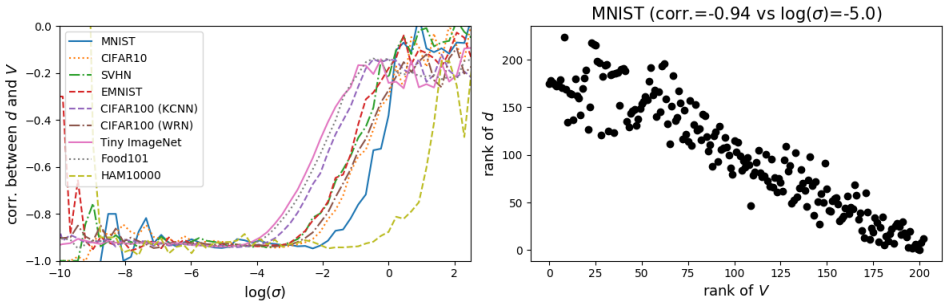

Figure 2: Empirical validation of Assumption 2. Left figure: Spearman's rank correlation coefficient between LPDR and the variation ratio in terms of σ showing that there exists a σ such that LPDR and the variation ratio have a strong rank correlation. Right figure: An example of strong negative correlation between both ranks when $\log(\sigma) = -5.0$. Samples with increasing LPDR or variation ratio are ranked from high to low.

where $\boldsymbol{x}^{(i)}$ is the ith sample for $i \in [m]$ and $\mathbb{I}[\cdot]$ is an indicator function. The h is sampled by sampling model parameter $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ where $\hat{\boldsymbol{w}}$ is the model parameter of $\hat{h}$, and the expectation of distances between $\boldsymbol{w}$ and $\hat{\boldsymbol{w}}$ depends on σ. The ρ_e is obtained by the average of 100 times for a fixed σ. The left-hand side of Figure 1 shows the relationship between ρ_e and σ on various datasets and deep networks. The ρ_e increases almost monotonically as σ increases. This implies that the order is preserved between the σ and ρ_e. Furthermore, the ρ_e is almost linearly proportional to $\log(\sigma)$ in the ascension of the graph, i.e., $\sigma \propto e^{\beta \rho_e}$ for some $\beta > 0$. The right-hand side of Figure 1 shows the variation ratio V of predicted labels by the sampled hypotheses with respect to σ for each unlabeled sample on MNIST. The sample distance to the decision boundary can be expressed as σ at which the variation ratio is not zero for the first time, where the indices of unlabeled samples in y-axis are ordered by LPDR. The variation ratio increases as the σ increases, and it is expected that the data point with short distance has the large variation ratio compared to the data point with long distance on a certain range of σ.

3.2 LPDR vs Variation Ratio: Assumption 2

The left-hand side of Figure 2 shows Spearman's rank correlation coefficient between LPDR and the variation ratio with respect to σ. The correlation is calculated using only unlabeled samples whose variation ratio is not 0. The strong rank correlation is verified when the σ has the appropriate value. Too larger value of σ generates hypotheses too far away from $\hat{h}$, which is not helpful to measure the distance. The right-hand side of Figure 2 shows an example of σ ($\log(\sigma) = -5.0$) makes that LPDR and the variation ratio have a strong negative correlation on MNIST, that is, the data point with larger variation is closer to the decision boundary. Results for various datasets and networks are presented in Appendix C.

Algorithm 1 Least Probable Disagree Region (LPDR)

Input:
$\mathcal{L}_0, \mathcal{U}_0$: Initial labeled and unlabeled samples
m, q : Size of pool data and number of queries
σ_0^2 : Initial variance for sampling
ρ^* : Target hypothesis distance $(= q/m)$

Procedure:

for step $t = 0, 1, 2, \ldots, T-1$
  Train parameters $\hat{\boldsymbol{w}}_t$ with $\mathcal{L}_t$, then evaluate its empirical error $\hat{\varepsilon}_t$ on $\mathcal{L}_t$
  $\sigma_t \to \sigma'_1$
  for $n = 1, 2, \ldots, N$
    Sample parameters $\boldsymbol{w}'_n \sim \mathcal{N}(\hat{\boldsymbol{w}}_t, \mathbf{I}\sigma'^2_n)$ for h'_n
    Compute $\gamma_n = e^{-(\varepsilon'_n - \hat{\varepsilon}_t)_+}$ where ε'_n is empirical error of $\boldsymbol{w}'_n$ on $\mathcal{L}_t$
    Compute $\rho'_n = \rho_e(\hat{h}_t, h'_n)$
    Update $\sigma'_{n+1} = \sigma'_n e^{-\beta(\rho'_n - \rho^*)}$ where $\beta > 0$
  end for
  $\sigma'_N \to \sigma_{t+1}$
  Compute $V_w(\boldsymbol{x}^{(i)}) = 1 - f_m^{(i)}/N$ where $f_m^{(i)} = \max_c \sum_{n=1}^N \gamma_n \mathbb{I}\left[h'_n(\boldsymbol{x}^{(i)}) = c\right]$
  Get $\mathcal{I}^* = \arg\max_{\mathcal{I} \subset \mathcal{I}_{\mathcal{P}_t}, |\mathcal{I}|=q} \sum_{i \in \mathcal{I}} V_w(\boldsymbol{x}^{(i)})$ where $\mathcal{I}_{\mathcal{P}_t} = \left\{j : \boldsymbol{x}^{(j)} \in \mathcal{P}_t\right\}$
  Update $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i \in \mathcal{I}^*}$ and $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus \left\{\boldsymbol{x}^{(i)}\right\}_{i \in \mathcal{I}^*}$
end for

The time complexity is discussed to validate the efficiency of using variation ratio. Let m, N and n_σ be the unlabeled sample size, $|\mathcal{H}'|$ and the number of grid for σ respectively. Ordering unlabeled samples in terms of LPDR by grid search with respect to σ requires the time complexity of $m \times N \times n_\sigma$ (see the right-hand side of Figure 1). However, using variation ratio for ordering unlabeled samples reduces the time complexity to $m \times N$. In the case of $n_\sigma = cN$ for some $c > 0$, then the time complexity can be reduced from $O(mN^2)$ to $O(mN)$.

4 Algorithm for LPDR

4.1 Framework

Let $\mathcal{L}_t$ and $\mathcal{U}_t$ be the labeled and unlabeled samples at step t. At step t, LPDR trains model parameters $\hat{\boldsymbol{w}}_t$ using labeled samples $\mathcal{L}_t$, and constructs $\mathcal{H}'$ by sampling the model parameters $\boldsymbol{w}'_n \sim \mathcal{N}(\hat{\boldsymbol{w}}, \mathbf{I}\sigma^2)$ for $n \in [N]$. Then, LPDR queries the top q unlabeled samples having highest variation ratio from the pool data $\mathcal{P}_t \subset \mathcal{U}_t$ of size m.

