On the Relationship between Data Efficiency and Error for Uncertainty Sampling

Stephen Mussmann [1] **Percy Liang** [1]

Abstract

While active learning offers potential cost savings, the actual data efficiency—the reduction in amount of labeled data needed to obtain the same error rate—observed in practice is mixed. This paper poses a basic question: when is active learning actually helpful? We provide an answer for logistic regression with the popular active learning algorithm, uncertainty sampling. Empirically, on 21 datasets from OpenML, we find a strong inverse correlation between data efficiency and the error rate of the final classifier. Theoretically, we show that for a variant of uncertainty sampling, the asymptotic data efficiency is within a constant factor of the inverse error rate of the limiting classifier.

1. Introduction

Active learning offers potential label cost savings by adaptively choosing the data points to label. Over the past two decades, a large number of active learning algorithms have been proposed (Seung et al., 1992; Lewis & Gale, 1994; Freund et al., 1997; Tong & Koller, 2001; Roy & McCallum, 2001; Brinker, 2003; Hoi et al., 2009). Much of the community's focus is on comparing the merits of different active learning algorithms (Schein & Ungar, 2007; Yang & Loog, 2016).

This paper is motivated by the observation that even for a *fixed* active learning algorithm, its effectiveness varies widely across datasets. Tong & Koller (2001) show a dataset where uncertainty sampling achieves 5x data efficiency, meaning that active learning achieves the same error rate as random sampling with one-fifth of the labeled data. For this same algorithm, different datasets yield a mixed bag of results: worse performance than random sampling (Yang & Loog, 2016), no gains (Schein & Ungar, 2007), gains of 2x (Tong & Koller, 2001), and gains of 3x (Brinker, 2003).

In what cases and to what extent is active learning superior to naive random sampling? This is an important question to address for active learning to be effectively used in practice. In this paper, we provide both empirical and theoretical answers for the case of logistic regression and uncertainty sampling, "the simplest and most commonly used" active learning algorithm in practice (Settles, 2010) and the best algorithm given in the benchmark experiments of Yang & Loog (2016).

Empirically, in Section 3, we study 21 binary classification datasets from OpenML. We found that the data efficiency for uncertainty sampling and inverse error achieved by training on the full dataset are correlated with a Pearson correlation of 0.79 and a Spearman rank correlation of 0.67.

Theoretically, in Section 4, we analyze a two-stage variant of uncertainty sampling, which first learns a rough classifier via random sampling and then samples near the decision boundary of that classifier. We show that the asymptotic data efficiency of this algorithm compared to random sampling is within a small constant factor of the inverse limiting error rate. The argument follows by comparing the Fisher information of the passive and active estimators, formalizing the intuition that in low error regimes, random sampling wastes many samples that the model is already confident about. Note that this result is different in kind than the $\log(1/\epsilon)$ versus $1/\epsilon$ rates often studied in statistical active learning theory (Balcan et al., 2009; Hanneke, 2014), which focuses on convergence rates as opposed to the dependence on error. Together, our empirical and theoretical results provide a strong link between the data efficiency and the limiting error rate.

2. Setup

Consider a binary classification problem where the goal is to learn a predictor f from input $x \in \mathbb{R}^d$ to output $y \in \{-1, +1\}$ that has low expected error (0-1 loss), $\text{Err}(f) = \Pr[f(x) \neq y]$ with respect to an underlying

[1] Stanford University, Stanford, CA. Correspondence to: Stephen Mussmann <mussmann@stanford.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

data distribution. In pool-based active learning, we start with a set of unlabeled input points $\mathcal{X}_\text{U}$. An active learning algorithm queries a point $x \in \mathcal{X}_\text{U}$, receives its label y, and updates the model based on (x, y). A passive learning algorithm (random sampling) simply samples points from $\mathcal{X}_\text{U}$ uniformly randomly without replacement, queries their labels, and trains a model on this data.

2.1. Logistic Regression

In this work, we focus on logistic regression, where $p_w(y \mid x) = \sigma(yx \cdot w)$, w is a weight vector, and $\sigma(z) = \frac{1}{1+\exp(-z)}$ is the logistic function. A weight vector w characterizes a predictor $f_w(x) = \operatorname{sgn}(x \cdot w)$. Given a set of labeled data points D (gathered either passively or actively), the maximum likelihood estimate is $\hat{w} = \operatorname{argmin}_w \sum_{(x,y)\in D} - \log p_w(y \mid x)$. Define the limiting parameters as the analogous quantity on the population: $w^* = \operatorname{argmin}_w \mathbb{E}[-\log p_w(y \mid x)]$. A central quantity in this work is the *limiting error*, denoted $\text{Err} = \text{Err}(f_{w^*})$. Note that we are interested in 0-1 loss (as captured by Err), though the estimator w^* minimizes the logistic loss.

2.2. Uncertainty Sampling

In this work, we focus on "the simplest and most commonly used query framework" (Settles, 2010), uncertainty sampling (Lewis & Gale, 1994). This is closely related to margin-based active learning in the theoretical literature (Balcan et al., 2007).

