Uncertainty Sampling is Preconditioned Stochastic Gradient Descent on Zero-One Loss

Abstract

Uncertainty sampling, a popular active learning algorithm, is used to reduce the amount of data required to learn a classifier, but it has been observed in practice to converge to different parameters depending on the initialization and sometimes to even better parameters than standard training on all the data. In this work, we give a theoretical explanation of this phenomenon, showing that uncertainty sampling on a convex loss can be interpreted as performing a preconditioned stochastic gradient step on a smoothed version of the population zero-one loss that converges to the population zero-one loss. Furthermore, uncertainty sampling moves in a descent direction and converges to stationary points of the smoothed population zero-one loss. Experiments on synthetic and real datasets support this connection.

1 Introduction

Active learning algorithms aim to learn parameters with less data by querying labels adaptively. However, since such algorithms change the sampling distribution, they can introduce bias in the learned parameters. While there has been some work to understand this (Schütze et al., 2006; Bach, 2007; Dasgupta & Hsu, 2008; Beygelzimer et al., 2009), the most common algorithm, “uncertainty sampling” (Lewis & Gale, 1994; Settles, 2010), remains elusive. One of the oddities of uncertainty sampling is that sometimes the bias is helpful: uncertainty sampling with a subset of the data can yield lower error than random sampling on all the data (Schohn & Cohn, 2000; Bordes et al., 2005; Chang et al., 2017). But sometimes, uncertainty sampling can vastly underperform, and in general, different initializations can yield different parameters asymptotically. Despite the wealth of theory on active learning (Balcan et al., 2006; Hanneke et al., 2014), a theoretical account of uncertainty sampling is lacking.

In this paper, we characterize the dynamics of a streaming variant of uncertainty sampling to explain the bias introduced. We introduce a smoothed version of the zero-one loss which approximates and converges to the zero-one loss. We show that uncertainty sampling, which minimizes a convex surrogate loss on all the points so far, is asymptotically performing a preconditioned stochastic gradient step on the smoothed (non-convex) population zero-one loss. Furthermore, each uncertainty sampling iterate in expectation moves in a descent direction of the smoothed population zero-one loss, unless the parameters are at an approximate stationary point. In addition, uncertainty sampling converges to a stationary point of the smoothed population zero-one loss. This explains why uncertainty sampling sometimes achieves lower zero-one loss than random sampling, since that is approximately the quantity it implicitly optimizes. At the same time, as the zero-one loss is non-convex, we can get stuck in a local minimum with higher zero-one loss (see Figure 1).

Empirically, we validate the properties of uncertainty sampling on a simple synthetic dataset for intuition as well as 22 real-world datasets. Our new connection between uncertainty sampling and
Figure 1: A typical run of uncertainty sampling. Each iteration, uncertainty sampling chooses to label a point close to the current decision boundary. (a) Random initialization of uncertainty sampling. (b) A point close to the decision boundary is added and the decision boundary is updated. (c) Several more points close the decision boundary are added until convergence. We see that uncertainty sampling uses only a fraction of the data, but converges to a local minimum of the zero-one loss. (d) There are three different local minima of the zero-one loss, where the horizontal linear classifier is much more preferable than the other two.

zero-one loss minimization clarifies the importance of a sufficiently large seed set, rather than using a single point per class, as is commonly done in the literature (Tong & Koller 2001; Yang & Loog 2016).

2 Setup

We focus on binary classification. Let \( z = (x, y) \) be a data point, where \( x \in \mathbb{R}^k \) is the input and \( y \in \{-1, 1\} \) is the label, drawn from some unknown true data distribution \( z \sim p^* \). Assume we have a scoring function \( S(x, \theta) \), where \( \theta \in \mathbb{R}^d \) are the parameters; for linear models, we have \( S(x, \theta) = \theta \cdot \phi(x) \), where \( \phi : \mathbb{R}^k \to \mathbb{R}^d \) is the feature map.

Given parameters \( \theta \), we predict 1 if \( S(x, \theta) > 0 \) and \(-1\) otherwise, and therefore err when \( y \) and \( S(x, \theta) \) have opposite signs. Define \( Z(\theta) \) to be the zero-one loss (misclassification rate) over the data distribution, the evaluation metric of interest:

\[
Z(\theta) \defeq \mathbb{E}_{(x,y) \sim p^*} [H(-yS(x, \theta))],
\]

where \( H \) is the Heaviside step function:

\[
H(x) \defeq \begin{cases} 
0 & x < 0, \\
\frac{1}{2} & x = 0, \\
1 & x > 0.
\end{cases}
\]

Note that the training zero-one loss is step-wise constant, and the gradient is 0 almost everywhere. However, assuming the probability density function (PDF) of \( p^* \) is smooth, the population zero-one loss is differentiable at most parameters, a fact that will be shown later.

Since minimizing the zero-one loss is computationally intractable (Feldman et al. 2012), it is common to define a convex surrogate \( \ell((x, y), \theta) = \psi(yS(x, \theta)) \) which upper bounds the zero-one loss; for example, the logistic loss is given by \( \psi(s) = \log(1 + e^{-s}) \). Given a labeled dataset \( \mathcal{D} = \{z_1, \ldots, z_n\} \), we can define the estimator that minimizes the sum of the loss plus regularization:

\[
\theta_D \defeq \arg \min_{\theta} \sum_{z \in \mathcal{D}} \ell(z, \theta) + \lambda \|\theta\|_2^2.
\]

This can often be solved efficiently via convex optimization.

**Passive learning: random sampling.** Define the population surrogate loss as

\[
L(\theta) \defeq \mathbb{E}_{z \sim p^*} [\ell(z, \theta)].
\]

In standard passive learning, we sample \( \mathcal{D} \) randomly from the population and compute \( \theta_D \). As \( |\mathcal{D}| \to \infty \), the parameters generally converge to the minimizer of \( L \), which is in general distinct from the minimizer of \( Z \).
Active learning: uncertainty sampling. In this work, we consider the streaming setting (Settles 2010) where a learner receives a stream of unlabeled examples (known $x$ drawn from $p^x$ with unknown $y$) and must decide to label each point or not. We analyze uncertainty sampling in this setting (Lewis & Gale 1994; Settles 2010), which is widely used for its simplicity and efficacy (Yang & Loog 2016).

Let us denote our label budget as $n$, the number of points we label. Uncertainty sampling (Algorithm 1) begins with $n_{seed} < n$ labeled points $D$ drawn from the beginning of the stream and minimizes the regularized loss (3) to obtain initial parameters. Then, the algorithm takes a point from the stream and labels it with probability $q(S(x, \theta)/r)$ for some acceptance function $q$ and scalar $r$. One example is $q(s) = 1 ||s|| \leq 1$, which corresponds to labeling points from the stream if and only if $|S(x, \theta)| \leq r$. As $r$ gets smaller, we choose points closer to the decision boundary. If we decide to label $x$, then we obtain the corresponding label $y$ and add $(x, y)$ to $D$. Finally, we update the model by optimizing (3). The process is continued until we have labeled $n$ points in total.

3 Theory

We present four types of theoretical results. First, in Section 3.1 we show how the optimal parameters change with the addition of a single point to the convex surrogate (e.g. logistic) loss. Then, in Section 3.2, we introduce a smoothed version of the zero-one loss and show that uncertainty sampling is preconditioned stochastic gradient descent on this smoothed zero-one loss. Finally, we show that uncertainty sampling iterates in expectation move in a descent direction in Section 3.3 and that the uncertainty sampling converges to a stationary point of the smoothed zero-one loss (Section 3.4).