4.2 Construction of Sampled Hypothesis Set

It is important to set an appropriate σ when constructing $\mathcal{H}'$ as variation ratios goes to 0 with decreasing σ (see the right-hand side of Figure 1) and the rank correlation goes to zero with increasing σ (see the left-hand side of Figure 2). Theoretically, let's consider the binary classification with logistic regression where the predicted label is defined as $y = \mathrm{sgn}(\boldsymbol{x}^\mathsf{T}\boldsymbol{w})$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|\boldsymbol{x}\|_\infty < \infty$. Then the following theorem holds and the proof is described in Appendix A.

Theorem 1. *Suppose that $\boldsymbol{w}'_n$ for $n = 1, \ldots, N$ are generated with the variance of σ^2. For all $\boldsymbol{x}$, the followings hold: 1) As $N \to \infty$, $1 - f_m^{(\boldsymbol{x})}/N$ goes to 0 in probability as σ^2 goes to 0, 2) As $N \to \infty$, $1 - f_m^{(\boldsymbol{x})}/N$ goes to $1/2$ in probability as σ^2 goes to ∞.*

The implication of Theorem 1 is that when σ is too small or too large, it would be difficult to compare the sample distances of unlabeled samples. In this active learning task, at least q most informative unlabeled samples must be identified. To meet this condition, the ρ'_n denoted in the Algorithm 1 must be on average at least $\rho^* = q/m$ for N hypotheses, and setting ρ'_n to ρ^*, which is not very small and is less than $1/2$ in general, maintains negative rank correlation between the sample distance and the variation ratio (see the left-hand side of Figure 2). This can be attained by

updating σ'_n as $\sigma'_{n+1} = \sigma'_n e^{-\beta(\rho'_n - \rho^*)}$ where $\beta > 0$ (see Appendix D). The Figure 9 in Appendix E shows the final test accuracy with respect to target ρ'_n on MNIST dataset. The LPDR performs best when the target ρ'_n is around $\rho^* = q/m$. In addition, the range of target ρ'_n, associated with the best performance, is wide; thus, LPDR is relatively robust against target ρ'_n. Furthermore, LPDR is robust against hyperparameters β, N and sampling layers (see Appendix F).

Meanwhile, the efficiency of querying samples in the disagreement region of the version space is well known both theoretically (Hanneke et al., 2014) and empirically (Beluch et al., 2018). When the version space exists, the sampled hypotheses are in the version space with high probability, but there are cases where they are outside the version space (see Appendix G). Thus, LPDR gives weight γ_n on the prediction of sampled hypothesis h'_n where $\gamma_n = e^{-\left(\varepsilon'_n - \hat{\varepsilon}\right)_+}$ is a function of $\hat{\varepsilon} = \mathrm{err}_{\mathcal{L}}(\hat{h})$ and $\varepsilon'_n = \mathrm{err}_{\mathcal{L}_t}(h'_n)$. Here, $(\cdot)_+$ is $\max\{0, \cdot\}$ and $\mathrm{err}_{\mathcal{L}}(h)$ is the empirical error of h on $\mathcal{L}$. Then, LPDR uses weighted variation ratio V_w as a function of the weighted frequency of the modal class f_m as defined below:

$$V_w(\boldsymbol{x}^{(i)}) = 1 - \frac{f_m^{(i)}}{N} \tag{4}$$

where $f_m^{(i)} = \max_c \sum_{n=1}^{N} \gamma_n \mathbb{I}\left[h'_n(\boldsymbol{x}^{(i)}) = c\right]$ and $\boldsymbol{x}^{(i)} \in \mathcal{P}_t$.

If $\mathcal{H}'$ is a subset of the version space as in the case of CIFAR10 in Appendix G, the sample complexity of LPDR follows Hanneke's theorem. Let Λ be the sample complexity defined as the smallest integer t such that for all $t' \geq t$, $\mathrm{err}(h_{t'}) \leq \epsilon$ where $\mathrm{err}(h) := \mathbb{P}_{\mathcal{D}}[h(X) \neq Y]$ with probability at least $1 - \delta$.

Theorem 2. *[Hanneke, 2014] LPDR achieves a sample complexity Λ such that, for $\mathcal{D}$ in the realizable case, $\forall \epsilon, \delta \in (0, 1)$,*

$$\Lambda(\epsilon, \delta, \mathcal{D}) \lesssim \xi \cdot \left(D \log \xi + \log \left(\frac{\log(1/\epsilon)}{\delta} \right) \right) \cdot \log \frac{1}{\epsilon}$$

where D and ξ are the VC-dimension of $\mathcal{H}$ and the disagree coefficient with respect to $\mathcal{H}$ and $\mathcal{D}$.

When $\xi = O(1)$, in terms of ϵ, the number of labeled samples required by LPDR is just $O(\log(1/\epsilon) \cdot \log\log(1/\epsilon))$, while the number of labeled samples by a passive learning is $\Omega(1/\epsilon)$. Therefore, in this case, LPDR provides an exponential improvement over passive learning in sample complexity (Hsu, 2010).

5 Experiments

This section discusses experimental results on 8 benchmark datasets: MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky et al., 2009), SVHN (Netzer et al., 2019), EMNIST (Cohen et al., 2017), CIFAR100 (Krizhevsky et al., 2009), Tiny ImageNet (subset of the ILSVRC dataset containing 200 categories rather than the usual 1000 categories; Russakovsky et al., 2015), Food101 (Bossard et al., 2014) and HAM10000 (Tschandl et al., 2018) datasets. For fair comparison with other active learning algorithms, simple two layered CNN, referred to as 'S-CNN' (Chollet et al., 2015) is used for MNIST and four layered CNN, referred to as 'K-CNN' (Chollet et al., 2015) is used for CIFAR10, SVHN, EMNIST and CIFAR100. Additionally, Wide-ResNet (WRN-16-8; Zagoruyko & Komodakis, 2016) is used for CIFAR100, Tiny ImageNet, Food101 and HAM10000.

Figure 3–6 show magnified plots of test accuracy to accentuate the difference in performance among different methods: initial labeled sample sizes are not shown in the figures. Figures that include initial labeled sample size are presented in Appendix H.

5.1 Experimental Settings

Experimental settings regarding total number of epochs, data size and acquisition size are summarized in Table 1, and other details concerning the model, optimizer, batch size, learning rate and hyperparameters are presented in Appendix B.

Table 1: Experimental settings for comparing the performance on various datasets are summarized. Epochs is the maximum number of training epochs. Data size denotes the sizes of datasets for training / validation / test. Acquisition size denotes the number of samples for the initial model + number of samples acquired in each step (from the number of samples in the pool data) → Maximum number of samples acquired during training.