Uncertainty sampling first samples n_seed data points randomly from $\mathcal{X}_\text{U}$, labels them, and uses that to train an initial model. For each of the next $n - n_\text{seed}$ iterations, it chooses an data point from $\mathcal{X}_\text{U}$ that the current model is most uncertain about (i.e., closest to the decision boundary), queries its label, and retrains the model using all labeled data points collected so far. See Algorithm 1 for the pseudocode (note we will change this slightly for the theoretical analysis).

2.3. Data Efficiency

Let $\hat{w}_\text{passive}$ and $\hat{w}_\text{active}$ be the two estimators obtained by performing passive learning (random sampling) and active learning (uncertainty sampling), respectively. To compare these two estimators, we use *data efficiency* (also known as statistical relative efficiency (van der Vaart, 1998) or sample complexity ratio), which is the reduction in number of labeled points that active learning requires to achieve error ϵ compared to random sampling.

More precisely, consider the number of samples for each estimator to reach error ϵ:

$$n_\text{active}(\epsilon) \overset{\text{def}}{=} \max\{n : \mathbb{E}[\text{Err}(\hat{w}_\text{active})] \geq \epsilon\}, \tag{1}$$

$$n_\text{passive}(\epsilon) \overset{\text{def}}{=} \max\{n : \mathbb{E}[\text{Err}(\hat{w}_\text{passive})] \geq \epsilon\}, \tag{2}$$

where the expectation is with respect to the unlabeled pool, the labels, and any randomness from the algorithm. Then the data efficiency is defined as the ratio:

$$\text{DE}(\epsilon) \overset{\text{def}}{=} \frac{n_{passive}(\epsilon)}{n_{active}(\epsilon)}. \tag{3}$$

2.4. Data Efficiency Dependence on Dataset

The data efficiency depends on properties of the underlying data distribution. In the experiments (Section 3), we illustrate this dependence on show a variety of real-world datasets. As a simple illustration, we show this phenomenon on a simple synthetic data distribution. Suppose data points are sampled according to

$$y \sim \text{Uniform}(\{-1, 1\}), \quad x \sim \mathcal{N}(y\mu e_1, I), \tag{4}$$

where $e_1 = [1, 0, \dots]$. This distribution over (x, y) is the standard Gaussian Naive Bayes model with means $-\mu e_1$ and μe_1 and covariance I. See Figure 1 for the learning curves when $\mu = 0.8$ and Figure 2 for when $\mu = 2.3$. We note that the data efficiency doesn't even reach 1.1 when $\mu = 0.8$, and the curves get closer with more data. On the other hand, when $\mu = 2.3$, the data efficiency exceeds 5 and increases dramatically. This illustrates the wildly different gains of active learning, depending on the dataset. In particular, the data efficiency is higher for the less noisy dataset, as the thesis of this work predicts.

Algorithm 1 Uncertainty Sampling

```
Input: Probabilistic model $p_w(y|x)$, unlabeled $X_\text{U}$, $n_\text{seed}$
Randomly sample $n_\text{seed}$ points without replacement from $\mathcal{X}_\text{U}$ and call them $\mathcal{X}_\text{seed}$.
$\mathcal{X}_\text{U} = \mathcal{X}_\text{U} \setminus \mathcal{X}_\text{seed}$
$\mathcal{D} = \emptyset$
for each $x$ in $\mathcal{X}_\text{seed}$ do
  Query $x$ to get label $y$
  $\mathcal{D} = \mathcal{D} \cup \{(x, y)\}$
end for
for $n - n_\text{seed}$ iterations do
  $\hat{w} = \operatorname{argmin}_w \sum_{(x,y)\in\mathcal{D}} - \log p_w(y|x)$
  Choose $x = \operatorname{argmin}_{x\in\mathcal{X}_\text{U}} |P_{\hat{w}}(y|x) - \frac{1}{2}|$
  Query $x$ to get label $y$
  $\mathcal{X}_\text{U} = \mathcal{X}_\text{U} \setminus \{x\}$
  $\mathcal{D} = \mathcal{D} \cup \{(x, y)\}$
end for
$\hat{w} = \operatorname{argmin}_w \sum_{(x,y)\in\mathcal{D}} - \log p_w(y|x)$ and return $\hat{w}$
```

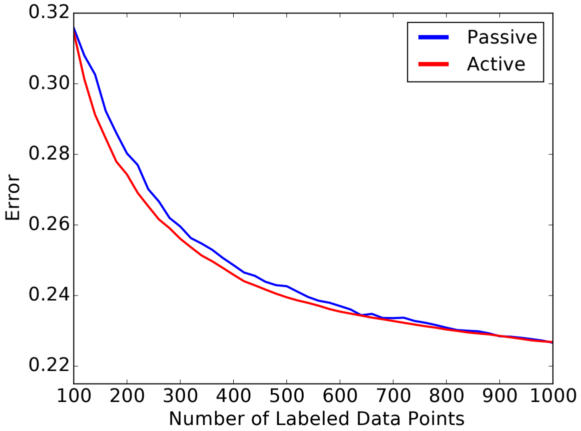

Figure 1. Active learning yields meager gains when the clusters are closer together ($\mu = 0.8$). The data efficiency is about 1x to get to 23% error; both algorithms require approximately the same amount of data to achieve that error.