3.1 Incremental Parameter Updates

First, we analyze how the sample convex surrogate loss minimizer changes with a single additional point, showing that the change is a preconditioned gradient step on the additional point. Let us assume the loss is convex and thrice differentiable with bounded derivatives:

**Assumption 1 (Convex Loss).** The loss $\ell(z, \theta)$ is convex in $\theta$.

**Assumption 2 (Loss Regularity).** The loss $\ell(z, \theta)$ is continuously thrice differentiable in $\theta$, and the first three derivatives are bounded by some constant $M_\ell$ in the Frobenius norm.

Consider any iterative algorithm that adds a single point each time and minimizes the regularized training loss at each iteration $t$:

$$L_t(\theta) \overset{\text{def}}{=} \sum_{i=1}^t \ell(z^{(i)}, \theta) + \lambda \|	heta\|_2^2$$

(5)

to produce $\theta_t$. Since $L_{t-1}$ and $L_t$ differ by only one point, we expect $\theta_{t-1}$ and $\theta_t$ to also be close. We can make this formal using Taylor’s theorem. First, since $\theta_t$ is a minimizer, we have $\nabla L_t(\theta_t) = 0$.

1Preconditioned refers to multiplication of a symmetric positive semidefinite matrix by the (stochastic) gradient for (stochastic) gradient descent (Li 2018; Klein et al. 2011). It is often chosen to approximate the inverse Hessian.
Then, since the loss is continuously twice-differentiable:

\[ 0 = \nabla L_t(\theta_t) = \nabla L_t(\theta_{t-1}) + \int_0^1 \nabla^2 L_t((1-u)\theta_{t-1} + u\theta_t)du (\theta_t - \theta_{t-1}). \tag{6} \]

Let \( P_t \) be the value of the integral. Since \( \ell \) is convex and regularizer is quadratic, \( \nabla^2 L_t \) is symmetric positive definite, and thus \( P_t \) is symmetric positive definite and thus invertible. Therefore, we can solve for \( \theta_t \):

\[ \theta_t = \theta_{t-1} - P_t^{-1} \nabla L_t(\theta_{t-1}). \tag{7} \]

Since \( \theta_{t-1} \) minimizes \( L_{t-1} \), we have \( \nabla L_{t-1}(\theta_{t-1}) = 0 \). Also note that \( L_t(\theta) = L_{t-1}(\theta) + \ell(z^{(t)}, \theta) \). Thus,

\[ \theta_t = \theta_{t-1} - P_t^{-1} \nabla \ell(z^{(t)}, \theta_{t-1}). \tag{8} \]

The update above holds for any choice of \( z^{(t)} \), in particular, when \( z^{(t)} \) is chosen by either random sampling or uncertainty sampling.

For random sampling, \( z^{(t)} \sim p^* \), so we have

\[ \mathbb{E}[\nabla \ell(z^{(t)}, \theta_{t-1})] = \nabla L(\theta_{t-1}), \tag{9} \]

from which one can interpret the iterates of random sampling as preconditioned SGD on the population surrogate loss \( L \).

### 3.2 Parameter Updates of Uncertainty Sampling

Whereas random sampling is preconditioned SGD on the population surrogate loss \( L \), we will now show that uncertainty sampling is preconditioned SGD on a smoothed version of the population zero-one loss \( Z \).

Recall that \( q \) is the acceptance function. Define \( Q \) as a normalized anti-derivative of \( q \), which converges to the Heaviside function when the domain is scaled by \( r \). First, we need to make an assumption that ensures \( q \) has an integral over the real line.

**Assumption 3** (Continuous, bounded, even). The function \( q(s) \)

- is continuous,
- has bounded support \((\exists M_q : |s| \geq M_q \implies q(s) = 0)\),
- is even \((q(s) = q(-s))\).

Now we are ready to define \( Z_r \), which is made by replacing the Heaviside step function with \( Q \) and scaling the domain by \( r \).

\[ Q_\infty = \int_{-\infty}^{\infty} q(u)du \]

\[ Q(s) = \frac{\int_{-\infty}^{s} q(u)du}{Q_\infty} \]

\[ Z_r(\theta) = \mathbb{E}_{(x,y)\sim p^*}[\mathbb{E}(\mathcal{S}(x, \theta)/r)] \]

We now show that \( Z_r \) converges to \( Z \) pointwise:

**Proposition 4.** For all \( \theta \), \( \lim_{r \to 0} Z_r(\theta) = Z(\theta) \).

*Proof.* This follows from noticing that \( \lim_{r \to 0} Q(s/r) = H(s) \) and applying the Dominated Convergence Theorem.

Before stating the dynamics of uncertainty sampling, we must first define another quantity, \( a_r(\theta) \) which is the probability of accepting a random point \( x \):

\[ a_r(\theta) \overset{\text{def}}{=} \mathbb{E}[q(S(x, \theta)/r)], \tag{13} \]

We must further make an assumption that \( \psi \) is locally exactly linear around 0, which is satisfied for the hinge loss and smoothed versions of the hinge loss.
Assumption 5 (Locally linear $\psi$). There is some neighborhood of 0 where $\psi$ is exactly linear:

$$\exists m_\psi : |s| \leq m_\psi \implies \psi'(s) = \psi'(0)$$ (14)

We also assume the score $S$ is smooth, and the support of $p^*$ is bounded:

Assumption 6 (Smooth Score). The score $S(x, \theta)$ is smooth, that is, all derivatives with respect to $x$ and $\theta$ exist.

Assumption 7 (Bounded Support). The support of $p^*$ is bounded.

We are ready to state the relationship between uncertainty sampling iterates (8), which are governed by $\nabla\ell(z, \theta)$, and the smoothed zero-one loss $Z_r$:

**Theorem 8.** Under assumptions 3, 5, 6, and 7 for any $\theta$, if $a_r(\theta) \neq 0$ and if $z$ is chosen via uncertainty sampling with parameters $\theta$ and scale $r \leq m_\psi / M_q$,

$$E[\nabla\ell(z, \theta)] = \eta \nabla Z_r(\theta), \quad \eta \overset{\text{def}}{=} \frac{-\psi'(0) Q_\infty r}{a_r(\theta)}.$$ (15)

**Proof.** First, note that

$$\nabla Z_r(\theta) = \nabla \int Q(-yS(x, \theta)/r) dp^*(x, y).$$ (16)

Because $Q$ is continuously differentiable, $S$ is smooth, and $p^*$ has bounded support, by the Leibniz Integral Rule, we can exchange the integral and derivative:

$$\nabla Z_r(\theta) = \int \frac{q(-yS(x, \theta)/r)}{Q_\infty} (-y/r) \nabla_\theta S(x, \theta) dp^*(x, y).$$ (17)

Because $q$ is even,

$$\nabla Z_r(\theta) = -\frac{1}{Q_\infty r} \int q(S(x, \theta)/r) y \nabla_\theta S(x, \theta) dp^*(x, y).$$ (18)

Now we are ready to evaluate $E[\nabla\ell(z, \theta)]$. From the definition of uncertainty sampling,

$$E[\nabla\ell(z, \theta)] = \frac{\int q(S(x, \theta)/r) \nabla\ell(z, \theta) dp^*(x, y)}{\int q(S(x, \theta)/r) dp^*(x, y)}$$

$$= \frac{1}{a_r(\theta)} \int q(S(x, \theta)/r) \psi'(yS(x, \theta)) y \nabla_\theta S(x, \theta) dp^*(x, y)$$ (19)

Notice that $q(S(x, \theta)/r) = 0$ for $|S(x, \theta)| \geq r M_q$ and that $\psi'(s) = \psi'(0)$ for $|s| \leq m_\psi$. Thus, for $r \leq m_\psi / M_q$,

$$E[\nabla\ell(z, \theta)] = \frac{1}{a_r(\theta)} \int q(S(x, \theta)/r) \psi'(0) y \nabla_\theta S(x, \theta) dp^*(x, y)$$

$$= -\frac{\psi'(0) Q_\infty r}{a_r(\theta)} \nabla Z_r(\theta).$$ (20)

Thus, if $z^{(t)}$ is drawn using uncertainty sampling, $E[\nabla\ell(z^{(t)}, \theta_{t-1})]$ is in the direction of $\nabla Z_r(\theta_{t-1})$, since the quantity in front of $\nabla Z_r$ is a scalar that is positive for all common losses. Similar to how we showed random sampling is preconditioned SGD on the population surrogate loss $L$ [9], uncertainty sampling is preconditioned SGD on the smoothed population zero-one loss $Z_r$.