Dataset	Model	Epochs	Data size train / val / test	Acquisition size		
MNIST	S-CNN	50	55,000 / 5,000 / 10,000	20	+20 (2K)	→ 1,020
CIFAR10	K-CNN	150	45,000 / 5,000 / 10,000	200	+100 (2K)	→ 10,200
SVHN	K-CNN	150	68,257 / 5,000 / 26,032	200	+100 (2K)	→ 10,200
EMNIST	K-CNN	100	75,000 / 5,000 / 10,000	235	+150 (2K)	→ 15,235
CIFAR100	K-CNN	150	45,000 / 5,000 / 10,000	2,000	+500 (5K)	→ 27,000
CIFAR100	WRN-16-8	100	45,000 / 5,000 / 10,000	5,000	+2,000 (10K)	→ 25,000
Tiny ImageNet	WRN-16-8	200	90,000 / 10,000 / 10,000	10,000	+5,000 (20K)	→ 50,000
Food101	WRN-16-8	200	60,600 / 15,150 / 25,250	6,000	+3,000 (15K)	→ 30,000
HAM10000	WRN-16-8	100	7,015 / 1,500 / 1,500	500	+300 (3K)	→ 3,500

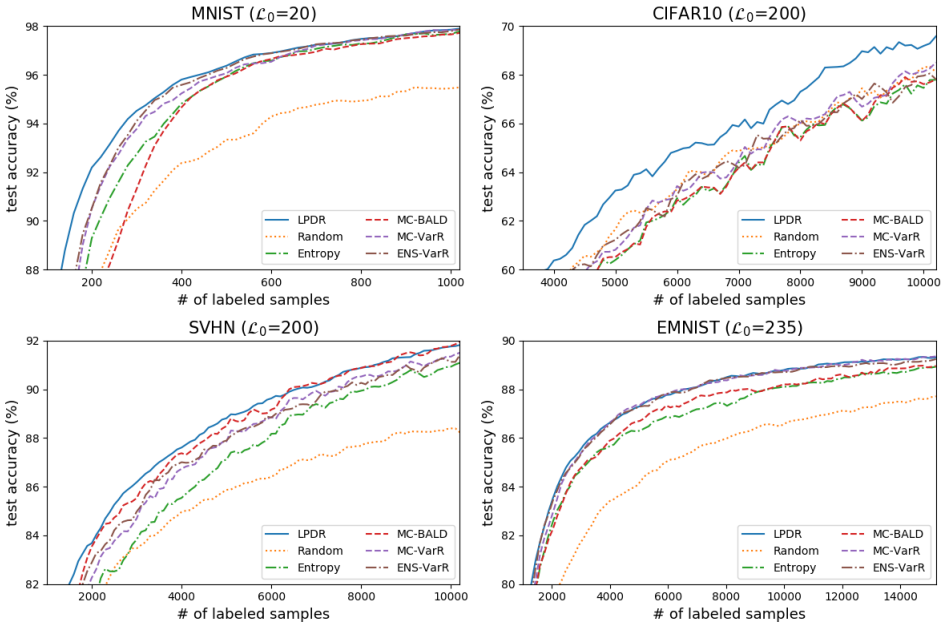

Figure 3: The performance comparison of LPDR with the uncertainty based active learning algorithms on MNIST, CIFAR10, SVHN and EMNIST datasets (Random: random sampling, Entropy: entropy based uncertainty sampling, MC-BALD: MC dropout sampling with BALD, MC-VarR: MC dropout sampling with variation ratio, ENS-VarR: ensemble network with variation ratio). Overall, LPDR consistently either performs best or comparable with all other algorithms regardless of dataset. The performance of all algorithms except LPDR tend to be data dependent.

5.2 Results for MNIST, CIFAR10, SVHN and EMNIST

A number of experiments are conducted to compare performance of LPDR with other high performing uncertainty based active learning algorithms on 8 datasets. Figure 3 shows the test accuracy with respect to the number of labeled samples on MNIST, CIFAR10, SVHN and EMNIST datasets. Each algorithm is denoted such as 'LPDR': the proposed algorithm, 'Random': random sampling, 'Entropy': entropy based uncertainty sampling, 'MC-BALD': MC dropout sampling using BALD, 'MC-VarR': MC dropout sampling using variation ratio (Ducoffe & Precioso, 2015) and 'ENS-VarR': ensemble method. Overall, LPDR either performs best or comparable with all other algorithms. Its performance is consistent regardless of the benchmark datasets. In the early step, LPDR

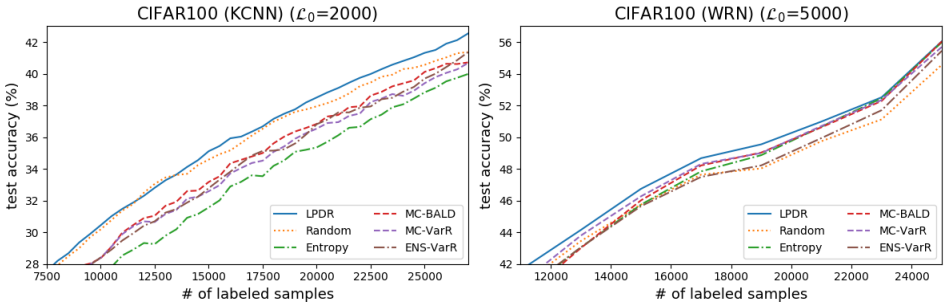

Figure 4: Performance comparison with respect to the network capacity on CIFAR100 dataset. The performances of all algorithms except LPDR are much worse than that of Random when using K-CNN, which has a relatively smaller network capacity than that of WRN-16-8. LPDR is able to perform consistently better than Random regardless of the network capacity.

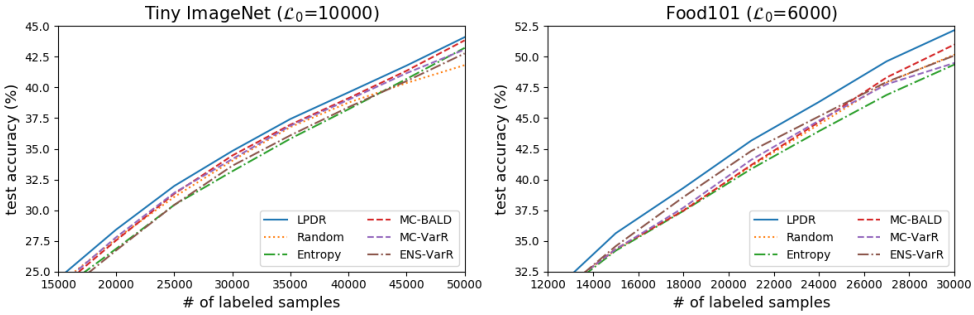

Figure 5: The performance comparison on Tiny ImageNet and Food101 datasets with WRN-16-8. LPDR outperforms all other algorithms in more difficult tasks.