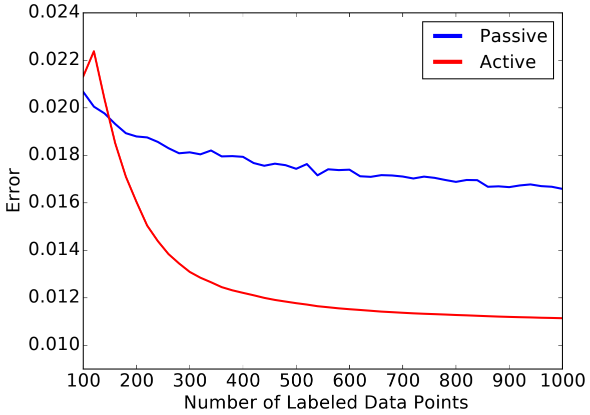

Figure 2. Active learning yields spectacular gains when the clusters are farther apart ($\mu = 2.3$). The data efficiency is about 5x to get to 16% error; passive learning requires about 1000 data points to achieve that error, while active learning only requires about 200.

3. Experiments

3.1. Datasets

We wish to study the data efficiency of active learning versus passive learning across a comprehensive set of datasets which are "typical" of real-world settings. Capturing a representative set of datasets is challenging, and we wanted to be as objective and transparent about the process as possible, so we detail the dataset selection process below.

We curated a set of datasets systematically from OpenML, avoiding synthetic or degenerate cases. In August 2017, we downloaded all 7968 datasets. We removed datasets with missing features or over 1 million data points. We wanted a large unlabeled pool (relative to the number of features) so we kept datasets where the number of features was less than 100 and the number of data points was at least 10,000. In our experiments, we allow each algorithm to query the label of $n = 1000$ points, so this filtering step ensures that $d \leq n/10$ and $n_{\text{pool}} \geq 10n$. We remark that more than 98% of the datasets were filtered out because they were too small (had fewer than 10,000 points). 138 datasets remained.

We further removed datasets that were synthetic, had unclear descriptions, or were duplicates. We also removed non-classification datasets. For multiclass datasets, we processed them to binary classification by predicting majority class versus the rest. Of the 138 datasets, 36 survived.

We ran standard logistic regression on training splits of these datasets. In 11 cases, logistic regression was less than 1% better than the classifier that always predicts the majority class. Since logistic regression was not meaningful for these datasets, we removed them, resulting in 25 datasets. On one of these datasets, logistic regression achieved 0% error with fewer than 40 data points. On another dataset, the performance of random sampling became *worse* as the number of labels increased. On two datasets, active learning achieved at least 1% error lower than the error with the *full training set*, a phenomenon that Schohn & Cohn (2000) calls "less is more"; this is beyond the scope of this work. We removed these four cases, resulting a total of 21 datasets.

The final 21 datasets has a large amount of variability, from healthcare, game playing, control, ecology, economics, computer vision, security, and physics.

3.2. Methodology

We used a random sampling seed of size $n_{\text{seed}} = 100$ and plotted the learning curves up until a labeling budget of $n = 1000$. We calculated the data efficiency at the lower of the errors achieved with the $n = 1000$ budget by active and passive learning. As a proxy for the limiting error, we use the error on the test set obtained by a classifer trained on the full training set.

3.3. Results

Figure 3 plots the relationship between data efficiency and the inverse error across all datasets. To remove outliers, we capped the inverse error at 50; this truncated the inverse error of three datasets which had inverse error of 190, 3200, and 27000 which corresponds to errors less than around 0.5%. The correlation (R^2 of the best linear fit) is 0.789. Further, the data efficiency and the inverse error have a Spearman rank correlation of 0.669.

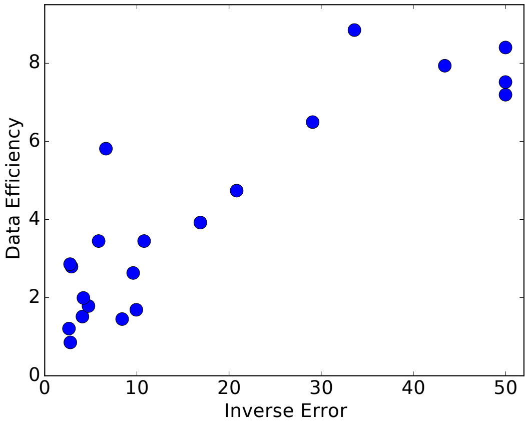

Figure 3. Scatterplot of the data efficiency of uncertainty sampling versus the inverse error using all the data. Line of best fit has 0.789 R^2, also known as the Pearson Correlation.

In summary, we note that data efficiency is closely tied to the inverse error. In particular, when the error is below 10%, the data efficiency is at least 3x and can be much higher.

4. Theoretical Analysis

In this section, we provide theoretical insight into the inverse relationship between data efficiency and limiting error. For tractability, we study the asymptotic behavior as the number of labels n tends to infinity.