The only assumption that is unorthodox is Assumption 5 which holds for a smoothed hinge loss, but not the logistic loss. If we remove this assumption, we would incur only a small additive vector term of $O(r^2/a_r(\theta))$ which goes to 0 quickly as $r \to 0$.

While $Z_r(\theta) \to Z(\theta)$ pointwise as $r \to 0$ (Proposition 4), showing that $\nabla Z_r(\theta) \to \nabla Z(\theta)$ requires additional assumptions (see Section 6.2 in the Appendix for more details).

$^2Q_\infty$, $r$, and $a_r(\theta)$ are all positive, and $\psi'(0)$ is negative for all reasonable losses.
3.3 Descent Direction

So far, we have shown that uncertainty sampling is preconditioned SGD on the smoothed population zero-one loss $Z_r$, by analyzing $\mathbb{E}[\nabla \ell(z, \theta)]$. To show that these updates are descent directions on $Z_r$, we need to also consider the preconditioner $P_r^{-1}$ appearing in (8). Due to quadratic regularization (5), the preconditioner is positive definite. However, we need to be careful since the preconditioner depends on the resulting iterate $\theta_t$. Because of this snag, we need to ensure that $\theta_t$ doesn’t change the preconditioner $P_r^{-1}$ too much, which we can accomplish by requiring $\|\nabla Z(\theta_{t-1})\| \geq \epsilon$ and large enough regularization.

**Theorem 9** (Uncertainty Sampling Descent Direction). Assume that Assumptions 1, 2, 3, 5, 6, 7 hold, and $\psi'(0) < 0$. For any $\epsilon > 0$ and $n$, for any

$$\lambda \geq \max \left( \frac{4M_r^3 n}{\epsilon (-\psi'(0)) Q_{\infty} r}, \frac{M_r^4 n^2}{\epsilon (-\psi'(0)) Q_{\infty} r} \right), \tag{23}$$

for all iterates of uncertainty sampling $\{\theta_t\}$ such that $\|\nabla Z(\theta_{t-1})\| \geq \epsilon$, we have

$$\nabla Z_r(\theta_{t-1}) \cdot \mathbb{E}[\theta_t - \theta_{t-1} | \theta_{t-1}] < 0. \tag{24}$$

It might seem that $\lambda = \Omega(n^{2/3})$ above is rather large. However, note that we don’t optimize the regularized loss until after $n_{\text{seed}}$ random points, so if $n_{\text{seed}} \geq \lambda$, then the regularization is always less than the number of data points that contribute to the loss, and the regularization will not dominate. Thus, this constraint on $\lambda$ can be intuitively thought of as a constraint on $n_{\text{seed}} = \Omega(n^{2/3})$.

3.4 Convergence

Having shown that uncertainty sampling iterates move in descent directions of $Z_r$ in expectation, we now turn to showing that they also converge to a stationary point of $Z_r$. To prove convergence, we will need to stay (with high probability) in regions where the assumption of Theorem 8 holds ($a_r(\theta) \neq 0$) and also ensure that the parameters stay bounded (with high probability).

As is standard in stochastic gradient convergence analyses, instead of showing convergence of the final parameter iterate $\theta_n$, we show the convergence of the parameters from a random iteration. For a budget $n$, let $t \in [n_{\text{seed}}, n]$ with $\mathbb{P}(\tilde{t} = t) \propto 1/t$. Then, define $\tilde{\theta}_n \overset{\text{def}}{=} \theta_{\tilde{t}}$ be the randomized parameters. We will show that $\tilde{\theta}_n$ converges to a stationary point of $Z_r$ as $n \rightarrow \infty$.

Define $\delta_{C,r,n}$ as the failure probability that any parameter iterate $\theta_t$ is too large or has zero acceptance probability:

$$\delta_{C,r,n} \overset{\text{def}}{=} \mathbb{P}[\exists t \in [n_{\text{seed}}, n] : \|\theta_t\| > C \vee a_r(\theta_t) = 0]. \tag{25}$$

We will assume that this failure probability converges to 0 as $n \rightarrow \infty$. As $n$ grows, the seed set size grows, the regularization grows, and the effective step size shrinks. Intuitively, this means we might expect that the parameter iterates become more stable. Unless the parameters diverge, we can choose $C$ large enough to contain the parameter iterates. Furthermore, the acceptance probability should be non-zero if there are not large regions of the space with zero probability density. Assuming the probability of these failure events goes to zero, the randomized parameters converge to a stationary point:

**Theorem 10** (Convergence to Stationary Points). Assume that Assumptions 1, 2, 3, 5, 6, 7 hold and that $\psi'(0) < 0$. Assume $\lambda = \Omega(n^{2/3})$ and $n_{\text{seed}} = o(n)$, and there exists some $C > 0$ such that $\delta_{C,r,n} \rightarrow 0$ as $n \rightarrow \infty$. Then for the randomized parameters produced by uncertainty sampling with fixed scale $r \leq m_\psi / M_\psi$, we have that as $n \rightarrow \infty$,

$$\nabla Z_r(\tilde{\theta}_n) \overset{P}{\rightarrow} 0. \tag{26}$$

These results shed light on the mysterious dynamics of uncertainty sampling which motivated this paper. In particular, uncertainty sampling can achieve lower zero-one loss than random sampling because it is implicitly descending on the smoothed zero-one loss $Z_r$. Furthermore, since $Z_r$ is non-convex, uncertainty sampling can converge to different values depending on the initialization.
Figure 2: Synthetic dataset based on a mixture of four Gaussians (left) and the associated learning curves for runs of uncertainty sampling with different initial seed sets (right). Depending on the seed set, uncertainty sampling can produce either better or worse parameters than random sampling.

It is important to note that the actual uncertainty sampling algorithm is unchanged—it is still performing gradient updates on the convex surrogate loss. But because its sampling distribution is skewed towards the decision boundary, we can interpret its updates as being on the smoothed zero-one loss with respect to the original data-generating distribution.

4 Experiments

We run uncertainty sampling on a simple synthetic dataset to illustrate the dynamics (Section 4.1) as well as 22 real datasets (Section 4.2). In both cases, we show how uncertainty sampling converges to different parameters depending on initialization, and how it can achieve lower asymptotic zero-one loss compared to minimizing the surrogate loss on all the data. Note that most active learning experiments are interested in measuring the rate of convergence (data efficiency), whereas this paper focuses exclusively on asymptotic values and the variation that we obtain from different seed sets. We evaluate only on the zero-one loss, but all algorithms perform optimization on the logistic loss.

4.1 Synthetic Data

Figure 2 (left) shows a mixture of Gaussian distributions in two dimensions. All the Gaussians are isotropic, and the size of the circle indicates the variance (one standard deviation for the inner circle, and two standard deviations for the outer circle). The points drawn from the two red Gaussian distributions are labeled $y = 1$ and the points drawn from the two blue ones are labeled $y = -1$. The percentages refer to the mixture proportions of the clusters. We see that there are four local minima of the population zero-one loss, indicated by the green dashed lines. Each minimum misclassifies one of the Gaussian clusters, yielding error rates of about 10%, 20%, 30%, and 40%. The black dotted line corresponds to the parameters that minimize the logistic loss, which yields an error of about 20%.