significantly outperforms all other algorithms on MNIST and CIFAR10 datasets. Of all the algorithms compared, Entropy performed the worst. MC-BALD performed well only on SVHN dataset: it seems that the performance of BALD is highly dependent on the dataset. With the query size set to 1, LPDR outperforms BatchBALD on MNIST dataset (see Appendix I). Although MC-VarR and ENS-VarR are based on different sampling methods, both perform similarily-both outperforming all others on EMNIST dataset, while showing a significant drop in performance compared to LPDR on SVHN and CIFAR10 datasets. It is observed that the performances of other algorithms have a relatively strong data dependency compared to LPDR. On CIFAR10 dataset, the performances of MC-VarR and ENS-VarR are no better than that of Random, and Entropy and MC-BALD have lower performance than Random. These results can be attributed to the low network capacity compared to the data complexity. This issue will be discussed in next section.

5.3 Results for CIFAR100 with K-CNN and Wide-ResNet

In order to compare the performance of the algorithms with respect to the network capacity, experiments are conducted using networks of different capacity but on the same dataset. Figure 4 shows the results of test accuracy with respect to the number of labeled samples on CIFAR100 dataset with K-CNN and WRN-16-8. The left-hand figure is the results of using K-CNN, which has a relatively smaller network capacity than that of WRN-16-8. With the exception of LPDR, the performances of all algorithms are much worse than that of Random. The right-hand figure is the result of using WRN-16-8, which has a relatively larger network capacity. In contrast to the results for K-CNN, most algorithms outperform Random. With a large network capacity, the performance gap between LPDR and the other algorithms is reduced, but LPDR still outperforms others. LPDR is able to perform consistently better than Random regardless of the network capacity, and it seems to be particularly effective with low capacity networks.

5.4 RESULTS FOR TINY IMAGENET AND FOOD101

Experiments on a more difficult task are conducted. Figure 5 shows test accuracy with respect to the number of labeled samples on Tiny ImageNet and Food101 datasets with WRN-16-8. Tiny ImageNet and Food101 are considered to be more difficult than CIFAR100. Even on more difficult tasks, LPDR outperforms all other algorithms.

5.5 RESULTS FOR HAM10000

Additional experiments are conducted to compare the performance of the algorithms on imbalanced HAM10000 dataset with WRN-16-8. Figure 6 shows the results of the test accuracy with respect to the number of labeled samples. The LPDR outperforms all other algorithms compared. Figure 14 in Appendix J shows the results of AUC with respect to the number of labeled samples. The LPDR performs comparable with all other algorithms.

To sum up the comparing algorithms across all experimental settings and repetitions, rank and Dolan-More curves are presented in Appendix K. The LPDR consistently achieves top rank for all steps and significantly outperforms the other algorithms in all experimental settings.

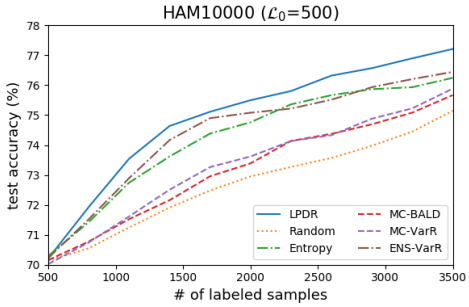

Figure 6: The performance comparison on HAM10000 dataset with WRN-16-8. LPDR outperforms all other algorithms on imbalanced dataset.

6 RELATED WORK

Other than uncertainty-based sampling framework (Culotta & McCallum, 2005; Scheffer et al., 2001; Mussmann & Liang, 2018; Lewis & Gale, 1994; Gal et al., 2017; Kirsch et al., 2019; Beluch et al., 2018) for active learning, decision-theoretic framework based methods such as expected model change (Settles et al., 2008) have certain relevance to the proposed LPDR as unlabeled samples nearer the decision boundary which LPDR is attempting to identify have larger gradients leading larger model change. Recently, adversarial approaches are proposed to discriminate labeled and unlabeled samples (Gissin & Shalev-Shwartz, 2019; Sinha et al., 2019; Zhang et al., 2020), and after performing adversarial learning, any unlabeled samples that is most confidently predicted as unlabeled is queried and used to retrain the network. Here adversarial learning is used to indirectly identify sample near the decision boundary.

7 CONCLUSION

This paper defines a theoretical distance of unlabeled sample to the decision boundary referred to as the least probable disagreement region (LPDR) containing the unlabeled sample for active learning. LPDR can be evaluated empirically with low computational load by making two assumptions regarding parameters of the hypothesis space, variation ratio and the LPDR. The two assumptions are empirically verified.

Experimental results on various datasets show that LPDR consistently outperforms all other high performing uncertainty based active learning algorithms and leads to state-of-the-art active learning performance on CIFAR10, CIFAR100, Tiny ImageNet, and Food101 datasets. In addition, LPDR is able to perform consistently better than random sampling regardless of the network capacity while all other algorithms compared fail to do so.

LPDR is simple enough to be applied to various classification tasks with deep networks: the implementation requires only sampling a subset of parameters (parameters in the last FC layer of the deep network). Additionally, LPDR is capable of quick and reliable performance in a variety of different settings with only a computational load that is not much higher than that of other uncertainty sampling methods. In conclusion, LPDR is an effective uncertainty based sampling algorithm in pool-based active learning.

REFERENCES

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.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

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.

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

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

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.

Elizabeth D Dolan and Jorge J Moré. Benchmarking optimization software with performance profiles. *Mathematical programming*, 91(2):201–213, 2002.

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.

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.

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.

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.

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.

Tom M Mitchell. Generalization as search. *Artificial intelligence*, 18(2):203–226, 1982.

Stephen Mussmann and Percy S Liang. Uncertainty sampling is preconditioned stochastic gradient descent on zero-one loss. In *Advances in Neural Information Processing Systems*, pp. 6955–6964, 2018.

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

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.

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

Ozan Sener and Silvio Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv*, 1708:1, 2017b.

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.

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.

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.

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.