Let $p(x, y)$ be the underlying data distribution over $\mathbb{R}^d \times \{-1, 1\}$. For uncertainty sampling, there are three data quantities: n_{seed}, the number of seed data points; n, the amount of labeled data (the budget); and n_{pool}, the number of unlabeled points in the pool. We will assume that n_{seed} and n_{pool} are functions of n, and we let n go to infinity. In particular, we wish to bound the value of $\lim_{\epsilon \to \text{Err}} \text{DE}(\epsilon)$ where Err is the limiting error defined in Section 2.1. Bounding $\text{DE}(\epsilon)$ for small ϵ is closely related to the statistical asymptotic relative efficiency (van der Vaart, 1998). We use data efficiency as it applies for finite n.

The asymptotic data efficiency only makes sense if the random sampling and uncertainty sampling both converge to the same error. Otherwise, the asymptotic data efficiency would either be 0 or ∞. While bias in active learning is an important topic of study (Liu et al., 2015), it is beyond the scope of this work. We will make an assumption that ensures this is satisfied if the model is well-specified in some small slab around the decision boundary.

4.1. Two-stage Variant of Uncertainty Sampling

Because of the complicated coupling between uncertainty sampling draws and updates, we analyze a two-stage variant: we gather an initial seed set using random sampling from the unlabeled dataset, and then gather the points closest to the decision boundary learned from the seed data. This two-stage approach is similar to other active learning work (Chaudhuri et al., 2015).

Thus, we only update the parameters twice: after the seed round we train on the seed data, and after we have collected all the data, we train on the data that was collected after the seed data. We do not update the parameters between draws closest to the decision boundary.

Also, instead of always choosing the point closest to decision boundary without replacement during the uncertainty sampling phase, with $\alpha > 0$ probability we randomly sample from the unlabeled pool and with $1 - \alpha$ probability we choose the point closest to the decision boundary. The random sampling proportion α ensures that the empirical data covariance is non-singular for uncertainty sampling.

4.2. Sketch of Main Result

Under assumptions that will be described later, our main result is that there exists some ϵ_0 such that for any $\text{Err} < \epsilon < \epsilon_0$,

$$DE(\epsilon) > \frac{s}{4\text{Err}}, \tag{5}$$

where s is a constant bounding a ratio of conditional covariances in the directions orthogonal to w^*. In particular, if the pdf factorizes into two marginal distributions (decomposition of x into two independent components), one along the direction of w^* and one in the directions orthogonal to w^*, then the conditional covariances orthogonal to w^* are equal, and $s = 1$. If the distribution is additionally symmetric across the decision boundary, we obtain

$$\frac{1}{4\text{Err}} < DE(\epsilon) < \frac{1}{2\text{Err}}. \tag{6}$$

We now give a rough proof sketch. The core idea is to compare the Fisher information of active and passive learning, similar to other work in the literature (Sourati et al., 2017). It is known that the Fisher information matrix for logistic regression is

$$\mathcal{I} = \mathbb{E}[\sigma(1 - \sigma)xx^\top], \tag{7}$$

where $\sigma = \sigma(yx \cdot w^*)$. Note that σ only depends on the part of x parallel to w^*. If the data decomposes into two independent components as mentioned above, then

$$\mathcal{I}_{\text{passive}} = \mathbb{E}[\sigma(1 - \sigma)]\mathbb{E}[xx^\top] \tag{8}$$

if we ignore the dimension of the Fisher information along w^* which doesn't end up mattering (it only changes the magnitude of w^* which is independent of the 0-1 loss). Additionally, since uncertainty sampling samples at the decision boundary where $w \cdot x^* = 0$, we have $\sigma = \frac{1}{2}$ and thus active learning achieves:

$$\mathcal{I}_{\text{active}} = \frac{1}{4}\mathbb{E}[xx^\top]. \tag{9}$$

The Fisher information determines the asymptotic rate of convergence of the parameters:

$$\sqrt{n}(w_n - w^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}^{-1}). \tag{10}$$

Intuitively, this convergence rate is monotonic with $\mathcal{I}^{-1}/n$ which means the ratio (abuse of notation, but true for any linear function of the inverse Fisher information) of the inverse Fisher information matrices gives the asymptotic relative rate,

$$\text{DE}(\epsilon) \approx \frac{\mathcal{I}_{\text{passive}}^{-1}}{\mathcal{I}_{\text{active}}^{-1}} \approx \frac{1/4}{\mathbb{E}[\sigma(1-\sigma)]}. \tag{11}$$

If the optimal logistic model is *calibrated*, meaning the model's predicted probabilities are on average correct, then

$$\frac{\text{Err}}{2} \le \mathbb{E}[\sigma(1-\sigma)] \le \text{Err}. \tag{12}$$

Putting these together, we get:

$$\frac{1}{4\text{Err}} \lessapprox DE(\epsilon) \lessapprox \frac{1}{2\text{Err}}. \tag{13}$$

Having given the rough intuition, we now go through the arguments more formally.