Figure 2 (right) shows learning curves for different seed sets, which consist of two points, one from each class. We see that the uncertainty sampling learning curves converge to four different asymptotic losses, corresponding to the four local minima of the zero-one loss mentioned earlier. The thick black dashed line is the zero-one loss for random sampling. We see that uncertainty sampling can actually achieve lower loss than random sampling, since the global optimum of the logistic loss does not correspond to the global minimum of the zero-one loss.

4.2 Real-World Datasets

We collected 22 datasets from OpenML (retrieved August, 2017) that had a large number of data points and where logistic regression outperformed the baseline classifier that always predicts the majority label. We further subsampled each dataset to have 10,000 points, which was divided into 7000 training points and 3000 test points. We created a stream of points by randomly selecting points one-by-one with replacement from the dataset. We ran uncertainty sampling on each dataset with
Figure 3: A scatter plot of the asymptotic zero-one loss for uncertainty sampling for two particular datasets for 13 seed sizes. The black line is the zero-one loss on the full dataset.

Figure 4: A plot showing the distribution of runs over the datasets (with 10 runs per dataset) of when uncertainty sampling converges to a lower zero-one loss than using the entire dataset.

Figure 5: A violin plot capturing the relative asymptotic zero-one loss compared to the zero-one loss on the full dataset. The plot shows the density of points with kernel density estimation. The red lines are the median losses. Each “violin” captures 220 points (10 runs over 22 datasets).

random seed sets of sizes that are powers of two from 2 to 4096 and then 7000. We stopped when uncertainty sampling either could not select a point ($\alpha_r(\theta) = 0$) or when a point had been repeatedly selected more than 10 times. For each dataset and seed set size, we ran uncertainty sampling 10 times, for a total of $22 \cdot 13 \cdot 10 = 2860$ runs.

In Figure 3 we see scatter plots of the asymptotic zero-one loss of 130 points: 13 seed set sizes, each with 10 runs. The dataset on the left was chosen to exhibit the wide range of convergence values of uncertainty sampling, some with lower zero-one loss than with the full dataset. In both plots, we see that the variance of the zero-one loss of uncertainty sampling decreases as the seed set grows. This is expected from theory since the initialization has less variance for larger seed set sizes (as the seed set size goes to infinity, the parameters converge). For most of the datasets, the behavior was more similar to the plot on the right, where uncertainty sampling has a higher mean zero-one loss than random sampling for most seed sizes.

To gain a more quantitative understanding of all the datasets, we summarized the asymptotic zero-one loss of uncertainty sampling for various random seed set sizes. In Figure 4 we show the proportions of the runs over the datasets where uncertainty sampling converges to a lower zero-one loss than using the entire dataset. In Figure 5 we show a “violin plot” for the distribution of the ratio between the asymptotic zero-one loss of uncertainty sampling and the zero-one loss using the full dataset. We note that the mean and variance of uncertainty sampling significantly drops as the size of the seed set grows larger. The initial parameters are poor if the seed set is small, and it is well-known that poor initializations for optimizing non-convex functions locally can yield poor results, as seen here.
5 Related Work and Discussion

The phenomenon that uncertainty sampling can achieve lower error with a subset of the data rather than using the entire dataset has been observed multiple times in the literature. In fact, the original uncertainty sampling paper (Lewis & Gale, 1994) notes that “For 6 of 10 categories, the mean [F-score] for a classifier trained on a uncertainty sample of 999 examples actually exceeds that from training on the full training set of 319,463”. Schohn & Cohn (2000) defines a heuristic that selects the point closest to the decision boundary of an SVM, which is similar to uncertainty sampling in our formulation. In the abstract, the authors note, “We observe... that a SVM trained on a well-chosen subset of the available corpus frequency performs better than one trained on all available data”. More recently, Chang et al. (2017) developed an “active bias” technique that emphasizes the uncertain points and found that it increases the performance compared to using a fully-labeled dataset.

There is also work showing the bias of active learning can harm final performance. Schütze et al. (2006) notes the “missed cluster effect”, where active learning can ignore clusters in the data and never query points from there; this is seen in our synthetic experiment. Dasgupta & Hsu (2008) has a section on the bias of uncertainty sampling and provides another example where uncertainty sampling fails due to sampling bias, which we can explain as due to local minima of the zero-one loss. Bach (2007) and Beygelzimer et al. (2009) note this bias issue and propose different importance sampling schemes to re-weight points and correct for the bias.

In this work, we showed that uncertainty sampling updates are preconditioned SGD steps on the population zero-one loss and move in descent directions for parameters that are not approximate stationary points. Note that this does not give any global optimality guarantees. In fact, for linear classifiers, it is NP-hard to optimize the training zero-one loss below \( \frac{1}{2} - \epsilon \) (for any \( \epsilon > 0 \)) even when there is a linear classifier that achieves just \( \epsilon \) training zero-one loss (Feldman et al., 2012).

One of the key questions in light of this work is when optimizing convex surrogate losses yield good zero-one losses. If the loss function corresponds to the negative log-likelihood of a well-specified model, then the zero-one loss \( Z \) will have a local minimum at the parameters that optimize the log-likelihood. If the loss function is “classification-calibrated” (which holds for most common surrogate losses), Bartlett et al. (2006) shows that if the convex surrogate loss of the estimated parameters converges to the optimal convex surrogate loss, then the zero-one loss of the estimated parameters converges to the global minimum of the zero-one loss (Bayes error). This holds only for universal classifiers (Micchelli et al., 2006), and in practice, these assumptions are unrealistic. For instance, several papers show how outliers and noise can cause linear classifiers learned on convex surrogate losses to suffer high zero-one loss (Nguyen & Sanner, 2013; Wu & Liu, 2007; Long & Servedio, 2010).

Other works connect active learning with optimization in rather different ways. Ramdas & Singh (2013) uses active learning as a subroutine to improve stochastic convex optimization. Guillory et al. (2009) shows how performing online active learning updates corresponds to online optimization updates of non-convex functions, more specifically, truncated convex losses. In this work, we analyze active learning with offline optimization and show the connection between uncertainty sampling and one particularly important non-convex loss, the zero-one loss.

In summary, our work is the first to show a connection between the zero-one loss and the commonly-used uncertainty sampling. This provides an explanation and understanding of the various empirical phenomena observed in the active learning literature. Uncertainty sampling simultaneously offers the hope of converging to lower error and the danger of converging to local minima (an issue that can possibly be avoided with larger seed sizes). We hope this connection can lead to improved active learning and optimization algorithms.

Reproducibility. The code, data, and experiments for this paper are available on the CodaLab platform at https://worksheets.codalab.org/worksheets/0x18dfe5bce1dc4081b54b3cc15a5abce8/.

Acknowledgments. This research was supported by an NSF Graduate Fellowship to the first author.
References

Bach, F. R. Active learning for misspecified generalized linear models. In Advances in neural information processing systems, pp. 65–72, 2007.

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

Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473):138–156, 2006.

Beygelzimer, A., Dasgupta, S., and Langford, J. Importance weighted active learning. In Proceedings of the 26th annual international conference on machine learning, pp. 49–56. ACM, 2009.

Bordes, A., Ertekin, S., Weston, J., and Bottou, L. Fast kernel classifiers with online and active learning. Journal of Machine Learning Research, 6(Sep):1579–1619, 2005.

Chang, H.-S., Learned-Miller, E., and McCallum, A. Active bias: Training more accurate neural networks by emphasizing high variance samples. In Advances in Neural Information Processing Systems, pp. 1003–1013, 2017.

Dasgupta, S. and Hsu, D. Hierarchical sampling for active learning. In Proceedings of the 25th international conference on Machine learning, pp. 208–215. ACM, 2008.