APPENDIX

A PROOF OF THEOREM 1

Assume that $\|\hat{\boldsymbol{w}}_t\| = 1$ without the loss of generality. The predicted label of $\boldsymbol{x}$ by $\boldsymbol{w}'_n$ disagrees with that by $\hat{\boldsymbol{w}}_t$ if $\operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right)$, here, $\operatorname{sgn}(0) = 1$. Note that

$$\boldsymbol{x}^\mathsf{T}\boldsymbol{w}'_n = \boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t + \sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n \quad \text{where} \quad \mathbf{e}'_n = (Z_{n1}, \ldots, Z_{n|\boldsymbol{w}|})^\mathsf{T},$$

and Z_{nk}s are independent random variables from $\mathcal{N}(0,1)$. The event of $\{\operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right)\}$ is equal to that of $\{\operatorname{sgn}\left(\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right), \left|\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right| > \left|\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right|\}$. This induces that the probability of disagreement is

$$\begin{aligned}
\mathbb{P}\left[\operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\boldsymbol{w}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right)\right] &= \mathbb{P}\left[\operatorname{sgn}\left(\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right)\right] \\
&\quad \times\mathbb{P}\left[\left|\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right| > \left|\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right| \,\middle|\, \operatorname{sgn}\left(\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right) \neq \operatorname{sgn}\left(\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right)\right] \\
&= \frac{1}{2}\mathbb{P}\left[\left|\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n\right| > \left|\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right|\right] \\
&= \mathbb{P}\left[\sigma\|\boldsymbol{x}\|Z > \left|\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right|\right] \\
&= 1 - \Phi\left(\frac{a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}_t\right)}{\sigma}\right)
\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$ and Φ is the cumulative distribution function of the normal distribution. The second and third lines are from $\sigma\boldsymbol{x}^\mathsf{T}\mathbf{e}'_n \sim \mathcal{N}(0, \sigma^2\|\boldsymbol{x}\|^2)$. Note that $1 - \Phi\left(a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}_t\right)/\sigma\right) \leq \frac{1}{2}$ since

$$0 < a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}_t\right) = \frac{\left|\boldsymbol{x}^\mathsf{T}\hat{\boldsymbol{w}}_t\right|}{\|\boldsymbol{x}\|} \leq 1.$$

Then, $\forall\boldsymbol{x}$, $\hat{h}_t(\boldsymbol{x})$ becomes the modal class of $\boldsymbol{x}$ and the frequency of the modal class is defined as:

$$f_m^{(\boldsymbol{x})} = \sum_{n=1}^{N} \mathbb{I}\left[\hat{h}_t(\boldsymbol{x}) = h'_n(\boldsymbol{x})\right].$$

As $N \to \infty$, $\forall\boldsymbol{x}$, the variation ratio

$$1 - \frac{f_m^{(\boldsymbol{x})}}{N} \to 1 - \Phi\left(\frac{a\left(\boldsymbol{x}, \hat{\boldsymbol{w}}_t\right)}{\sigma}\right)$$

in probability since $\operatorname{Var}(f_m^{(\boldsymbol{x})}/N) \to 0$. By the smoothness of Φ,

$$1 - \frac{f_m^{(\boldsymbol{x})}}{N} \to 1 - \Phi(\infty) = 0 \quad \text{as} \quad \sigma^2 \to 0$$

and

$$1 - \frac{f_m^{(\boldsymbol{x})}}{N} \to 1 - \Phi(0) = \frac{1}{2} \quad \text{as} \quad \sigma^2 \to \infty.$$

This completes the proof. □

B EXPERIMENTAL SETTINGS

B.1 DATASETS

MNIST (LeCun et al., 1998) is a dataset of handwritten digits which has a training set of $60,000$ samples and a test set of $10,000$ 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 labeled subsets of the 80 million tiny images dataset which have a training set of $50,000$ samples and a test set of $10,000$ 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 image dataset which has a training set of $73,257$ samples and a test set of $26,032$ samples in 10 classes. Each sample is a color image and 32×32 in size.

EMNIST (Cohen et al., 2017) is a dataset of handwritten character digits which has a training set of $80,000$ samples and a test set of $10,000$ samples in 47 classes. Each sample is a black and white image and 28×28 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: a training set of $90,000$ samples and a test set of $10,000$ samples.

Food101 (Bossard et al., 2014) is a fine grained dataset which has a training set of $75,750$ samples and a test set of $25,250$ samples in 101 classes. Each sample is a color image and resized to 75×75.

HAM10000 (Tschandl et al., 2018) is a imbalanced 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: a training set of $8,515$ samples and a test set of $1,500$ samples.

All datasets are used without any preprocessing of images.

B.2 Settings

S-CNN, which is the Keras MNIST CNN implementation (Chollet et al., 2015), consists of [$3 \times 3 \times 32$ conv - $3 \times 3 \times 64$ conv - 2×2 maxpool - dropout (0.25) - 128 dense - dropout (0.5) - #class dense - softmax] layers. K-CNN, which is the Keras CIFAR CNN implementation (Chollet et al., 2015), consists of [two $3 \times 3 \times 32$ conv - 2×2 maxpool - dropout (0.25) - two $3 \times 3 \times 64$ conv - 2×2 maxpool - dropout (0.25) - 512 dense - dropout (0.5) - #class dense - softmax] layers. WRN-16-8 is a wide residual network that has 16 convolutional layers and a widening factor 8 (Zagoruyko & Komodakis, 2016). The optimizer, initial learning rate, learning rate schedule and batch size for each experimental setting are described in Table 2. He normal initialization is used for all models. All experiments are run for a fixed number of acquisition steps until a certain amount of training data is labeled. Results are averaged over 5 repetitions. For all datasets, the initial labeled samples for each repetition are randomly sampled according to the distribution of the training set. For MC dropout we use 100 forward passes, and ensemble consists of 5 networks of identical architecture but different random initialization and random batches. For LPDR, we set $\sigma_0 = 0.01$, $\beta = 1$, $N = 100$ and parameter sampling is applied to the last dense layer of each network.

Table 2: Settings for Training

Dataset	Model	Optimizer	Learning Rate	Learning Rate Schedule ×decay [epoch schedule]	Batch Size
MNIST	S-CNN	Adam	0.001	-	32
CIFAR10	K-CNN	Adam	0.0001	-	64
SVHN	K-CNN	Adam	0.0001	-	64
EMNIST	K-CNN	Adam	0.0001	-	64
CIFAR100	K-CNN	Adam	0.0001	-	64
CIFAR100	WRN-16-8	Nesterov	0.05	×0.2 [60, 80]	128
Tiny ImageNet	WRN-16-8	Nesterov	0.1	×0.2 [60, 120, 160]	128
Food101	WRN-16-8	Nesterov	0.1	×0.2 [60, 120, 160]	128
HAM10000	WRN-16-8	Nesterov	0.05	×0.2 [60, 80]	64

C Rank Correlation between LPDR and Variation Ratio

Figure 7 shows the relationship between the sample distance rank and the variation ratio rank for all experimental settings. For all results, there is strong negative correlation between the distance rank and the variation rank. The σ is selected to satisfy $\rho'_n = \rho^* = q/m$ at initial step.