4.3. Notation

Let w^* be the limiting parameters, Let w_0 be the weights after the seed round for active learning, and w_n be the weights at the end of learning with n labels.

We include a bias term for logistic regression by inserting a coordinate at the beginning of x that is always 1. Thus, $x_0 = 1$ and w_0^* is the bias term of the optimal parameters. As a simplification of notation, the pdf $p(x)$ is only a function of the non-bias coordinates (otherwise, such a pdf wouldn't exist).

Since logistic regression is invariant to translations (we can appropriately change the bias) and rotations (we can rotate the non-bias weights), without loss of generality, we will assume that $w^* = \|w^*\|e_1$, that the bias term is 0, and that the data is mean 0 for all directions orthogonal to w^*, ($\mathbb{E}[x_{2:}] = 0$).

4.4. Assumptions

We have four types of assumptions: assumptions on the values of n_{seed} and n_{pool}, assumptions on the distribution of x, assumptions on the distribution of y, and non-degeneracy assumptions. As an example, all these assumptions are satisfied if $n_{\text{seed}} = \sqrt{n}$, $n_{\text{pool}} = n\sqrt{n}$, x is a mixture of truncated, mollified Gaussians, and y is well-specified for non-zero weights.

4.4.1. ASSUMPTIONS RELATING $n_{\text{SEED}}, n, n_{\text{POOL}}$

Recall that n_{seed} is the number of labels for the seed round, n is the labeling budget, and n_{pool} is the number of unlabeled data points.

Assumption 1 (Data Pool Size)**.** $n_{pool} = \omega(n)$.

Assumption 2 (Seed Size)**.** $n_{seed} = \Omega(n^\rho)$ *for some* $\rho > 0$ *and* $n_{seed} = o(n)$.

We need the size n_{pool} of the unlabeled pool has to be large enough so that uncertainty sampling can select points close to the decision boundary. We require that the seed for uncertainty sampling is large enough to make the decision boundary after the seed round converge to the true decision boundary, and we require that it is small enough so that it doesn't detract from the advantages of uncertainty sampling.

4.4.2. ASSUMPTION ON x DISTRIBUTION

We assume that the distribution on x has a pdf ("continuous distribution"), and the following two conditions hold:

Assumption 3 (Bounded Support)**.**

$$\exists B > 0 : \Pr[\|x\| > B] = 0 \tag{14}$$

Assumption 4 (Lipschitz)**.** *The pdfs and conditional pdfs* $p(x), p(x|w \cdot x = b), p(x|w_1 \cdot x_1 = b_1, w_2 \cdot x_2 = b_2)$ *are all Lipschitz.*

4.4.3. ASSUMPTIONS ON x, y DISTRIBUTION

These next three assumptions (Assumptions 5–7) are implied if the logistic regression model is well-specified ($\Pr[y|x] = \sigma(yx \cdot w^*)$), but they are strictly weaker. If the reader is willing to assume well-specification, this section can be skipped.

Assumption 5 (Local Expected Loss is Zero)**.** *There exists* λ *such that for* $\|w - w^*\| \le \lambda$,

$$\mathbb{E}_{w \cdot x = 0}[\nabla_w(-\log p_{w^*}(x, y))] = 0 \tag{15}$$

Assumption 5 is satisfied if model is well-specified in any thin slab around the decision boundary defined by w^*. We need this assumption to conclude that our two-stage uncertainty sampling algorithm converges to w^*.

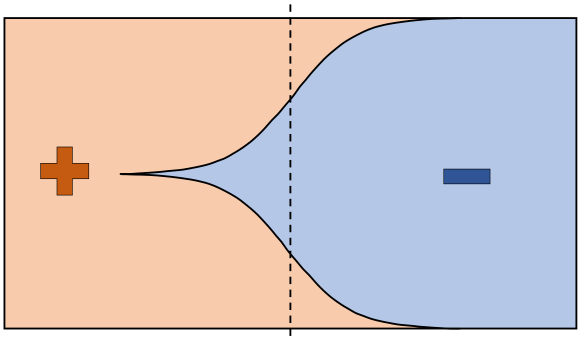

Figure 4. Example of distribution with deterministic labels which is calibrated but not well-specified for logistic regression.

Assumption 6 (Conditions on Zero-One Loss)**.** *Let $Z(w) = \Pr_{x,y}[yx \cdot w < 0]$ be the zero-one loss of the classifier defined by the weights w. Then,*

- *Z is twice-differentiable at w^*,*
- *Z has a local optimum at w^*, and*
- *$\nabla^2 Z(w^*) \neq 0$.*

In order to conclude that convergence to the optimal parameters implies convergence in error, we need Assumption 6. The strongest requirement is the local optimum part. The twice differentiable condition is a regularity condition and the Hessian condition is generically satisfied.

Assumption 7 (Calibration)**.**

$$\Pr[y|w^* \cdot x = a] = \sigma(ya) \tag{16}$$

We say a model is *calibrated* if the probability of a class, conditioned on the model predicting probability p, is p. Assumption 7 amounts to assuming that the logistic model with the optimal parameters w^* is calibrated. Note that this is significantly weaker than assuming that the model is well-specified ($\Pr[y|x] = \sigma(yx \cdot w^*)$). For example, the data distribution in Figure 4 is calibrated but not well-specified.