Feldman, V., Guruswami, V., Raghavendra, P., and Wu, Y. Agnostic learning of monomials by halfspaces is hard. SIAM Journal on Computing, 41(6):1558–1590, 2012.

Guillory, A., Chastain, E., and Bilmes, J. Active learning as non-convex optimization. In Artificial Intelligence and Statistics, pp. 201–208, 2009.

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

Hoveijn, I. Differentiability of the volume of a region enclosed by level sets. arXiv preprint arXiv:0712.0915, 2007.

Klein, S., Staring, M., Andersson, P., and Pluim, J. P. Preconditioned stochastic gradient descent optimisation for monomodal image registration. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 549–556. Springer, 2011.

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.

Li, X.-L. Preconditioned stochastic gradient descent. IEEE transactions on neural networks and learning systems, 29(5):1454–1466, 2018.

Long, P. M. and Servedio, R. A. Random classification noise defeats all convex potential boosters. Machine learning, 78(3):287–304, 2010.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. Journal of Machine Learning Research, 7(Dec):2651–2667, 2006.

Nguyen, T. and Sanner, S. Algorithms for direct 0–1 loss optimization in binary classification. In International Conference on Machine Learning, pp. 1085–1093, 2013.

Ramdas, A. and Singh, A. Algorithmic connections between active learning and stochastic convex optimization. In International Conference on Algorithmic Learning Theory, pp. 339–353. Springer, 2013.

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

Schütze, H., Velipasaoglu, E., and Pedersen, J. O. Performance thresholding in practical text classification. In Proceedings of the 15th ACM international conference on Information and knowledge management, pp. 662–671. ACM, 2006.
Settes, B. Active learning literature survey. *Computer Sciences Technical Report*, 1648, 2010.

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.

Wu, Y. and Liu, Y. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

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

The appendix has two main sections. In Section 6.1, we prove the results about the descent direction and SGD convergence of uncertainty sampling. In Section 6.2, we show that under some conditions, $\nabla Z_r \rightarrow \nabla Z$.

6.1 Descent Direction and Convergence

We prove two lemmas about the parameter updates. First, we show that a single step of the parameter iterates don’t change much due to regularization (Lemma 11). Second, we show that $\theta_t - \theta_{t-1}$ is approximately equal to $-\left[\nabla^2 L_{t-1}(\theta_{t-1})\right]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1})$ (Lemma 12) and bound the error in approximation. This is important because it takes the dependency on iterate $t$ only through the gradient of the loss at $z^{(t)}$, the point selected at iterate $t$. With these two lemmas and Theorem 8, the descent direction (Theorem 9) is straightforward and the SGD convergence (Theorem 10) follows a standard SGD convergence argument.

6.1.1 Parameter Update Lemmas

Lemma 11.

$$\|\theta_t - \theta_{t-1}\| \leq \frac{M_t}{\lambda}$$

Proof. As in the main text, we have

$$L_t(\theta) = \sum_{i=1}^{t} \ell(z^{(i)}, \theta) + \lambda\|\theta\|^2_2.$$  (28)

Thus, $L_t(\theta) = L_{t-1}(\theta) + \ell(z^{(t)}, \theta)$ and further $\nabla L_t(\theta_t) = 0$. Together, this implies that $\nabla L_t(\theta_{t-1}) = \nabla \ell(z^{(t)}, \theta_{t-1})$.

Using the Taylor expansion,

$$0 = \nabla L_t(\theta_t) = \nabla \ell(z^{(t)}, \theta_{t-1}) + P_t(\theta_t - \theta_{t-1}),$$  (29)

where

$$P_t = \int_{0}^{1} \nabla^2 L_t((1-u)\theta_{t-1} + u\theta_t)du.$$  (30)

Since the loss is convex with quadratic regularization,

$$\lambda_{\text{min}}(P_t) \geq \int_{0}^{1} \lambda_{\text{min}}(\nabla^2 L_t((1-u)\theta_{t-1} + u\theta_t))du$$

$$\geq \int_{0}^{1} \lambda_{\text{min}}(\lambda I)du$$

$$\geq \lambda$$

$$\|P_t^{-1}\| \leq \frac{1}{\lambda}$$  (34)

Therefore,

$$\theta_t - \theta_{t-1} = -[P_t]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1})$$  (35)

$$\|\theta_t - \theta_{t-1}\| \leq \frac{M_t}{\lambda}$$  (36)
Lemma 12.
\[ \|\theta_t - \theta_{t-1} + [\nabla^2 L_{t-1}(\theta_{t-1})]^{-1}\nabla \ell(z^{(t)}, \theta_{t-1})\| \leq \frac{M_t^2}{\lambda^2} + \frac{M_t^3 n}{2\lambda^3} \] (37)

Proof. From a Taylor expansion,
\[ 0 = \nabla L_t(\theta_t) = \nabla \ell(z^{(t)}, \theta_{t-1}) + \nabla^2 L_t(\theta_{t-1})(\theta_t - \theta_{t-1}) + Q, \]
where
\[ Q_t = (\theta_t - \theta_{t-1})^T \int_0^1 (1-u)[\nabla^3 L_t(u\theta_t + (1-u)\theta_{t-1})]_i du \] (39)

We want to solve for \( \theta_t - \theta_{t-1} \), but in order to do this, we need to bound \( Q \).
\[ \|Q\| \leq \|\theta_t - \theta_{t-1}\| \int_0^1 (1-u)\|\nabla^3 L_t(u\theta_t + (1-u)\theta_{t-1})\|E \| du \|\theta_t - \theta_{t-1}\| \] (40)

Using Lemma 11
\[ \leq \frac{M_t}{\lambda} \int_0^1 (1-u)du M_t \frac{M_t}{\lambda} \]
\[ \leq \frac{M_t n M_t M_t}{\lambda} \]
\[ \leq \frac{M_t^3 n}{2\lambda^2} \] (43)

Solving for \( \theta_t - \theta_{t-1} \) in the Taylor expansion,
\[ \theta_t - \theta_{t-1} = -[\nabla^2 L_t(\theta_{t-1})]^{-1}(\nabla \ell(z^{(t)}, \theta_{t-1}) + Q) \] (44)
\[ \theta_t - \theta_{t-1} + [\nabla^2 L_t(\theta_{t-1})]^{-1}\nabla \ell(z^{(t)}, \theta_{t-1}) = -[\nabla^2 L_t(\theta_{t-1})]^{-1}Q \] (45)
\[ \|\theta_t - \theta_{t-1} + [\nabla^2 L_t(\theta_{t-1})]^{-1}\nabla \ell(z^{(t)}, \theta_{t-1})\| \leq \|\|\nabla^2 L_t(\theta_{t-1})]^{-1}||Q\| \]
\[ \leq \frac{1}{\lambda} \frac{M_t^3 n}{2\lambda^2} \] (47)
\[ \leq \frac{M_t^3 n}{2\lambda^3} \] (48)
\[ \leq \frac{M_t^3 n}{2\lambda^3} \] (49)