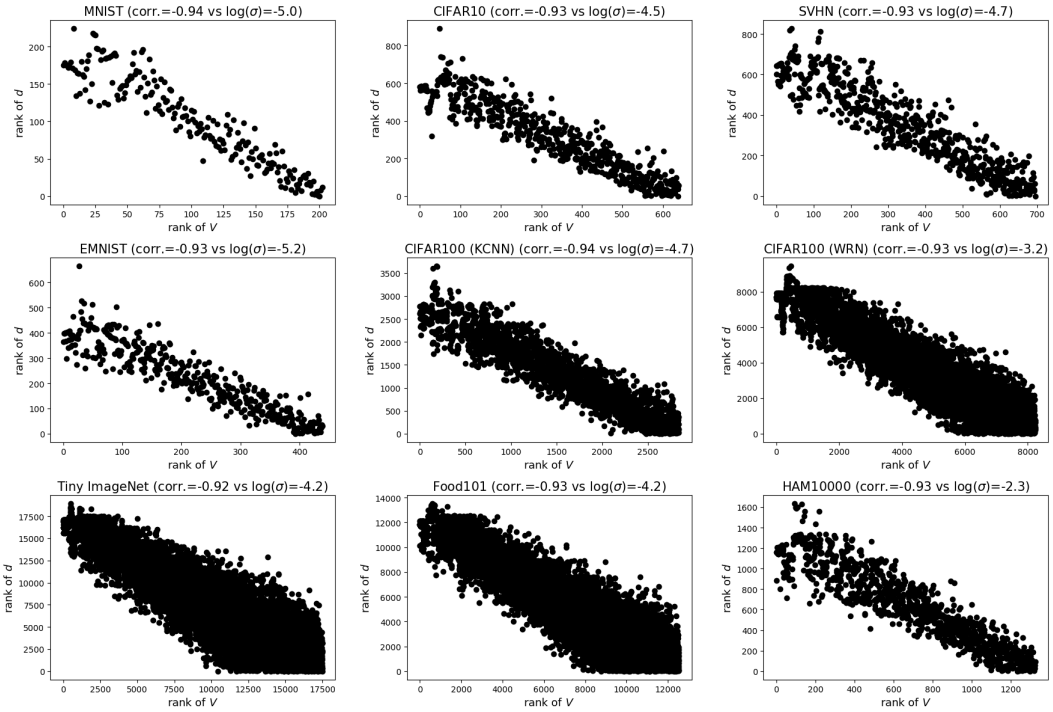

Figure 7: The relationship between distance rank and variation rank for all experimental settings. There is strong negative correlation between both for all cases.

D REGULATING ρ'_n BY THE VARIANCE OF SAMPLING

The left-hand side of Figure 8 shows the ρ'_n with respect to the active learning progress. For all experiments, LPDR reliably guides the ρ'_n to be $\rho^* = q/m$ (MNIST: 0.01, CIFAR10: 0.05, SVHN: 0.05, EMNIST: 0.075, CIFAR100 (KCNN): 0.1, CIFAR100 (WRN): 0.2, Tiny ImageNet: 0.25, Food101: 0.2 and HAM10000: 0.1) after the initial few steps. The right-hand side of Figure 8 shows $\log(\sigma)$ with respect to the active learning progress. For all experiments, the variance of sampling increases as the labeling proceeds. This is because larger variance is required to make the $\rho'_n = \rho^*$ since unlabeled samples move away from the learned decision boundary from labeled samples due to an increase in network confidence as the number of labeled samples increases.

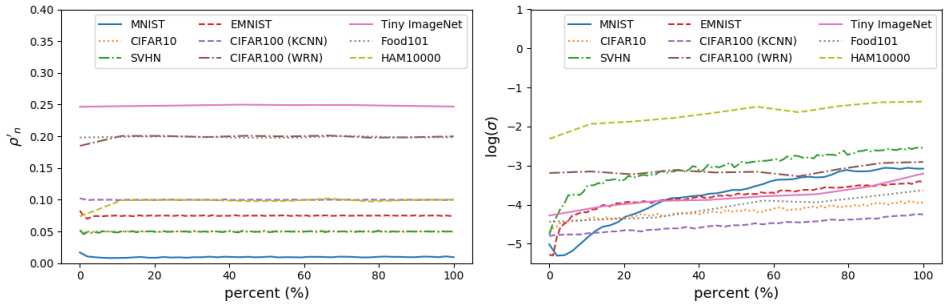

Figure 8: The ρ'_n and σ with respect to the labeling proceeds for all experimental settings. LPDR reliably guides the ρ'_n to be the target value by increasing the variance of sampling as the number of labeled samples increases.

E FINAL TEST ACCURACY VS TARGET ρ'_n

The Figure 9 shows the final test accuracy with respect to target ρ'_n on MNIST dataset. The results show that at around $\rho^*(= 0.02)$, it performs the best for $q = 20$ and $m = 1000$. In addition, the range of target ρ'_n, associated with the best performance, is wide ($0.01 \sim 0.1$); thus, LPDR is robust against the target ρ'_n in the wide range.

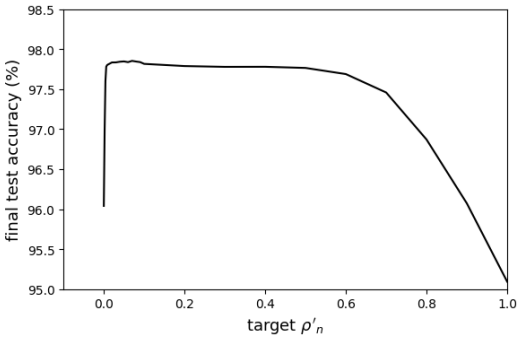

Figure 9: The final accuracy with respect to the target ρ'_n on MNIST dataset. LPDR performs best in a wide range of the target ρ'_n.

F ROBUSTNESS OF LPDR AGAINST HYPERPARAMETERS

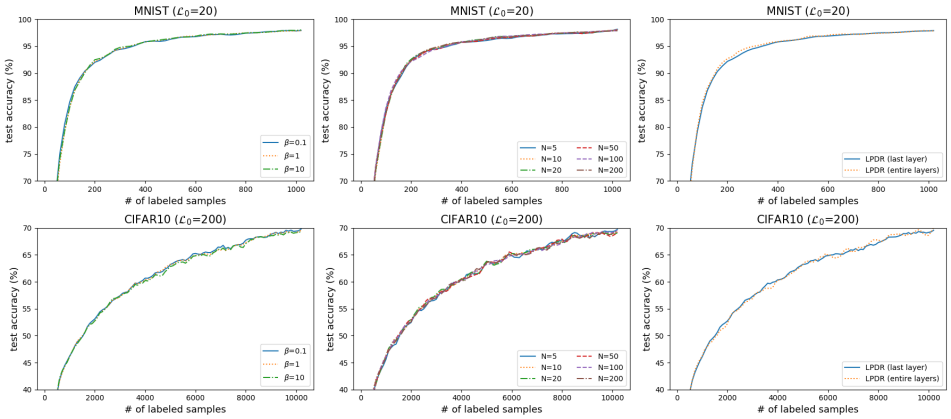

Figure 10: The performance comparison with respect to the hyperparameters of LPDR on MNIST and CIFAR10 datasets. LPDR is robust against β and N, and has no significant performance difference whether the sampling is applied to the parameters of last layer or all layers.