These three assumptions all hold if the logistic distribution is well-specified, meaning $\Pr[y|x] = \sigma(yx \cdot w^*)$.

4.4.4. Non-degeneracy

Define $p_0 = \int_{w^* \cdot x = 0} p(x)$ as the marginal probability *density* of selecting a point at the decision boundary. More precisely, p_0 is the probability density $p(x)$ integrated over the $d - 1$ dimensional hyperplane manifold defined by $w^* \cdot x = 0$. Equivalently, p_0 is the probability density of the random variable $\frac{w^*}{\|w^*\|} \cdot x$ at 0.

Assumption 8 (Non-degeneracies)**.**

$$p_0 \neq 0, \quad \|w^*\| \neq 0, \quad \mathit{Err} \neq 0, \quad \det(\mathbb{E}[xx^\top]) \neq 0 \tag{17}$$

Let us interpret these four conditions. We assume that the probability *density* at the decision boundary is non-zero, $p_0 \neq 0$, otherwise uncertainty sampling will not select points close to the decision boundary (note this is not an assumption about the probability *mass*). We assume that $\|w^*\| \neq 0$, meaning that the classifier is not degenerate, with all points on the decision boundary. We assume $\mathrm{Err} \neq 0$, meaning the logistic parameters do not achieve 0% error. Finally, we assume $\det(\mathbb{E}[xx^\top]) \neq 0$, meaning that the data covariance is non-singular, or equivalently, that the parameters are identifiable.

4.5. Proofs

We will first prove a condition on the convergence rate of the error based on a quantity Σ closely related to the Fisher Information. However, we can't rely on the usual Fisher information analysis, which does not connect to the zero-one loss, but rather to the asymptotic normality of the parameters. Thus, our conditions for this key lemma are slightly stronger than the asymptotic normality result of Fisher Information.

4.5.1. Rates Lemma in Terms of Σ

The logistic loss (negative log-likelihood) for a single data point under logistic regression is

$$\ell_w(x, y) = \log(1 + \exp(-w \cdot yx)). \tag{18}$$

Further, the gradient and Hessian are,

$$\nabla \ell_w(x, y) = -yx\sigma(-w \cdot yx) \tag{19}$$
$$\nabla^2 \ell_w(x, y) = \sigma(w \cdot yx)\sigma(-w \cdot yx)xx^\top \tag{20}$$

Note that $\sigma(-x) = 1 - \sigma(x)$.

Following the Fisher Information asymptotic normality analysis, note that

$$\sqrt{n}(w_n - w^*) = A_n^{-1} b_n, \tag{21}$$

where

$$A_n = \frac{1}{n} \sum_i \nabla^2 \ell_{w'}(x_i, y_i), \tag{22}$$

$$b_n = \frac{1}{\sqrt{n}} \sum_i \nabla \ell_{w^*}(x_i, y_i), \tag{23}$$

with $\|w' - w^*\| \leq \|w_n - w^*\|$. This is justified by Taylor's theorem since the logistic loss is smooth.

From these, we can define the key quantity Σ that is equivalent to the inverse Fisher Information under stronger conditions.

Definition 4.1. *If $A_n \xrightarrow{P} A$ (non-singular and symmetric) and $\mathbb{E}[b_n b_n^T] \to B$ exists, then define*

$$\Sigma = A^{-1} B A^{-1}. \tag{24}$$

This quantity is important because of the following lemma, which translates comparisons in the asymptotic variances to comparisons of data efficiency. Recall that without loss of generality, we let $w^* = \|w^*\| e_1$. Define A_{-1} as the matrix A without the first row and column.

Lemma 4.1. *If we have two estimators with asymptotic variances Σ_a and Σ_b, and for any $\epsilon > 0$ and both estimators, $n \Pr[\|A_n - A\| \geq \epsilon] \to 0$ and $n \Pr[\|w_n - w^*\| \geq \epsilon] \to 0$, then*

$$\Sigma_{a,-1} \succ c \Sigma_{b,-1} \tag{25}$$

implies that for some ϵ_0 and any Err $< \epsilon < \epsilon_0$,

$$n_a(\epsilon) \geq c n_b(\epsilon). \tag{26}$$

The proof is in the appendix. This lemma only requires Assumption 6, the condition on Z at w^*, and is possibly of independent interest.

Note that with the bias term, our weight vector is $d + 1$ dimensional, so Σ is a square $d+1$ dimensional matrix. However, without the first row and column, Σ_{-1} is a square d dimensional matrix. The fact that the rates depend on Σ_{-1} instead of Σ is necessary for our results. Intuitively, the first coordinate (in direction of w^*) has slow convergence for uncertainty sampling since we are selecting points near the decision boundary which have small projection onto w^* and thus we gain little information about the dependence of y on x_1. However, because our analysis is in terms of the convergence of the 0-1 error rather than convergence of the parameters, the above lemma doesn't depend on the convergence rate of the first coordinate.