Looking at the theorem statement, we are almost done. The only difference between the theorem statement and the equation above is that the theorem statement has \( \nabla^2 L_{t-1} \) while the equation above has \( \nabla^2 L_t \). We can use the triangle inequality and bound the difference.
\[ \|\theta_t - \theta_{t-1} + [\nabla^2 L_{t-1}(\theta_{t-1})]^{-1}\nabla \ell(z^{(t)}, \theta_{t-1})\| \]
\[ \leq \|\|\nabla^2 L_{t-1}(\theta_{t-1})]^{-1} - [\nabla^2 L_t(\theta_{t-1})]^{-1}||\nabla \ell(z^{(t)}, \theta_{t-1})\| + \frac{M_t^3 n}{2\lambda^3} \] (51)
\[ \leq \|\|\nabla^2 L_{t-1}(\theta_{t-1})]^{-1} - [\nabla^2 L_t(\theta_{t-1})]^{-1}||\nabla \ell(z^{(t)}, \theta_{t-1})\| \]
\[ \leq \|\nabla^2 L_{t-1}(\theta_{t-1})]^{-1} - [\nabla^2 L_t(\theta_{t-1})]^{-1}||\nabla \ell(z^{(t)}, \theta_{t-1})\| \]
\[ \leq \frac{1}{\lambda} M_t \frac{1}{\lambda} M_t + \frac{M_t^3 n}{2\lambda^3} \]
\[ \leq \frac{M_t^2}{\lambda^2} + \frac{M_t^3 n}{2\lambda^3} \] (54)
6.1.2 Descent Direction

**Theorem 9:** Assume that Assumptions 1 and 2 are met. Further, assume \( \psi'(0) < 0 \). For any \( \epsilon > 0 \) and \( n \), for any sufficiently large

\[
\lambda \geq \max \left( \frac{4M^2}{\epsilon} \sqrt{\frac{n}{Q_{\infty}}}, \epsilon \left( \frac{M^2}{Q_{\infty}} \right)^3 \right),
\]

(55)

for all iterates of uncertainty sampling \( \{\theta_t\} \) such that \( \|\nabla Z(\theta_{t-1})\| \geq \epsilon \), then,

\[
\nabla Z_r(\theta_{t-1}) \cdot \mathbb{E}[\theta_t - \theta_{t-1}|\theta_{t-1}] < 0.
\]

(56)

**Proof.** The first thing to note is that if \( \|\nabla Z_r(\theta_{t-1})\| > 0 \), then \( a_r(\theta_{t-1}) > 0 \).

\[
\|\nabla Z_r(\theta_{t-1})\| = \frac{1}{Q_{\infty}^{\epsilon r}} \int q(S(x, \theta)/r) y \nabla_\theta S(x, \theta) dp^*(x, y)
\]

(57)

\[
\leq \frac{1}{Q_{\infty}^{\epsilon r}} \int q(S(x, \theta)/r) \|\nabla_\theta S(x, \theta)\| dp^*(x, y)
\]

(58)

Because \( p^* \) has bounded support and since \( S \) is smooth, there exists a constant \( C_0 \) such that \( \|\nabla S(x, \theta)\| \leq C_0 \).

\[
\|\nabla Z_r(\theta_{t-1})\| \leq \frac{C_{\theta_{t-1}}}{Q_{\infty}^{\epsilon r}} a_r(\theta).
\]

(59)

And thus, if \( \|\nabla Z_r(\theta_{t-1})\| \geq \epsilon > 0 \), then \( a_r(\theta_{t-1}) > 0 \).

This will allow us to use Theorem 8 later in the proof.

Using Lemma 12

\[
\nabla Z_r(\theta_{t-1}) \cdot (\theta_t - \theta_{t-1})
\]

\[
\leq -\nabla Z_r(\theta_{t-1})^T [\nabla^2 L_{t-1}(\theta_{t-1})]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1}) + \|\nabla Z_r(\theta_{t-1})\| \left( \frac{M^2}{\lambda^2} + \frac{M^3 n}{2\lambda^3} \right).
\]

(61)

Note that the only part that is dependent on the \( t \) iteration is the \( \nabla \ell \) term, which we can evaluate on the right by Theorem 8. Thus,

\[
\nabla Z_r(\theta_{t-1}) \cdot \mathbb{E}[\theta_t - \theta_{t-1}|\theta_{t-1}]
\]

\[
\leq -\nabla Z_r(\theta_{t-1})^T [\nabla^2 L_{t-1}(\theta_{t-1})]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1}) + \|\nabla Z_r(\theta_{t-1})\| \left( \frac{M^2}{\lambda^2} + \frac{M^3 n}{2\lambda^3} \right)
\]

(63)

\[
\leq -\|\nabla Z_r(\theta_{t-1})\|^2 \frac{1}{nM_\epsilon} (-\psi'(0)Q_{\infty}r) + \|\nabla Z_r(\theta_{t-1})\| \left( \frac{M^2}{\lambda^2} + \frac{M^3 n}{2\lambda^3} \right)
\]

(64)

\[
\leq -\|\nabla Z_r(\theta_{t-1})\| \frac{\epsilon}{nM_\epsilon} (-\psi'(0)Q_{\infty}r) + \|\nabla Z_r(\theta_{t-1})\| \left( \frac{M^2}{\lambda^2} + \frac{M^3 n}{2\lambda^3} \right)
\]

(65)

\[
\leq -\|\nabla Z_r(\theta_{t-1})\| \frac{\epsilon}{nM_\epsilon} (-\psi'(0)Q_{\infty}r) \left( 1 - \frac{M^3 n}{\lambda^2 \epsilon(-\psi'(0))Q_{\infty}r} - \frac{M^4 n^2}{2\lambda^3 \epsilon(-\psi'(0))Q_{\infty}r} \right)
\]

(67)

If the last term is positive then the whole expression is less than 0 and the theorem is proved. A sufficient condition for this to be the case is that,
\[
\frac{M^3_t n}{\lambda^2 \epsilon(-\psi'(0)) Q_{\infty} r} \leq \frac{1}{4},
\]

(68)

and

\[
\frac{M^2_t n^2}{2 \lambda^3 \epsilon(-\psi'(0)) Q_{\infty} r} \leq \frac{1}{2},
\]

(69)

which are both satisfied for

\[
\lambda \geq \max \left( \sqrt{\frac{4M^3_t n}{\epsilon(-\psi'(0)) Q_{\infty} r}}, \sqrt{\frac{M^2_t n^2}{\epsilon(-\psi'(0)) Q_{\infty} r}} \right).
\]

(70)

### 6.1.3 Convergence

**Theorem 10** Assume that Assumptions 1, 2, 3, 5, 6, 7 hold and that \( \psi'(0) < 0 \). If \( \lambda = \Omega(n^{2/3}) \) and \( n_{\text{seed}} = o(n) \), and if there exists some \( C > 0 \) such that \( \delta_{\text{C}, r, n} \to 0 \) as \( n \to \infty \), then for randomized uncertainty sampling parameters with fixed scale \( r \leq m_{\psi}/M_Z \), as \( n \to \infty \),

\[
\nabla Z_r(\hat{\theta}_n) \to 0.
\]

(71)

**Proof.** Assume that the parameter iterates are bounded by \( C \) and the acceptance probability \( a_r(\theta) \) is non-zero for all parameter iterates. This will occur with probability going to 1 and so if we can show convergence in probability under this condition, then unconditional convergence in probability follows. The set of parameters that are bounded are a compact set and thus \( \nabla Z_r(\theta) \) and \( \nabla^2 Z_r(\theta) \) are bounded by some constant, call it \( M_Z \).