LPDR has four hyperparameters: 1) the initial variance of sampling σ_0; 2) the positive hyperparameter for regulating the variance of sampling β; 3) the number of sampled hypotheses N, and 4) the layer index of the network to which sampling is applied. The σ_0 has no significant effect on the performance of LPDR since σ is adaptively regulated based on the ρ'_n while sampling the sampled hypothesis. Thus, σ_0 is not examined in detail. Figure 10 shows the performance comparison with respect to the hyperparameters of LPDR on MNIST and CIFAR10 datasets. The left figures show that there is no significant difference in the performance of LPDR for various $\beta \in \{0.1, 1, 10\}$ on both datasets. The robustness of LPDR against β is based on the sufficient buffer for regulating σ since the range of target ρ'_n associated with the best performance is wide. The middle figures show that there is no significant difference in the performance of LPDR for various $\beta \in \{0.1, 1, 10\}$ on both datasets. The robustness of LPDR against N is based on the sufficient discrimination in the variation ratio for identifying q most informative unlabeled samples with a small number of sampled hypotheses by setting $\rho^* = q/m$. The right figures show that there is no significant difference in the performance of LPDR for the sampling to the parameters of last layer and to the parameters of all layers of the networks on both datasets.

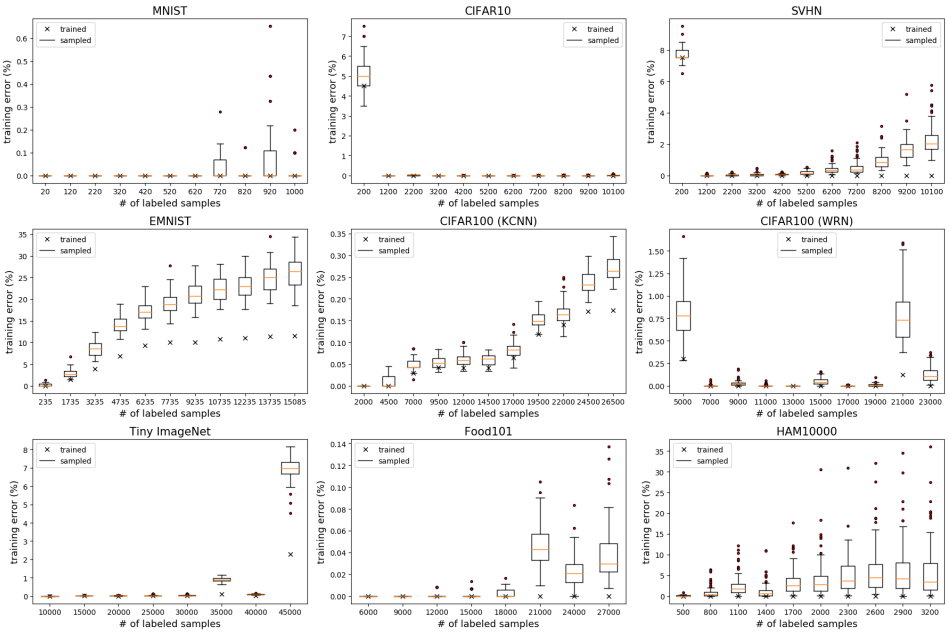

Figure 11: The empirical errors of the learned and the sampled hypotheses with respect to the acquisition step for all experimental settings. It is observed that the empirical error of the learned hypothesis or the sampled hypothesis is not zero.

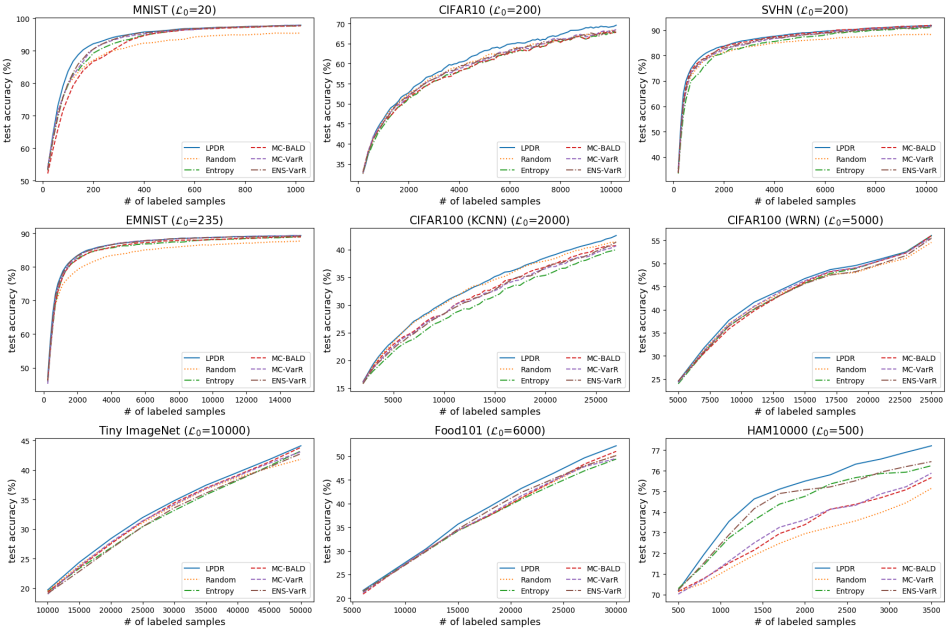

Figure 12: The test accuracy with respect to the number of labeled samples from initial to final step for all experimental settings.

G Empirical Errors of Learned and Sampled Hypotheses

Figure 11 shows the empirical error of the learned and the sampled hypotheses with respect to the acquisition step for all experimental settings. In many cases, the empirical error of the learned hypothesis becomes zero, thus it is placed in the version space, while the sampled hypothesis is often placed outside the version space, e.g., in SVHN dataset. Even in the cases of EMNIST and CIFAR100 with K-CNN datasets, as the number of labeled samples increases, the empirical error of

the learned and the sampled hypothesis increases. To address this situation, LPDR incorporates the weighted hypotheses based on the prediction error difference between the learned and the sampled hypotheses, and it works well empirically.

H PLOTS FOR TEST ACCURACY

Figure 12 shows the test accuracy with respect to the number of labeled samples from initial to final step for all experimental settings.