From this lemma, it follows that if

$$c_1 \Sigma_{\text{active},-1} \prec \Sigma_{\text{passive},-1} \prec c_2 \Sigma_{\text{active},-1}, \tag{27}$$

then for sufficiently small error,

$$c_1 \leq DE(\epsilon) \leq c_2. \tag{28}$$

4.5.2. Specific Calculations for Algorithms

In proving the later results, it's useful to first establish the consistency of our algorithms. Assumption 5 is used here.

Lemma 4.2. *Both two-stage uncertainty sampling and random sampling converge to w^*.*

Next, we need our two algorithms satisfy the conditions of Lemma 4.1.

Lemma 4.3. *For our active and passive learning algorithms, for any $\epsilon > 0$, $n \Pr[\|A_n - A\| \geq \epsilon] \to 0$ and $n \Pr[\|w_n - w^*\| \geq \epsilon] \to 0$.*

Now, we are ready for the computation of Σ (Definition 4.1), the quantity closely related to the inverse Fisher Information.

Lemma 4.4.

$$\Sigma_{\textit{passive}} = \mathbb{E}[\sigma(1-\sigma) x x^\top]^{-1} \tag{29}$$

The proof is in the appendix. The proof relies on calibration, Assumption 7, to ensure that $\mathbb{E}[\nabla^2 \ell_w(x, y)] = \mathrm{Cov}(\nabla \ell_w(x, y))$, which is always true for well-specified models.

This lemma gives Σ as exactly the inverse Fisher information that was mentioned earlier. It is the expected value of $\nabla^2 \ell_{w^*}(x, y) = \sigma(1-\sigma) x x^\top$.

Lemma 4.5.

$$\Sigma_{\textit{active}} = \tag{30}$$

$$\left((1-\alpha)\mathbb{E}_{x_1=0}[\sigma(1-\sigma)xx^\top] + \alpha \mathbb{E}[\sigma(1-\sigma)xx^\top]\right)^{-1}$$

The proof is in the appendix. The proof relies on the assumptions of bounded support and Lipshitz pdf, Assumptions 3 and 4.

Because we randomly sample for α proportion, a factor of α times Σ_{passive} shows up. Additionally, we get a $1-\alpha$ factor for the expected value of $\nabla^2 \ell_{w^*}(x, y) = \sigma(1-\sigma)xx^\top$ at the decision boundary. We will almost surely never sample exactly at the decision boundary, but as $n \to \infty$, the seed round weights $w_0 \to w^*$ and $n_{\text{pool}}/n \to \infty$, we sample closer and closer to the decision boundary.

4.5.3. Results

Here, we define s that quantifies how much the covariance at the decision boundary differs from the covariance for the rest of the distribution, which is a key dependency of our most general theorem. Denote x_{-1} as the vector x without the first index. Recall that without loss of generality, $w^* = \|w^*\| e_1$.

Definition 4.2. *We define s in terms of C_0 and C_1,*

$$C_0 = \mathbb{E}_{x_1=0}[x_{-1} x_{-1}^\top] \tag{31}$$

$$C_1 = \frac{\mathbb{E}[\sigma(1-\sigma) x_{-1} x_{-1}^\top]}{\mathbb{E}[\sigma(1-\sigma)]} \tag{32}$$

$$\frac{1}{s} = \|C_0^{-1/2} C_1 C_0^{-1/2}\|_2 \tag{33}$$

We can give an interpretation to these constants. Define $C(a) = \mathbb{E}[x_{-1}x_{-1}^T | x_1 = a]$ as the covariance of the directions orthogonal to w^* at the slice $x_1 = a$. Then, C_0 is simply $C(0)$, the covariance at the decision boundary.

Further, define a variable B that is x_1 weighted by $\sigma(1-\sigma)$:

$$p(B = b) \propto \sigma(\|w^*\|b)(1 - \sigma(\|w^*\|b))p(x_1 = b). \quad (34)$$

Then, $C_1 = \mathbb{E}[C(B)]$, the covariance over the whole distribution, but weighted higher near the decision boundary with exponential tails. Finally, s compares how much these two covariances differ.

Intuitively, we need this parameter to handle the case where the covariance at the decision boundary (a factor of the Σ for active learning) is small relative to the average covariance.

Here is our main theorem which is proved by showing that

$$\Sigma_{\text{passive},-1} \succ \frac{s}{4\text{Err}} \Sigma_{\text{active},-1} \quad (35)$$

and then using Lemma 4.1.

Theorem 4.1. *For sufficiently small constant α (that depends on the dataset) and for $Err < \epsilon < \epsilon_0$,*

$$DE(\epsilon) > \frac{s}{4Err}. \quad (36)$$

We can also get an upper bound on the data efficiency if we make an additional assumption that the pdf of x factorizes into two marginal distributions (decomposition of x into two independent components), one along the direction of w^* and one in the directions orthogonal to w^*.