From a Taylor expansion, for some \( \theta' \) between \( \theta_{t-1} \) and \( \theta_t \),

\[
Z_r(\theta_t) = Z_r(\theta_{t-1}) + \nabla Z_r(\theta_{t-1})^T (\theta_t - \theta_{t-1}) + (\theta_t - \theta_{t-1})^T \nabla^2 Z_r(\theta') (\theta_t - \theta_{t-1})
\]

(72)

\[
\leq Z_r(\theta_{t-1}) + \nabla Z_r(\theta_{t-1})^T (\theta_t - \theta_{t-1}) + \frac{M_t}{\lambda} M_Z \frac{M_t}{\lambda}
\]

(73)

\[
\leq Z_r(\theta_{t-1}) - \nabla Z_r(\theta_{t-1})^T \left[ \nabla^2 L_{t-1}(\theta_{t-1}) \right]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1}) + M_Z \left( \frac{M_t^2}{\lambda^2} + \frac{M_t^3 n}{2 \lambda^3} \right)
\]

(74)

\[
\leq Z_r(\theta_{t-1}) - \nabla Z_r(\theta_{t-1})^T \left[ \nabla^2 L_{t-1}(\theta_{t-1}) \right]^{-1} \nabla \ell(z^{(t)}, \theta_{t-1}) + O(1/n)
\]

(75)

Taking an expectation conditioned on \( \theta_{t-1} \),

\[
\mathbb{E}[Z_r(\theta_t)|\theta_{t-1}] \leq Z_r(\theta_{t-1}) - \frac{\psi'(0)Q_{\infty} r}{a_r(\theta_{t-1})} \nabla Z_r(\theta_{t-1})^T \left[ \nabla^2 L_{t-1}(\theta_{t-1}) \right]^{-1} \nabla Z(\theta_{t-1}) + O(1/n)
\]

(76)

\[
\leq Z_r(\theta_{t-1}) - (\psi'(0)Q_{\infty} r) \|\nabla Z_r(\theta_{t-1})\|^2 \frac{1}{M_t(t-1)} + O(1/n)
\]

(77)

Taking the expectation with respect to the randomness in the entire algorithm,
\[
E[Z_r(\theta_t)] \leq E[Z_r(\theta_{t-1})] - \frac{-\psi'(0)Q_\infty r}{M_t} \mathbb{E}[\|\nabla Z_r(\theta_{t-1})\|^2] \frac{1}{t-1} + O(1/n).
\] (79)

Rearranging and summing over iterations \( t \in [n_{\text{seed}} + 1, n] \),

\[
-\frac{-\psi'(0)Q_\infty r}{M_t} \sum_{t=n_{\text{seed}}}^{n-1} \frac{1}{t} \mathbb{E}[\|\nabla Z_r(\theta_t)\|^2] \leq E[Z_r(\theta_{n_{\text{seed}}})] - E[Z_r(\theta_n)] + O(1).
\] (80)

Since \( Z_r \) is bounded between 0 and 1, then, for some constant \( C \),

\[
\sum_{t=n_{\text{seed}}}^{n-1} \frac{1}{t} \mathbb{E}[\|\nabla Z_r(\theta_t)\|^2] \leq C
\] (81)

\[
\frac{\sum_{t=n_{\text{seed}}}^{n-1} \frac{1}{t} \mathbb{E}[\|\nabla Z_r(\theta_t)\|^2]}{\sum_{t=n_{\text{seed}}}^{n-1} \frac{1}{t}} \leq \frac{C}{\ln(\frac{n-1}{n_{\text{seed}}})}
\] (82)

\[
E[\|\nabla Z_r(\hat{\theta}_n)\|^2] \leq \frac{C}{\ln(\frac{n-1}{n_{\text{seed}}})}
\] (83)

Then, because \( n_{\text{seed}} = o(n) \), the right side converges to 0. So,

\[
E[\|\nabla Z_r(\hat{\theta}_n)\|^2] \to 0
\] (84)

\[
\|\nabla Z_r(\hat{\theta}_n)\|^2 \overset{P}{\to} 0
\] (85)

\[
\|\nabla Z_r(\hat{\theta}_n)\| \overset{P}{\to} 0
\] (86)

\[
\nabla Z_r(\hat{\theta}_n) \overset{P}{\to} 0
\] (87)

6.2 Convergence of gradient

In this section, we wish to show that

\[
\lim_{r \to 0} \nabla Z_r(\theta) = \nabla Z(\theta)
\] (88)

Along with the results in the main paper, this result will enable us to say that uncertainty sampling is asymptotically optimizing the zero-one loss. First, we must define some concepts. Each parameter vector \( \theta \) defines a decision boundary:

**Definition 13 (Decision Boundary).**

\[
B_\theta \overset{\text{def}}{=} \{ x : S(x, \theta) = 0 \}.
\] (89)

If \( S(x, \theta) \) is differentiable with respect to \( x \) and \( \nabla_x S(x, \theta) \neq 0 \) for all \( x \in B_\theta \), then by the implicit function theorem, \( B_\theta \) is a \((d - 1)\)-dimensional differentiable manifold and has measure zero.

**Proposition 14.** If \( S(x, \theta) \) is differentiable with respect to \( x \) and \( \nabla_x S(x, \theta) \neq 0 \) throughout \( B_\theta \), \( B_\theta \) is an \((d - 1)\)-dimensional differentiable manifold and has measure zero.

**Proof.** For any point \( b \in B_\theta \), since \( \nabla_x S(x, \theta) \neq 0 \), there is some direction where \( \nabla_x S(x, \theta) \) is non-zero. By the implicit function theorem, this means that there is a differentiable mapping from a
subset of $\mathbb{R}^{d-1}$ to a neighborhood of $b$ within $B_{\theta}$. Thus, $B_{\theta}$ is a $(d-1)$-dimensional differentiable manifold. Further, in $\mathbb{R}^{d}$, every open cover has a countable subcover. Thus, there is a countable family of local patches (with local differentiable charts). Since each local patch is a continuous mapping from a measure zero set $\mathbb{R}^{d-1}$, the local patches have measure zero. Since a countable union of measure zero sets has measure zero, $B_{\theta}$ has measure zero.

When this condition is satisfied, the decision boundary is well behaved, and $Z$ has nice properties. For these reasons, we will denote the set of parameters that meet this condition as the regular parameters, $\Theta_{\text{regular}}$.

**Definition 15 (Regular Parameters).**

$$\Theta_{\text{regular}} \overset{\text{def}}{=} \{ \theta : \forall x \in B_{\theta}, \nabla_x S(x, \theta) \neq 0 \}. \quad (90)$$

For logistic regression with identity features ($\phi(x) = x$), $\nabla_x S(x, \theta) = \theta$, so the only point not in $\Theta_{\text{regular}}$ is $\theta = 0$. For logistic regression with quadratic features, $\theta \cdot \phi(x) = x^\top A x + b^\top x + c$ (the parameters are $A$, $b$, and $c$), parameters where $A$ is non-singular and $c \neq \frac{1}{4} b^\top A^{-1} b$ are in $\Theta_{\text{regular}}$. Thus, the parameters not in $\Theta_{\text{regular}}$ have measure zero.

The key quantity of interest will require an additional assumption.

**Assumption 16 (Smooth PDF).** $p^*(x, y)$ has a smooth (all derivatives exist) probability density function for both values of $y$.

Recall that the decision boundary $B_{\theta}$ has measure zero for $\theta \in \Theta_{\text{regular}}$. Assumption [16] implies that there is zero probability mass on all decision boundaries corresponding to $\theta \in \Theta_{\text{regular}}$ ($\mathbb{P}(x \in B_{\theta}) = 0$ for $x \sim p^*$). This assumption and the ones from the main text ensure that the following quantity exists for $\theta \in \Theta_{\text{regular}}$.

$$\lim_{r \to 0} \frac{1}{2r} \int_{|S(x, \theta)| \leq r} (-y \nabla_x S(x, \theta)) dp^*(x, y) \quad (91)$$

This is a corollary of the following lemma,

**Lemma 17.** Suppose $\theta \in \Theta_{\text{regular}}$ and Assumption [16] holds. If $g(x)$ is smooth and has bounded support, $F(s) = \int_{S(x, \theta) < s} g(x) dx \quad (92)$

is smooth at 0.