I LPDR VS MC-BATCHBALD

Figure 13 shows the performance comparison between LPDR and MC-BALD on MNIST dataset using S-CNN when the query size is 1 or 20. LPDR significantly outperforms MC-BatchBALD on MNIST dataset when $q = 1$ such that MC-BatchBALD is completely identical to MC-BALD. LPDR is also expected to outperform MC-BatchBALD even when $q > 1$: LPDR with $q > 1$ performs better than MC-BALD with $q = 1$ that MC-BatchBALD with $q > 1$ does not exceed (Kirsch et al., 2019).

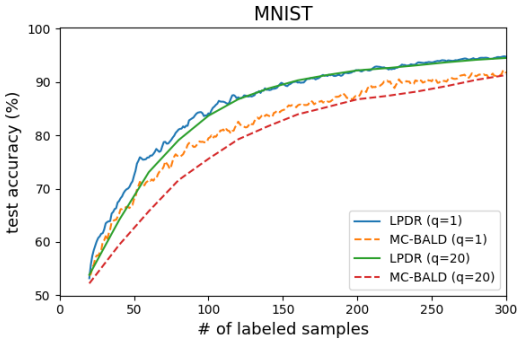

Figure 13: The comparison of performance between LPDR and MC-BALD on MNIST dataset where the query size is 1 or 20. The performance of BatchBALD with $q > 1$ does not exceed that of MC-BALD ($q = 1$) and LPDR ($q = 20$) outperforms MC-BALD ($q = 1$).

J AUC OF HAM10000 DATASET

On imbalanced dataset, the performance comparison is performed not only for accuracy but also for AUC. Figure 14 shows the results of AUC with respect to the number of labeled samples on HAM10000 dataset. LPDR performs comparable with Entropy or ENS-VarR performing better than other algorithms.

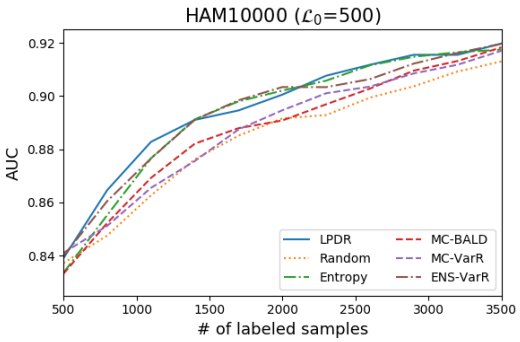

Figure 14: The comparison of AUC on HAM10000 dataset. LPDR performs comparable with the best performing algorithms.

K RANK AND DOLAN-MORE CURVES

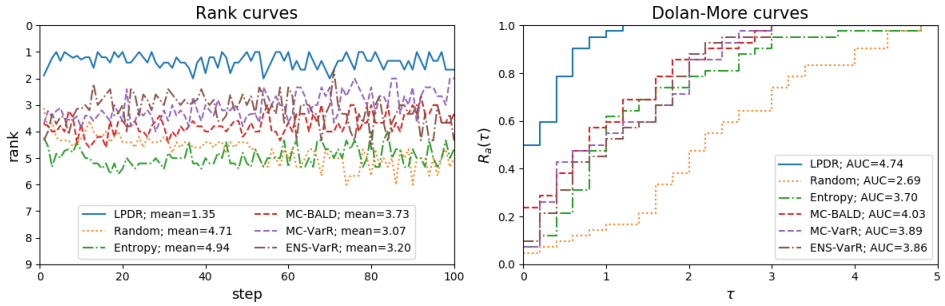

Figure 15: The rank and Dolan-More curves of the algorithms across all experimental settings and repetitions. The left figure shows rank curve which is the mean of ranks on all datasets at each step. LPDR consistently is top-ranked for all steps. The right figure shows the each algorithm's Dolan-More curves which present the fraction of problems in which the algorithm has the performance gap from the best competitor. LPDR maintains the highest value for all τ.

Table 3: The mean ($\pm$ standard deviation) of performance gap from the best competitor for all steps of each algorithm on each dataset. LPDR significantly outperforms the other algorithms on all datasets.

	MNIST	CIFAR10	SVHN	EMNIST	CIFAR100	CIFAR100-W	T. ImageNet	Food101	HAM10000
LPDR	**0.17±0.04**	**0.22±0.07**	**0.15±0.02**	**0.13±0.04**	**0.23±0.10**	**0.50±0.15**	**0.42±0.18**	**0.13±0.09**	**0.29±0.11**
Random	3.45±0.46	1.56±0.07	3.08±0.16	2.80±0.08	0.62±0.17	1.67±0.46	1.41±0.49	1.74±0.48	2.52±0.39
Entropy	1.20±0.27	2.28±0.19	1.92±0.15	0.95±0.13	3.19±0.34	1.19±0.28	1.82±0.36	2.17±0.03	0.98±0.40
MC-BALD	1.91±0.48	2.25±0.16	0.45±0.04	0.78±0.18	1.80±0.35	1.17±0.10	1.00±0.22	1.53±0.09	2.04±0.29
MC-VarR	0.86±0.30	1.76±0.46	1.06±0.11	0.21±0.09	2.20±0.44	0.92±0.27	1.06±0.37	1.72±0.01	1.92±0.30
ENS-VarR	0.65±0.34	1.74±0.18	0.97±0.06	0.22±0.04	2.13±0.28	1.57±0.34	1.85±0.26	1.31±0.31	0.81±0.29

Rank curves and Dolan-More curves are used to compare the performance of the algorithms across all experimental settings and repetitions. Figure 15 shows the rank and Dolan-More curves for all algorithms considered in the experiment. The rank curve of each algorithm in the left-hand figure represents the mean of ranks on all datasets at each steps of active learning. LPDR consistently is top-ranked for all steps.

The right-hand figure shows Dolan-More curves defined as follows (Dolan & Moré, 2002). Let acc_a^p be the final test accuracy of the a algorithm on the p problem. After defining the performance gap as $\Delta_a^p = \max_x(\mathrm{acc}_x^p) - \mathrm{acc}_a^p$, we can define Dolan-More curve $R_a(\cdot)$ as a function of the performance gap factor τ:

$$R_a(\tau) = \frac{\#(p : \Delta_a^p \leq \tau)}{n_p}$$

where n_p is the total number of evaluations for the problem p. Thus, $R_a(\tau)$ is the ratio of problems with performance gap between algorithm a and the best performing competitor not more than τ. Note that $R_a(0)$ is the ratio of problems on which algorithm a performs the best. LPDR has the highest value $R_a(0) = 43.3\%$, and LPDR maintains the highest $R_a(\tau)$ for all τ.

Table 3 presents the mean and the standard deviation of performance gap from the best competitor for all steps of each algorithm on each dataset. Consistent with all the results so far, LPDR significantly outperforms the other algorithms in all experimental settings.