Theorem 4.2. *If $p(x) = p(x_1)p(x_{-1})$ and $p(x_1) = p(-x_1)$, then for sufficiently small constant α (that depends on the dataset), and for Err $< \epsilon < \epsilon_0$,*

$$\frac{1}{4Err} < DE(\epsilon) < \frac{1}{2Err}. \quad (37)$$

We can therefore see from these results that there is an inverse relationship between the asymptotic data efficiency and the population error, shedding light and giving a theoretical explanation to the empirical observation made in Section 3.

5. Discussion and Related Work

The conclusion of this work, that data efficiency is inversely related to limiting error, has been hinted at by a couple sentences in empirical survey papers. Schein & Ungar (2007) states "the data sets sort neatly by noise, with [uncertainty] sampling failing on more noisy data ... and performing at least as well as random [sampling] for [less noisy] data sets." Yang & Loog (2016) states "For the [less noisy] datasets, random sampling does not achieve the best performance ..., which may indicate that we need only consider random sampling on relatively [noisy] tasks".

Additionally, this conclusion has evidence from statistical active learning theory (Hanneke, 2014). While not mentioned in the work, the ratio between the passive and active bounds points to a $1/\text{Err}$ factor (though with Err being the optimal error over classifiers, not the MLE classifier). More specifically, the ratio between the passive and active lower bounds converges to $\Theta(1/\text{Err})$ as $\epsilon \to \text{Err}$. Additionally, the ratio of active and passive algorithms converge to $\Theta(1/\text{Err})$; however with a factor of a disagreement coefficient which has a dimension dependence for linear classifiers and a $\log\log 1/\epsilon$ factor which "is sometimes possible to remove" (Hanneke, 2014).

This conclusion can be used in practice in at least two possible ways. First, a pilot study or domain knowledge can be used to get a rough estimate of the final error and if the error is low enough (less than around 10%), uncertainty sampling can be used. Additionally, random sampling could be run until the test error is below 10% and then a switch be made to uncertainty sampling.

Does our conclusion hold for other models? Because of the mathematical similarity to SVM, it's likely it also holds for hinge loss. It is possible that it also holds for neural networks with a a softmax layer, since the softmax layer is mathematically equivalent to logistic regression. In fact, Geifman & El-Yaniv (2017) performs experiments with deep neural networks and multiclass classification on MNIST (1% error, 6x data efficiency), CIFAR-10 (10%, 2x), and CIFAR-100 (35%, 1x) and finds results that are explained well by our conclusion.

In conclusion, we make an observation, clearly define a phenomenon, demonstrate it empirically, and analyze it theoretically. The thesis of this work, that the data efficiency of uncertainty sampling on logistic regression is inversely proportional to the limiting error, sheds light on the appropriate use of active learning, enabling machine learning practitioners to intelligently choose their data collection techniques, whether active or passive.

Reproducibility

The code, data, and experiments for this paper are available on the CodaLab platform at

https://worksheets.codalab.org/worksheets/0x8ef22fd3cd384029bf1d1cae5b268f2d/.

Acknowledgments

This research was supported by NSF grant DGE-1656518.

References

Balcan, M.-F., Broder, A., and Zhang, T. Margin based active learning. In *International Conference on Computational Learning Theory*, pp. 35–50. Springer, 2007.

Balcan, M.-F., Beygelzimer, A., and Langford, J. Agnostic active learning. *Journal of Computer and System Sciences*, 75(1):78–89, 2009.

Brinker, K. Incorporating diversity in active learning with support vector machines. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 59–66, 2003.

Chaudhuri, K., Kakade, S. M., Netrapalli, P., and Sanghavi, S. Convergence rates of active learning for maximum likelihood estimation. In *Advances in Neural Information Processing Systems*, pp. 1090–1098, 2015.

Freund, Y., Seung, H. S., Shamir, E., and Tishby, N. Selective sampling using the query by committee algorithm. *Machine learning*, 28(2):133–168, 1997.

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

Hanneke, S. *Statistical Theory of Active Learning*. Now Publishers Incorporated, 2014.

Hoi, S. C., Jin, R., Zhu, J., and Lyu, M. R. Semisupervised svm batch mode active learning with applications to image retrieval. *ACM Transactions on Information Systems (TOIS)*, 27(3):16, 2009.

Lewis, D. D. and Gale, W. A. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–12. Springer-Verlag New York, Inc., 1994.

Liu, A., Reyzin, L., and Ziebart, B. D. Shift-pessimistic active learning using robust bias-aware prediction. In *AAAI*, pp. 2764–2770, 2015.

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

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

Schohn, G. and Cohn, D. Less is more: Active learning with support vector machines. In *ICML*, pp. 839–846, 2000.

Settles, B. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2010.

Seung, H. S., Opper, M., and Sompolinsky, H. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 287–294. ACM, 1992.

Sourati, J., Akcakaya, M., Leen, T. K., Erdogmus, D., and Dy, J. G. Asymptotic analysis of objectives based on fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017.

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

van der Vaart, A. W. *Asymptotic statistics*. Cambridge University Press, 1998.

Yang, Y. and Loog, M. A benchmark and comparison of active learning for logistic regression. *arXiv preprint arXiv:1611.08618*, 2016.