**Proof.** For this proof, we rely heavily on the arguments in [Hoveijn 2007]. Since $g(x)$ has bounded support, for $||x|| \geq M_x$, $g(x) = 0$. Intuitively, this means we can define a function that is equal to $S(x, \theta)$ for $||x|| \leq M_x$ and is a small value $||x|| \geq M_x$ and mollify to make it smooth. More precisely, let $S_{\min} = \min(-2, \min_{|x| < 2M_x} S(x, \theta))$. Define $f(x)$ to be equal to $S(x, \theta)$ inside a ball of radius $2M_x$ and equal to $S_{\min}$ outside. Then mollify the function between balls of radius $M_x$ and $2M_x$. If we shift the function by $S_{\min}$, the function is smooth, always positive, and vanishes at infinity. Thus, it satisfies the Shifted class C functions of Definition 2 of [Hoveijn 2007].

Then, we can examine the function

$$G(s) = \int_{-1 < f(x) < s} g(x) dx, \quad (93)$$

which will have the same derivatives (if they exist) as $F(s)$ around 0. Note that $S_{\min} \leq -2 < -1$, so the integration between the level sets is well-defined.
0 is a regular value because $\theta \in \Theta_{regular}$. Further, we don’t need the non-degeneracy conditions of Hoveijn (2007) because $\nabla_x S(x, \theta)$ is continuous (Assumption 6) on a compact set (the support of $g(x)$) and thus is bounded below. And thus, a neighborhood around 0 are regular values.

We can use the flow box and diffeomorphism argument from Hoveijn (2007) to express the volume function as an integral with $h$ as the upper limit (see Proposition 7 of Hoveijn (2007)). While Hoveijn (2007) uses 1 as the integrand, the same argument holds for $g(x)$ as the integrand, and we recover that since $g(x)$ is smooth, the integral is smooth.

**Proposition 18.** If Assumptions [\ref{assumption6}, \ref{assumption7}, and \ref{assumption16} hold, then for $\theta \in \Theta_{regular}$,

$$\lim_{r \to 0} \frac{1}{2r} \int_{|S(x,\theta)| \leq r} (-y \nabla_\theta S(x, \theta)) dp^*(x, y)$$

exists.

**Proof.** Let $g_1(x) = \nabla_\theta S(x, \theta)p^*(x, y = -1)$ and $g_2(x) = \nabla_\theta S(x, \theta)p^*(x, y = 1)$. Let the corresponding integrals according to Lemma \ref{lemma17} be $F_1(s)$ and $F_2(s)$.

$$\lim_{r \to 0} \frac{1}{2r} \int_{|S(x,\theta)| \leq r} (-y \nabla_\theta S(x, \theta)) dp^*(x, y)$$

$$= \lim_{r \to 0} \frac{1}{2r} \int_{|S(x,\theta)| \leq r} \nabla_\theta S(x, \theta) dp^*(x, y = -1) - \lim_{r \to 0} \frac{1}{2r} \int_{|S(x,\theta)| \leq r} \nabla_\theta S(x, \theta) dp^*(x, y = 1)$$

$$= F_1'(s) - F_2'(s)$$

Now it remains to show that both $\lim_{r \to 0} \nabla Z_r(\theta)$ and $\nabla Z(\theta)$ are equal to that quantity for $\theta \in \Theta_{regular}$.

First, to show the result about $\nabla Z_r$, we will assume $q$ is smooth.

**Assumption 19.** $q$ is smooth, that is, all derivatives exist.

**Proposition 20.** If Assumptions [\ref{assumption3}, \ref{assumption6}, \ref{assumption7}, \ref{assumption16}, and \ref{assumption19} hold, then for $\theta \in \Theta_{regular}$,

$$\lim_{r \to 0} \nabla Z_r(\theta) = \lim_{r \to 0} \frac{1}{2r} \int_{|S(x,\theta)| \leq r} (-y \nabla_\theta S(x, \theta)) dp^*(x, y)$$

**Proof.** Recall that $q$ is an even function with bounded support. Thus, let $M_q$ be a number such that $|x| \geq M_q \implies q(x) = 0$. Then,

$$q(x) = q(|x|)$$

$$= \int_{|x|}^{M_q} q'(u) du$$

$$= - \int_{|x|}^{M_q} q'(u) du$$

$$= - \int_{0}^{M_q} 1[u \geq |x|]q'(u) du$$

From the proof of Theorem \ref{theorem8}}
\[
\n\nabla Z_r(\theta) = \frac{1}{Q_\infty r} \int q(S(x, \theta)/r)(-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
= -\frac{1}{Q_\infty r} \int_0^{M_q} \left[ u \geq \left| \frac{S(x, \theta)}{r} \right| \right] q'(u)(-y \nabla_{\theta} S(x, \theta)) du dp^*(x, y) \\
= -\frac{1}{Q_\infty r} \int_0^{M_q} \left[ |S(x, \theta)| \leq ru \right] q'(u)(-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
= -\frac{1}{Q_\infty} \int_0^{M_q} 2uq'(u) \frac{1}{2ru} \int_{|S(x, \theta)| \leq ru} (-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) du \\
\]

(103)

(104)

(105)

(106)

We can can the limit as \( r \to 0 \) of both sides. Further, we can use the dominated convergence theorem to bring the limit inside the outer integral.

\[
\lim_{r \to 0} \nabla Z_r(\theta) = -\frac{1}{Q_\infty} \int_0^{M_q} 2uq'(u) \lim_{r \to 0} \frac{1}{2ru} \int_{|S(x, \theta)| \leq ru} (-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) du \\
= -\frac{1}{Q_\infty} \int_0^{M_q} 2uq'(u) du \lim_{r \to 0} \frac{1}{2r} \int_{|S(x, \theta)| \leq r} (-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
\]

(107)

(108)

(109)

Now the only thing left to show is that the integral with respect to \( u \) is 1 and the result will follow. We can use integration by parts:

\[
-\frac{1}{Q_\infty} \int_0^{M_q} 2uq'(u) du = -\frac{2}{Q_\infty} \left( q(u)|_0^{M_q} - \int_0^{M_q} q(u) du \right) \\
= -\frac{2}{Q_\infty} \left( 0 \cdot M_q - q(0) \cdot 0 - \frac{Q_\infty}{2} \right) \\
= 1 \\
\]

(110)

(111)

(112)

(113)

\[
\]

Proposition 21. If Assumptions 6, 7, and 16 hold, then for \( \theta \in \Theta_{\text{regular}} \),

\[
\nabla Z(\theta) = \lim_{r \to 0} \frac{1}{2r} \int_{|S(x, \theta)| \leq r} (-y \nabla_{\theta} S(x, \theta)) dp^*(x, y). \\
\]

(114)

Proof. Noting that the derivative of the Heaviside step function is the Dirac delta function and that the Dirac delta function is even,

\[

Z(\theta) = \int H(-y S(x, \theta)) dp^*(x, y) \\

(115) \\

\nabla Z(\theta) = \int \delta(-y S(x, \theta))(-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
= \int \delta(S(x, \theta))(-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
= \int \delta(S(x, \theta))(-y \nabla_{\theta} S(x, \theta)) dp^*(x, y) \\
\]

(116) 

(117)
\[
\lim_{r \to 0} \frac{1}{2r} \int_{|S(x, \theta)| \leq r} (-y \nabla_\theta S(x, \theta)) dp^*(x, y) = \lim_{r \to 0} \int \frac{1}{2r} [-r \leq S(x, \theta) \leq r] (-y \nabla_\theta S(x, \theta)) dp^*(x, y) = \int \delta(S(x, \theta)) (-y \nabla_\theta S(x, \theta)) dp^*(x, y) = \nabla Z(\theta).
\]

Combining the previous propositions, we find that,

**Theorem 22.** If Assumptions 3, 6, 7, 16, and 19 hold, then for \( \theta \in \Theta_{\text{regular}}, \)

\[
\lim_{r \to 0} \nabla Z_r(\theta) = \nabla Z(\theta).
\]

20